Detecting and Tracking Regions of Interest for Remote Measurement of Vital Parameters

Estimation and Tracking of Keypoints Using Object Detection in Visual and Thermal Footage

Master's thesis in Physics

MADELEINE MÜLLER

DEPARTMENT OF PHYSICS
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2022
www.chalmers.se

Master's thesis 2022

Detecting and Tracking Regions of Interest for Remote Measurement of Vital Parameters

Estimation and Tracking of Keypoints Using Object Detection in Visual and Thermal Footage

MADELEINE MÜLLER

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Physics
Chalmers University of Technology
Gothenburg, Sweden 2022

Detecting and Tracking Regions of Interest for Remote Measurement of Vital Parameters
Estimation and Tracking of Keypoints Using Object Detection in Visual and Thermal Footage.
MADELEINE MÜLLER

Supervisor: Farzad Kamrani and Marianela Garcia, Swedish Defence Research Agency
Examiner: Christian Forssén, Department of Physics

Master's Thesis 2022
Department of Physics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Illustration of the facial grid implemented to model the facial keypoints spatial relation in the thermal domain.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2022

Detecting and Tracking Regions of Interest for Remote Measurement of Vital Parameters Using Deep Learning
Estimation and Tracking of Keypoints Using Object Detection in Visual and Thermal Footage
Madeleine Müller
Department of Physics
Chalmers University of Technology

Abstract

The initial assessment of a mass casualty incident is essential to effectively conduct a rescue operation. The survival rate is affected by the complexity of the incident, and it is therefore imperative to enhance the operational capacities of emergency medical services and civil protection agencies in mass casualty incidents. This thesis investigates the possibilities for an unmanned aerial vehicle (UAV) to detect and track regions of interest for remote measurement of vital parameters in visual and thermal footage for first response triage purposes. The regions of interest are the nose, mouth, and chest, and the UAV characteristic taken under consideration in this thesis is image blur due to random camera motion. In this thesis, we take an object detection approach and implement the keypoint estimation framework KAPAO and the tracking algorithm SORT in several different experimental setups. Using KAPAO and SORT, we achieve a good result. For the detection in the thermal domain, the model created by transferring knowledge from the visual to the thermal domain achieves the highest performance. We also consider adversarial training on random motion blur, however the result shows a minimal impact on the model performance in the presence of characteristic low-altitude UAV motion blur. Regarding the tracking of the regions of interest, the result concludes that the SORT algorithm improves the performance compared to assigning tracking identification numbers based on frame-to-frame differences. The result shows that the distance to the subjects and the image quality impacts the performance. Compared with previous work on remote measurement of vital parameters, the algorithms of this thesis achieve a nearly perfect score on corresponding distances. If the distances are realizable in a UAV triage application is however unknown and has to be investigated further. Moreover, the work of this thesis problematizes the low-altitude UAV motion blur which poses a potential limitation in a potential UAV triage application. An alternative could hence be to use optical stabilization measurement for blur reduction.

Acknowledgements

I would like to thank my supervisors at FOI, Farzad Kamrani and Marianela Garcia, for their participation, encouragement, and guidance. Furthermore, I would also like to thank my examiner at Chalmers Prof. Christian Forssén for his advice and support during the project. Finally, I would also like to thank the ones helping me review my work.

Madeleine Müller, Stockholm, October 2022

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis project listed in alphabetical order:

ANN	Artificial Neural Networks
CNN	Convolutional Neural Network
COCO	Microsoft Common Objects in Context
FOI	Swedish Defence Research Agency
IOU	Intersection of Union
KAPAO	Keypoints And Poses As Objects
MOTA	Multiple Object Tracking Accuracy
NMS	Non-maximum Suppression
OKS	Object Keypoint Similarity
PPG	Photoplethysmography
ROI	Region of Interest
rPPG	Remote Photoplethysmography
SORT	Simple Online Realtime Tracking
TFW	Thermal Faces in the Wild
UAV	Unmanned Aerial Vehicle
YOLO	You Only Look Once

Contents

List of Acronyms **ix**

1 Introduction **1**
1.1 Problem Definition and Purpose . . . 1
1.2 Research Question . . . 2
1.3 Scope and Limitations . . . 2
1.4 Swedish Defence Research Agency . . . 2
1.5 Report Overview . . . 3

2 Background **5**
2.1 Prehospital Triage . . . 5
2.2 Remote Measurement of Vital Parameters . . . 5
2.2.1 Remote Measurement of Heart Rate . . . 6
2.2.2 Remote Measurement of Body Temperature . . . 6
2.2.3 Remote Measurement of Respiratory Rate . . . 6
2.3 Related Work . . . 7

3 Theory **9**
3.1 A Brief Introduction to Deep Learning . . . 9
3.1.1 Artificial Neural Networks . . . 9
3.1.2 Convolutional Neural Networks . . . 11
3.1.2.1 Convolutional Layers . . . 11
3.1.2.2 Pooling Layers . . . 11
3.2 Object Detection . . . 12
3.2.1 You Only Look Once . . . 12
3.3 Keypoint Estimation . . . 14
3.3.1 Multi-Person Keypoint Estimation . . . 15
3.3.2 Modeling Keypoints and Poses as Objects . . . 15
3.4 Thermal Imaging . . . 16
3.5 Transfer Learning . . . 17
3.6 Adversarial Perturbations . . . 17
3.6.1 Motion Blur . . . 17
3.7 Object Tracking . . . 18
3.7.1 Simple Online Realtime Tracking . . . 19
3.8 Evaluation Metrics . . . 19
3.8.1 Precision and Recall . . . 19

3.8.2 Average Precision 20
3.8.3 Multiple Object Tracking Accuracy 20

4 Method **23**
4.1 Keypoint Estimation Datasets 23
4.1.1 The Common Objects in Context Dataset 23
4.1.2 Thermal Faces in the Wild 25
4.2 Video Datasets for Tracking 26
4.2.1 300 Videos in the Wild 26
4.2.2 RGBT234 Dataset 27
4.3 Motion Blur Augmentation 28
4.4 Triangulation of the Forehead and Chest 29
4.5 Network Architecture 29
4.6 Experimental Setup 31
4.6.1 Detection 31
4.6.2 Tracking 32
4.7 Evaluation 33

5 Results **35**
5.1 Blur Impact 35
5.2 Detection in the Thermal Domain 35
5.3 Tracking of ROIs 36

6 Discussion **39**
6.1 Detection in the Visual Domain 39
6.2 Detection in the Thermal Domain 40
6.3 Dataset Biases 40
6.4 Tracking of ROIs 41
6.4.1 Evaluation on the 300-VW Dataset 41
6.4.2 Evaluation on the RGBT234 Dataset 42
6.5 Adaption to UAV Applications 42

7 Conclusions **43**
7.1 Future Work 43

Bibliography **45**

1
Introduction

A mass casualty incident is referred to as an event where the number of casualties exceeds the available medical resources. Mass injuries can, for example, be caused by transportation accidents, terrorism, fires, or natural disasters. The initial assessment of a mass casualty incident is essential to effectively conduct a rescue operation. As a first response, triage is performed by medical staff at the scene. Triage is a medical procedure of evaluating and classifying injured based on their vital signs to prioritize patient care. This becomes a rather demanding task as the magnitude and complexity of the mass casualty incident increases, which can decrease the survival rate [1].
To address the difficulties faced in mass casualty incident operations the Swedish Defence Research Agency (FOI) is involved in a research project led by the Swedish Transport Administration. The project aims to develop an unmanned aerial vehicle (UAV) system integrated with artificial intelligence for search and rescue purposes, to simplify operations and increase the survival rate. The idea is to use a UAV equipped with sensors such as a standard color (RGB) camera and a thermographic camera to collect data from the scene of the incident. As a part of the assessment, the UAV needs to be able to measure vital parameters, and in order to measure a vital parameter, the UAV has to be able to detect and track the corresponding region of interest (ROI) where the measurement can be performed.

1.1 Problem Definition and Purpose

This thesis project aims to investigate the possibilities for a UAV to detect and track ROIs for remote measurement of vital parameters. Specific ROIs are the forehead, nose, mouth, and chest, where body temperature, respiration rate, and pulse can be measured for triage [2]. There are state-of-the-art deep learning frameworks for the detection of human body features. However, they are not able to detect all the ROIs mentioned, nor are they robust enough to be implemented in a UAV application [3]. Therefore, this thesis aims to develop and evaluate a robust method for detecting and tracking the mentioned ROIs in visual and thermal footage in multi-person scenarios. Robust is here defined as the ability to handle perturbation caused by characteristic low-altitude UAV motion. As there to this point are no publicly available low-altitude visible or thermal UAV data for detection and tracking of mentioned ROIs, the UAV motion characteristics are to be simulated on still images. The task at hand is to be accomplished using deep learning combined with adversarial training.

1.2 Research Question

This thesis project aims to answer the research question:

- To what extent is it possible to detect and track regions of interest for remote measurement of vital parameters in RGB and thermal footage and in the presence of characteristic low-altitude UAV motion blur?

1.3 Scope and Limitations

- This project aims to detect and track regions of interest for the measurement of vital parameters in RGB and thermal footage.
- The ROIs for this project are the forehead, mouth, nose, and chest. Due to a lack of annotations, the chest will only be evaluated for detection in the visual domain.
- The project focuses on implementing deep learning models for keypoint estimation for detection and tracking of mentioned ROIs.
- The UAV characteristic taken under consideration in this project is UAV motion blur. For example, considering different UAV pitch angles are outside the scope of this project.
- The camera motion blur is to be simulated from still images and evaluation of the effectiveness with respect to real UAV motion blur is outside the scope of this project.
- This project does not employ real data collected in a mass casualty incident. Real data collected in a mass casualty incident is considered sensitive as it contains medical information, and using such data would by Swedish law require an ethical analysis preformed by the Ethics Review Authority[1]. By excluding sensitive data, this thesis can focus on the technical aspects rather than the ethical aspects of data usage.
- Field testing and performing measurements of vital parameters are outside the scope of this project. Such activities would be required before deploying the final system. Moreover, it would also require the approval of the Ethics Review Authority due to the ethical aspects concerning human test subjects.

1.4 Swedish Defence Research Agency

This thesis is carried out in collaboration with FOI. FOI is a defence research institute and a government authority operating under the Swedish Ministry of Defence. The agency conducts research within the area of safety and security of the society, including strategic decision-making and crisis management [5]. On behalf of

[1]The mission of the Swedish Ethics Review Authority is to protect the individuals and the human values within research, and any research carried out on human requires their approval. The Swedish Ethics Review Authority considers the ethical aspects concerning the project and only approves a project and whether the possible benefits of the results exceed the endangerment of the persons subjected for investigation [4].

the Swedish Transport Administration and the EU project Nightingale, FOI is developing a UAV triage application for mass casualty incidents with the ultimate goal of increasing the survival rate. This thesis project aims to contribute to FOI's UAV triage application research. Furthermore, one should note that this is a civil project.

1.5 Report Overview

This report is organized as follows: Chapter 2 provides background knowledge of triage and remote vital parameter measurement, as well as related works; Chapter 3 provides the theoretical foundation of the project; Chapter 4 presents the methodology and the experiments conducted; Chapter 5 presents the results; Chapter 6 provides a discussion of the results and the research question along with some ethical aspects in Section 6.3, and lastly Chapter 7 which presents the conclusion and future works.

2
Background

This chapter is to provide a background for remote triage measurement. Section 2.1 describes the fundamentals of prehospital triage performed at the scene of the incident. Section 2.2 presents existing methods for remote measurement of the vital parameters presented in Section 2.1 along with corresponding regions of interest in this thesis. Finally, in Section 2.3 related work concerning the detection and tracking of ROIs for vital parameter measurement is presented.

2.1 Prehospital Triage

Triage is the medical procedure of evaluating and classifying injured to prioritize patient care implemented when the demand exceeds the available resource. There are different types of triage, and prehospital triage is referred to as the first response triage performed on the scene of the incident. Prehospital triage aims to assess the situation in order to effectively conduct a rescue operation.
Indicators of which prehospital triage assessments are built upon are:

- ambulatory,
- clear airways,
- respiratory rate,
- radial/peripheral pulse,
- and level of consciousness [6].

Ambulatory, the ability to walk, is a primary divider. Walking requires a sufficient central nervous system and blood pressure, and walking persons will thereby be down-prioritized. The second point of assessment is clear airways. If the person cannot breathe, it will get down-prioritized due to a low survival rate. If clear airways are the assessment of if the person is breathing, the respiratory rate is the assessment of how the person is breathing. The respiratory rate is an indicator of trauma to the airways and lungs, and an abnormal breathing pattern will get prioritized. The pulse parameter is used to estimate blood pressure to detect life-threatening internal and external bleeding. The last point is the level of consciousness which is assessed by whether the person can follow commands or not. Verbal and motor responses are indicators of neurological function [6].

2.2 Remote Measurement of Vital Parameters

This section presents different methods for remote measurement of vital parameters of interest for triage assessment. The ROI in this thesis is limited to the ROI for

remote measurement of heart rate, body temperature, and respiration rate, which are presented below.

2.2.1 Remote Measurement of Heart Rate

The commonly implemented method for remote measurement of heart rate and oxygen saturation is remote photoplethysmography (rPPG) [2]. Remote PPG operates on the same basis as traditional PPG, which illuminates the skin and measures the variation of light absorption. The light is absorbed by the blood in the capillaries, and the absorption is correlated to the dilation and constriction of the capillaries from which the heart rate can be measured [7]. The forehead is a region of interest for rPPG measurement due to its large vascularization and thin skin [8]. Therefore, the forehead constitutes an ROI in this project.

2.2.2 Remote Measurement of Body Temperature

The forehead is also an ROI for remote measurement of body temperature as the forehead temperature is highly correlated to the internal body temperature due to the large vascularization and thin skin. The temperature at the forehead can therefore be used to detect hyperthermia and hypothermia. This correlation cannot be seen in other parts of the body; the temperature in limbs can be highly different from the core body temperature [8].

2.2.3 Remote Measurement of Respiratory Rate

Different types of sensor data can be used to measure the respiratory rate remotely. The respiratory rate can be estimated remotely by acoustic analysis, analysis of temperature fluctuation due to respiration, chest motion analysis, and rPPG.
The acoustic-based methods utilize the breathing sound to determine the respiratory rate. Such methods have proven effective but also sensitive to background noise [2]. The sensitivity poses a difficulty in unconstrained environments, which makes acoustic analysis unsuitable for UAV triage applications.
The respiration rate can be estimated remotely by analyzing the temperature difference in inhaling and exhaling air in thermal footage [2]. Detecting and tracking of mouth and nose in thermal images are hence of interest in this thesis. Temperature fluctuation-based methods have proven to be sensitive to motion [2] which make the tracking aspect particularly relevant.
Estimation of respiration rate can also be performed through respiration motion analysis. Chest motion analysis has proven to be a robust method for respiration rate estimation in multi-person scenarios [2]. Moreover, it can be performed on different types of sensor data such as RGB, depth, and thermal footage. Detection of the chest is hence within the scope of this project. The chest area is a large ROI which poses an advantage. However, clothing can potentially obstruct the analysis of chest motion, which can become problematic in outdoor applications.
Respiration rates can be observed in rPPG signals as respiratory arrhythmia affects the heart rate and blood volume signal. Hence, the respiration rate can be

modulated from the rPPG signal [9]. The fundamentals of rPPG are described in Section 2.2.1.

2.3 Related Work

In the aftermath of the COVID-19 pandemic, the detection of ROIs for remote measurement of infection indicators has gained more attention. Previous research by Rodriguez-Lozano et al. [8] and Muller et al. [10] propose two methods for segmentation of the forehead and the nose region in thermal images. Rodriguez-Lozano et al. implement a trigonometric segmentation method where the forehead is segmented by fitting an ellipse onto the face and extracting the upper part of the ellipse [8].
In difference from Rodriguez-Lozano et al., Muller et al. implement a deep learning-based method for segmenting the forehead, nose, mouth, lower face, and eyebrows. Muller et al. implement a conditional generative adversarial network (cGAN), trained on a custom annotated database they have created. The dataset is, however, small and limited to controlled lab environments. In addition to segmentation, Muller et al. successfully extract the temperature levels at the different ROIs [10].
Another approach for detecting and tracking ROIs is triangulation from existing keypoints. Djeldjli [11] presents a method for remote assessment of blood pressure and arterial stiffness in RGB footage. The measurements are performed remotely at the forehead, triangulated from the eyebrows and the facial bounding box, using the facial keypoint estimator Dlib [12] and the OpenCV [13] face detector.
Yang et al. [14] propose a method for remote measurement of blood pressure, heart rate, and respiration rate in RGB and thermal footage, including detection of the forehead and nostrils. Yang et al. implement RetrinaFace [15], a deep learning framework for detecting facial keypoints, from which the ROIs are triangulated based on different facial feature distances. Although vital parameters are measured in visual and thermal domains, the ROIs are only detected in the RGB videos. The defections are transferred to corresponding thermal videos via an image alignment process made possible as the videos are recorded simultaneously.
Compared with previous work, this thesis is not limited to controlled settings. In this thesis, we implement data collected under uncontrolled settings in outdoor environments that represent a potential UAV triage application. Moreover, we are considering more considerable distances to the subject than in previous work.

3
Theory

This chapter contains the theory for this thesis focusing on the computer vision elements. Section 3.1 provides a brief introduction to deep learning and convolutional neural network. Section 3.3 the fundamentals of keypoint estimation are described along with a keypoint estimation framework of interest for this thesis project. As motion blur is to be considered in this thesis, Section 3.6 provides the theory regarding adversarial perturbation and motion blur kernels. Furthermore, Section 3.7 is dedicated to object tracking and the tracking algorithm implemented in this thesis.

3.1 A Brief Introduction to Deep Learning

Deep learning algorithms are based upon artificial neural network architectures, and deep learning models are consequently often referred to as deep neural networks. Deep neural networks are extensive models characterized by their complexity. Training a deep learning model requires a considerable amount of data and computational power [16]. This section describes some fundamental deep learning concepts of interest for this thesis.

3.1.1 Artificial Neural Networks

Artificial neural networks (ANN), or simply neural networks, are computational models inspired by biological neural networks. An ANN consists of artificial neurons, or nodes, which are connected and arranged in layers. An illustration of a basic ANN is illustrated in Figure 3.1. The first and last layers of an ANN correspond to the input and output layers, while the intermediate layers are referred to as the hidden layers. Moreover, the depth of a neural network is given by the number of hidden layers [16].
Mathematically, a node is a weighted summation passed through an activation function. Given the input values $x_1, ...x_m$, the corresponding weights $w_1, ...w_m$, and an activation function $\phi(\cdot)$ the output of the node is given by

$$y = \phi\left(\sum_{i=1}^{m} x_i w_i + b\right).$$

The weights and the bias therm b added to the weighted sum are learnable parameters determined by training. The weights control the strength of the connection of the nodes, i.e., how the inputs influence the output, and the bias terms regulate

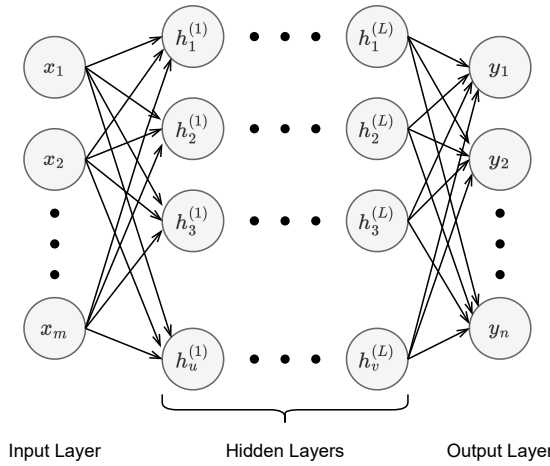


Figure 3.1: An example of a feed-forward ANN. The circles correspond to nodes which are arranged in layers.

the flexibility of the nodes by shifting the activation functions. Different functions can be employed for activation. A commonly implemented activation function is the sigmoid function

$$\phi(z) = \frac{1}{1 + e^{-z}},$$

which returns a value between 0 and 1. The functionality of a node is further illustrated in Figure 3.2 where z corresponds to the weighed summation.

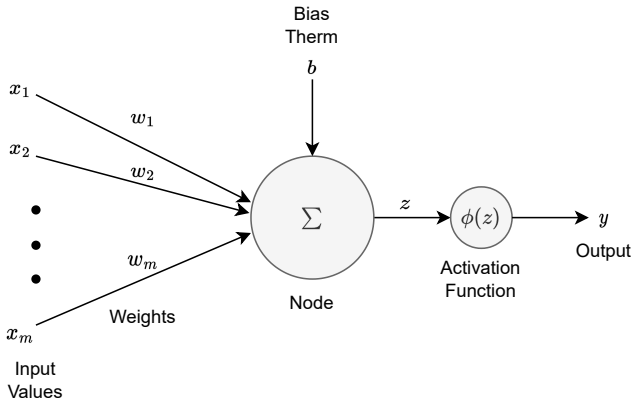


Figure 3.2: Illustration of an artificial neuron.

An ANN can be described as a non-linear transformation $\boldsymbol{F}$ which maps an input $\boldsymbol{X} \in \mathbb{R}^{m \times N}$ on to an output $\boldsymbol{Y} \in \mathbb{R}^{n \times N}$; $\boldsymbol{F} : \mathbb{R}^{m \times N} \to \mathbb{R}^{n \times N}$ where N denotes the number of samples while m and n correspond to the input and output dimensions respectively. The transformation is governed by the weights and the biases, $\theta = \{\boldsymbol{W}_l, \boldsymbol{b}_l\}_{l=1}^{L+1}$, of the trainable layers. There are $L + 1$ trainable layer corresponding to the the hidden layers L along with the output layer. The output is generated by

propagating the input through the ANN and the output can therefore be described as $\boldsymbol{Y} = \boldsymbol{F}(\boldsymbol{X}, \theta) = \phi_{L+1}(\phi_L(...(\phi_1(\boldsymbol{X}))))$ [17].
Training a neural network is the process of tuning the parameters θ to improve model performance. The objective is to minimize the difference between the actual and the predicted output described by the loss function $\mathcal{L}(\boldsymbol{F}(\boldsymbol{X}, \theta), \hat{\boldsymbol{Y}})$. Accordingly, the training process is a minimization problem where the optimal parameter is given by

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{F}(\boldsymbol{X}, \theta), \hat{\boldsymbol{Y}}).$$

The training of a neural network is an iterative process where the parameters are learned through stochastic gradient descent optimization [17].

3.1.2 Convolutional Neural Networks

Convolutional neural network (CNN) is a frequently implemented architecture within computer vision. CNN neurons are arranged in three dimensions; the height and width corresponding to the spatial dimension and the depth equivalent to the depth of the input. The three-dimensional representation makes CNNs suitable for image processing. CNN architecture comprises three types of layers; the convolutional layers and pooling layer for feature learning, and the fully connected layers for classification [18]. The fully connected layers are conventional ANN layers described above, while convolutional layers and pooling layers are described below.

3.1.2.1 Convolutional Layers

Convolutional layers pose as filters for feature extraction, which consist of convolutional kernels corresponding to matrices of weights learned by training. The kernels are passed sequentially over the input for which the kernel is multiplied with each subregion, followed by a summation to obtain a feature map for activation. Hence, the size of the acquired feature map is determined by the size of the kernel, as well as the division of subregions. A convolutional operation is illustrated in Figure 3.3. A neuron in a convolutional layer is only connected to neurons in the preceding layer within reach of the convolutional kernel. Due to the local connectivity of the conditional layers, the initial layers of a CNN will capture local features while the latter ones will capture global features [17, 18].

3.1.2.2 Pooling Layers

A pooling layer can follow a convolutional layer to reduce the spatial dimension further. This is known as downsampling, which reduces the complexity of the model. There are two pooling operators: the max pooling operator, which extracts the maximum value from a specific region, and the average pooling operator, which extracts the average value. The dimension is reduced by dividing a feature map into subregions and by letting a pooling operation act on the regions. Hence, the reduction is determined by the size of the regions [17, 18]. The downsampling process is illustrated in Figure 3.4. Note that the pooling layer acts on each feature map independently to maintain the depth.

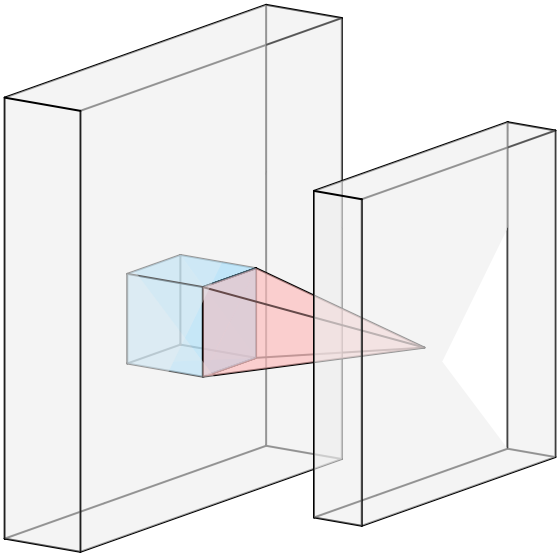


Figure 3.3: Illustration of a convolutional operation. The filter acts on the input layer and the size of the output is determined by the dimension of the filter and how the filter is traversed.

3.2 Object Detection

Object detection is the computer vision task of detecting and localizing objects within images and videos. Deep learning-based object detectors are commonly divided into one-stage and two-stage models. The two-stage detectors consist of a region proposal step for the prediction of object bounding boxes and a classification step for the classification of the predicted bounding boxes. In contrast to the two-stage detectors, the one-stage detectors perform region proposal and classification simultaneously [19]. The one-stage detectors are known for their inference speed, making them suitable for UAV applications, compared with the two-stage detectors, which demand greater computational resources [20]. The following section is dedicated to the one-stage object detector You Only Look Once implemented in this thesis.

3.2.1 You Only Look Once

You Only Look Once [21, 22, 23, 24, 25] (YOLO) is a state-of-the-art one-stage object detector known for its inference speed, making it suitable for real-time applications. As a one-stage detector, YOLO performs object proposal and classification simultaneously by dividing the input images into an $S \times S$ grid. Within each cell, B object bounding boxes are predicted and assigned a confidence score reflecting upon the accuracy of the corresponding bounding box. This process yields a surplus of bounding boxes which are passed through a non-maximum suppression (NMS) filter to determine the final detections. The NMS removes bounding boxes with confidence scores below a given threshold and then suppresses the bounding boxes with a high intersection over union (IOU) with high confidence class-specified bounding boxes. The YOLO pipeline is illustrated in Figure 3.5.

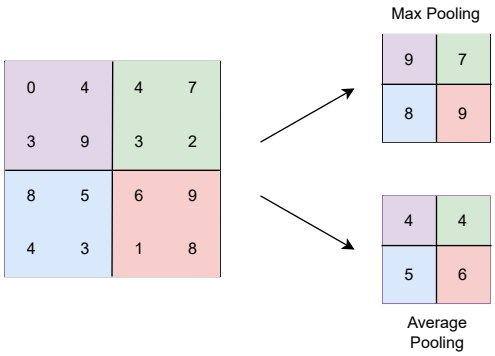


Figure 3.4: Max pooling and average pooling operations. The max pooling operator extracts the maximum value of the region, and the average pooling operation yields the average value of the region.

The YOLO model was initially proposed by Redmon et al. in 2015, after which four subsequent versions have been released. The original version, YOLOv1 [21], consists of a CNN of 24 convolutional layers for feature extraction followed by two fully connected layers for the prediction of bounding boxes. The updated second version, YOLOv2 [23], released the year after, uses a Darknet-19 backbone for feature extraction. Darknet-19 is a CNN architecture of 19 convolutional layers and five max pooling layers. Moreover, in the second version, the fully connected layers are removed and replaced with anchor boxes for the prediction of bounding boxes. The architectural updates show an increment in inference speed.

In the third version, YOLOv3 [22], also by Redmon et al., the backbone is further updated to a Darknet-53 architecture (to be described in more detail in Section 4.5). Moreover, YOLOv3 employs a feature pyramid network (FPN) for feature fusion which concatenates the feature maps from different layers of the backbone. Furthermore, in the third version, residual block, skip connections, and up-sampling are introduced. These changes show a significant improvement in accuracy and performance on small objects.

The fourth version, YOLOv4 [24], was released by Bochkovskiy et al. in 2020. Compared to its successor, YOLOv4 relies on a CSPDarknet53 backbone, and the FPN is replaced with a spatial pyramid pooling (SPP) layer and path aggregation network (PANet) for feature aggregation at multiple scales. Bochkovskiy et al. introduce a “Bag of Freebies” and a “Bag of Specials” to further improve the performance. The bag of freebies is data augmentation techniques to expand the dataset artificially. Data augmentation techniques employed are photometric, such as brightness, contrast, hue, and saturation distortion, and geometric distortions such as random scaling, cropping, and flipping. The most recent version as of March 2022, is the fifth version YOLOv5 which is an open-source project maintained by Ultralytics[1]. YOLOv5 is a PyTorch implementation of YOLOv4, making it more user-friendly [25].

[1] `https://ultralytics.com/yolov5`

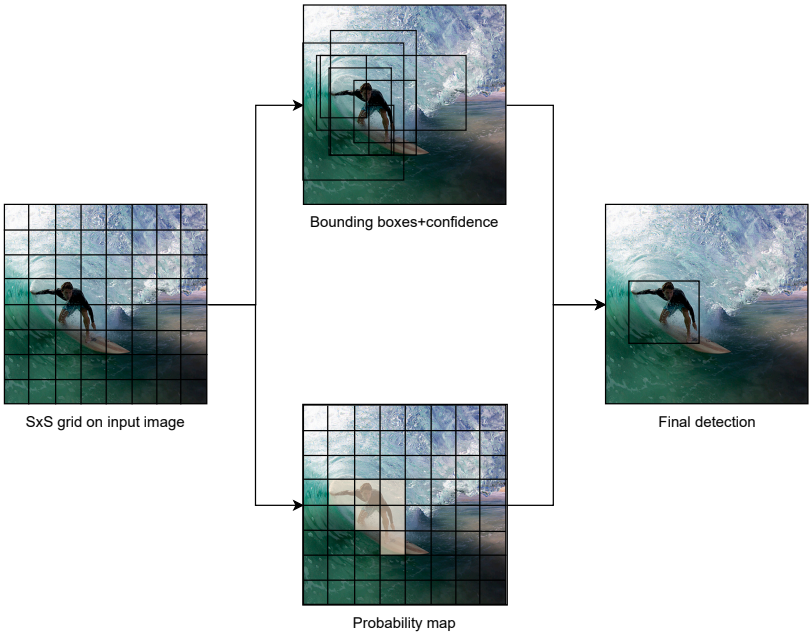


Figure 3.5: Illustration over the YOLO pipeline. The input image is divided into an $S \times S$ grid. For each cell, B object bounding boxes are predicted and assigned a confidence score along with the class probabilities. The predictions are then fused with NMS to obtain the final predictions.

3.3 Keypoint Estimation

Vision-based keypoint estimation aims to detect and localize points of interest in visual input data such as images and videos. In particular, person keypoint detection aims to estimate human body landmarks. Conventional human body landmarks annotations are: eyes, ears, nose, shoulders, elbows, hands, hips, knees, and feet, which are visualized in Figure 3.6.

To this point, CNN-based heatmap regression is the prevalent approach for keypoint estimation. In a keypoint heatmap, the grid constitutes the input image, and the pixels encode the probability of the corresponding pixel being a keypoint. Given that the model aims to determine N keypoints, there will be N keypoint heatmaps containing the spatial confidence distribution of the corresponding keypoint. CNN-based heatmap regression leverage CNN to regress the heatmaps onto the input images, and the final keypoint predictions are obtained by extracting the maximum indices of the heatmaps [26, 27].

The heatmap-based methods have become the dominating keypoint estimation approach due to their performance. However, generating and processing the heatmaps are computationally costly. High-resolution heatmaps are required to achieve accurate results, and the accuracy/speed trade-off is one of the major drawbacks of the heatmap-based approaches [26].

Figure 3.6: Human body keypoint annotations.

3.3.1 Multi-Person Keypoint Estimation

Multi-person keypoint estimation can be categorized into top-down and bottom-up methods. The challenge in a multi-person setting is associating the detected keypoints with the corresponding persons, which becomes especially challenging when people occlude each other.
Top-down approaches employ a person detector and a single-person keypoint estimator. The detector generates human detection bounding boxes, and the single-person keypoint estimator is then applied to each bounding box separately to create a multi-person keypoint estimation. Top-down approaches are thereby classified as two-stage processes. The bottom-up methods simultaneously detect all keypoints and assign them to the corresponding person [27]. Hence, bottom-up methods are considered to be one-stage approaches. The bottom-up approaches are generally faster; however, they are less accurate than their top-down counterparts [26, 27]. On the other hand, the top-down approaches are more sensitive to person occlusion as occlusion makes the person detection more likely to fail [27].

3.3.2 Modeling Keypoints and Poses as Objects

To address the heatmap drawbacks, McNally et al. introduced the heatmap-free multi-person keypoint detector KAPAO [26] (Keypoints And Poses As Objects). KAPAO models keypoint and their spatial relations as objects within a YOLOv5 framework. This heatmap-free development has significantly improved the accuracy and inference speed compared to heatmap-based state-of-the-art keypoint estimation frameworks [26].
KAPAO is an object detection-based keypoint estimator which treats keypoints and poses as objects. The keypoint objects are defined by a detection bounding box of equal width and height and its center coordinates. In contrast, the pose objects comprise of a traditional object bounding box along with its associated set

of keypoint objects. The keypoint objects are dedicated to keypoints with strong local features such as nose and eyes. Unlike the keypoint objects, the pose objects contain spatial relation information. The global understanding of the pose object makes it suitable for estimation of keypoint lacking in local features such as hips and shoulders [26].
KAPAO employs a YOLO-based network for the detection of keypoint and pose objects. The network takes an input image and maps it onto output grids containing the predicted keypoints and pose objects. Rather than using a bottom-up approach, the detection of keypoints and pose objects occur simultaneously. Subsequently, the output grids are passed to NMS to eliminate redundant proposals by suppressing candidates with low confidence scores by comparing the overlaps. The NMS is applied on the keypoints and pose objects separately before being fused with a matching algorithm to obtain the final body pose prediction [26]. The pipeline is illustrated in Figure 3.7.

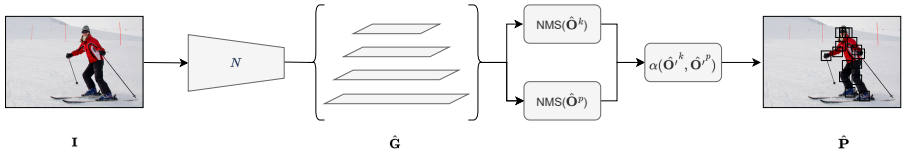


Figure 3.7: An illustration of the KAPAO framework. The detection network N maps the input image $\mathbf{I}$ onto a set of output grids $\hat{\mathbf{G}}$ containing both the predicted keypoint $\hat{O}^k$ and pose objects $\hat{O}^p$. The predictions are then separately filtered through an NMS before being fused by the matching algorithm α, which produces the final prediction $\hat{\mathbf{P}}$.

3.4 Thermal Imaging

The visual and thermal domains have been taken under consideration in this project, and this section presents the basics of thermal imaging. The electromagnetic spectrum is divided into subspectrums based on energy levels. The visible range includes the wavelengths visible to the naked eye, and the infrared (IR) region includes wavelengths of 0.7–1000 μm [28]. Furthermore, the IR spectrum is divided into subcategories; near, short, mid, long, and far wavelength IR. According to Wien's law, the human body emits a wavelength of 9.4 μm, which falls within the range of long wavelength infrared radiation (LWIR) of 8-15 μm [28]. Hence, LWIR sensors can be used for human detection.
The electromagnetic field can be used to create a visual representation of an object. RGB, or true color imaging, refers to imaging within the visible spectra. Likewise, thermal imaging can be applied to visualize thermal energy, i.e., IR radiation. A thermal heat signature consists of a spectral and spatial intensity distribution from apparent temperature differences to the background [29].
In contrast with visual imaging, thermal imaging is less sensitive to illumination degradation and visual obstruction, which poses an advantage in poor weather conditions and at night. On the other hand, thermal imaging depends on the surrounding

environment and background temperatures. Lighting conditions, geographic locations, weather conditions, time of capture, and materials reflective properties are examples of factors that impact the surrounding environment transmission. Moreover, one should be aware that atmospheric transmission caused by, e.g., humidity and aerosols also affects the thermal image resolution [29].

3.5 Transfer Learning

Transfer learning within machine learning aims to transfer knowledge gained from solving one problem to solving another, different but related problem. Transfer learning can be used for generalization across domains and transferring knowledge from a data-rich to a data-poor domain. In the context of this thesis, transfer learning can be employed for transferring knowledge from the visual to the thermal domain [30].
Rather than initializing a model with random weights, knowledge can be transferred by implementing the weights of a pre-trained model. Knowledge can be transferred by freezing layers and by fine-tuning. Freezing methods employs the pre-trained model without modifications. The weights of the pre-trained model are said to be frozen, meaning that they will not be adjusted during training. The model is adapted by adding additional layers to be modified by further training on the target data. Fine-tuning, rather than freezing the weights, adjusts the pre-trained models by further training the model on the target data [30].

3.6 Adversarial Perturbations

Computer vision models are known to be sensitive to adversarial attacks, i.e., misclassification due to perturbations in input data. This is a well-documented issue that poses a threat in real-world implementations [3, 31, 32, 33]. An adversarial perturbation is a noise that causes misclassification when added to an image. It can be quasi-imperceptible, i.e., invisible to a human eye, and it can be universal and cause misclassification when added to any arbitrary image [32]. Motion blur is an example of an adversarial that can cause misclassification.
There are two common defense strategies for dealing with adversarial perturbations: 1) model alternation and 2) image prepossessing [32, 34, 35]. Adversarial training is a commonly implemented defense strategy for increasing model robustness by including adversarial samples in the training dataset. Image prepossessing, on the other hand, aims to remove the adversarial perturbation rather than alternating the model [35].

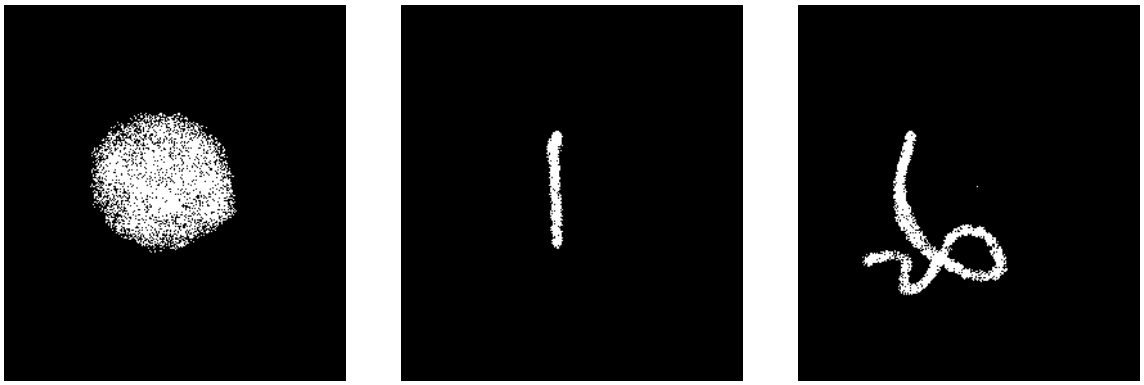
3.6.1 Motion Blur

State-of-the-art networks are typically trained and evaluated on large high-quality artifact-free datasets. Using perfect data for training causes a decrease in performance in the presence of quality distortion such as motion blur [33, 36]. Possible

solutions for handling motion blur are adversarial training on motion blur and deblurring, which both require knowledge about the blur kernel.
An image subjected to motion blur can mathematically be defined as

$$y = k * x + n, \tag{3.1}$$

where x corresponds to the equivalent sharp image, k the blur kernel, n some additive noise, and $*$ is the convolution operator [37, 38]. As the equation implies, if the kernel is known, it is possible to extract the sharp images from their blurry counterpart. However, determining the true kernel is not a trivial task.
By the definition in Equation 3.1, motion blur can be simulated by implementing synthetic motion blur kernels. A selection of different synthetic blur kernels is presented in Figure 3.8. The disk kernel in Figure 3.8 (a) can be used to simulate defocus blur, and the oriented box kernel in Figure 3.8 (b) can be applied to simulate linear motion blur [33]. The kernel in Figure 3.8 (c) is a random kernel that can be used for simulating random motion [33, 38, 39, 40]. In order to capture the blur kernel for a UAV, one has to take the complex motion pattern of a UAV into account, including irregular motion caused by hovering, wind, or turbulence.



(a) Defocus. **(b)** Linear motion. **(c)** Random motion.

Figure 3.8: Different synthetic blur kernels.

Vasiljevic et al. [33] investigate the image blur impact on state-of-the-art detection networks. They implement the types of synthetic blur kernels presented in Figure 3.8 and observe a decrease in performance. Moreover, they are able to improve the performance by fine-tuning and adversarial training. Vasiljevic et al. observe that similar static distribution tends to generalize among different kernels, i.e., fine-tuning on defocus blur increases the model performance on camera shake blur and vice versa. It is a significant finding, however, in a more recent study by Sayed et al. [31] conclude the opposite.

3.7 Object Tracking

Object tracking within computer vision refers to the task of tracking objects across consecutive video frames. Tracking-by-detection is a common approach to multiple object tracking relying on an object detector. A joint tracking and detection model

employs an object detection on the video sequence frames for detection and a tracking algorithm for data association across the frames to obtain the trajectories of the objects. Keypoint tracking and object tracking share the same objectives, and as object trackers, keypoint trackers typically take a two-stage approach [41]. To the best of this author's knowledge, there is no de-facto standard approach for keypoint tracking. Since we are taking an object-based approach to keypoint detection in this thesis, we will do the same for keypoint tracking.

3.7.1 Simple Online Realtime Tracking

Simple Online Realtime Tracking (SORT) [42] is a multiple object tracking framework introduced by Bewley et al. in 2016. SORT relies on an object detector, a state estimation model, and a data association algorithm. The estimation model is a Kalman filter that estimates the next position of an object by extrapolating the motion of the object. SORT implements the Hungarian optimization algorithm [43] to associate objects across frames. The Hungarian algorithm operates on an assignment cost matrix of the IOU distances among the detected objects in frame t and previously detected objects in frame $t-1$.

3.8 Evaluation Metrics

This section presents different metrics for evaluation used in this thesis. The evaluation metrics are based upon the binary prediction results presented in the confusion matrix in Figure 3.9.

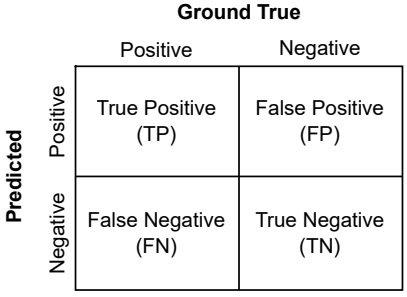


Figure 3.9: The confusion matrix for the definition of the classification metrics. The columns correspond to the true labels and the rows the predicted labels. TP corresponds to a hit (a correct classified positive), TN to a correct rejection (a correct classified negative), while FP is a false alarm (incorrectly positive-classified), and FN a miss (incorrectly negative-classified).

3.8.1 Precision and Recall

Precision and Recall are key metrics for the evaluation of binary classification models. The Precision score (P) is defined as the fraction of true positives and actual

positives:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}. \tag{3.2}$$

The Recall score (R) corresponds to the true positive rate:

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{3.3}$$

According to Equations 3.2 and 3.3, a high Precision score corresponds to a low FP rate, and a high Recall score corresponds to a low FN rate. A model should preferably achieve both. The trade-off between the Precision and Recall is usually visualized with a precision-recall curve from which an optimal threshold can be determined.

3.8.2 Average Precision

The Average Precision (AP) is an accuracy measurement commonly used for benchmarking keypoint estimators. The AP corresponds to the area under the precision-recall curve, which can be calculated by summarizing the weighted mean of the Precision and Recall scores at each threshold n,

$$\mathrm{AP} = \sum_n (\mathrm{R}_n - \mathrm{R}_{n-1})\mathrm{P}_n.$$

A high-performance model should ideally achieve a high AP score. There are different adaptations of the AP score. The AP score is calculated separately for each class, i.e., the classes are averaged independently. Another variant is the mean Average Precision (mAP) which corresponds to the average AP over all classes [27].

3.8.3 Multiple Object Tracking Accuracy

Multiple Object Tracking Accuracy (MOTA) is an evaluation metric for multiple object tracking algorithms. The MOTA metric incorporates three error rates; the ratio of misses, the ratio of false alarms, and the ratio of ID mismatches [44]. An ID mismatch is equivalent to an ID switch and is denoted as IDSW. The MOTA is defined as

$$\mathrm{MOTA} = 1 - \frac{\sum_t \mathrm{FN}_t + \mathrm{FP}_t + \mathrm{IDSW}_t}{\sum_t \mathrm{GT}_t} \tag{3.4}$$

where error rates and the number of true objects GT are calculated per frame t [45]. A high-performance tracker should obtain a high MOTA score.

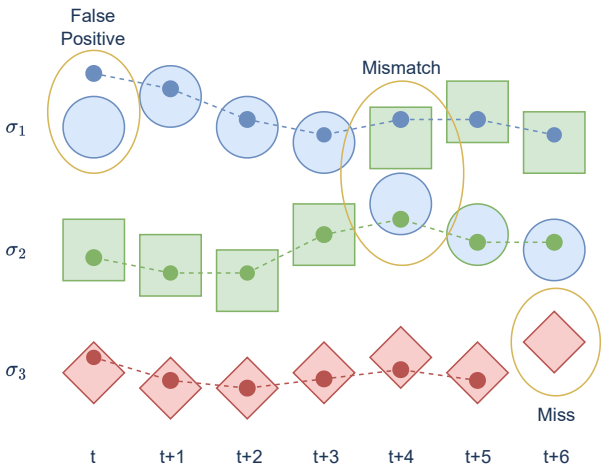


Figure 3.10: Illustration of the different MOTA labels. The different shapes represent objects captured over seven frames where σ_i are the corresponding ID. The False Positive corresponds to a detection where there is no object, the mismatch occurs when the ID switches object, and the miss is when an object goes undetected.

4
Method

This chapter presents the methodology behind experiments conducted to answer the research question *"To what extent is it possible to detect and track regions of interest for remote measurement of vital parameters in RGB and thermal footage and in the presence of characteristic low-altitude UAV motion blur?"*. The experiments consist of six use cases described in Section 4.6. Due to the lack of public available keypoint annotated data collected by UAVs, synthetic motion blur have been applied to still images to simulate UAV motion characteristics focusing on the irregular motion pattern of a UAV. The datasets implemented for keypoint estimation are presented in Section 4.1 while the blur augmentation is described in Section 4.3.

4.1 Keypoint Estimation Datasets

This section presents the datasets used for the keypoint detection task and the alterations made for customizing the mouth keypoint. The datasets are summarized in Table 4.1. The forehead and chest are not included in the standard keypoint annotations and have been triangulated from the existing annotations. The triangulation process is further described in Section 4.4.

Table 4.1: Technical information about the datasets used for the detection task. *Keypoints* refers to the number of annotated keypoint labels.

Dataset	Modality	Keypoints (Total/Facial)	Number of Images		
			Train	Val	Test
COCO [46]	RGB	17/5	118 287	5000	40 671
COCO-WholeBody [47]	RGB	113/68	118 287	5000	40 671
TFW outdoor [48]	Thermal	5/5	5916	664	1600
TFW indoor [48]	RGB+Thermal	5/5	7200	864	2160

4.1.1 The Common Objects in Context Dataset

The original KAPAO model [26] is trained on the Microsoft Common Objects in Context (COCO) dataset [46]; an established large-scale object recognition dataset of everyday objects in the wild. COCO is an RGB dataset annotated with 17 human keypoints. A selection of images from the COCO dataset is presented in Figure 4.1, while the COCO keypoints are shown in Figure 3.6. COCO is annotated with five

facial keypoints: eyes, ears, and nose. There is an extension of the COCO dataset, which is the COCO-WholeBody dataset [47]. The data remains the same, but COCO-WholeBody is annotated with an additional 133 human keypoints, of which 68 are facial keypoints. The additional facial landmarks of COCO-WholeBody are presented in Figure 4.2.

Figure 4.1: Sample images from the COCO dataset.

In this thesis, a custom dataset has been created by adding a mouth keypoint to the original COCO dataset. This mouth keypoint has been derived from the COCO-WholeBody dataset. The COCO keypoints are annotated in standard COCO format, where each keypoint is given by a pixel location, x- and y-coordinate, along with a visibility criterion. The COCO keypoint visibility flags are; $v = 0$: not labeled (in which case $x = y = 0$), $v = 1$: labeled but occluded, $v = 2$: labeled and visible [49]. COCO-WholeBody incorporates other visibility flags representing a reliability criterion that can either be True or False. Compared to COCO, COCO-WholeBody does not differ between occluded and non-labeled scenarios, and keypoints with visibility $v > 0$ are considered reliable [47]. In this thesis, the additional mouth keypoint has been derived by averaging the COCO-WholeBody mouth keypoints with a non-zero visibility criterion. The extracted mouth keypoints have been given visibility $v = 2$ to match the standard COCO format.

In addition to the keypoint labels, COCO includes a skeleton label to relate/connect the keypoint and to create a spatial understanding. The original COCO keypoint skeleton is visualized in Figure 4.3. One should note that this skeleton does not directly translate into the KAPAO pose object. In this thesis, the additional mouth keypoint has been connected directly to the nose to create the custom skeleton.

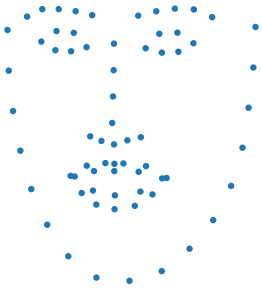


Figure 4.2: The additional 68 COCO-WholeBody facial keypoints.

Figure 4.3: The standard COCO skeleton.

4.1.2 Thermal Faces in the Wild

Thermal Faces in the Wild (TFW) [48] published in 2022 is a thermal dataset annotated with facial keypoint. TFW contains data collected in controlled indoor, semi-controlled indoor, and uncontrolled outdoor settings. The outdoor data are multi-person scenarios collected in different environments under unconstrained settings and are hence the most representative of a potential UAV triage application. Two images from the TFW dataset collected under uncontrolled outdoor settings are presented in Figure 4.4 for exemplification. The TFW dataset has been manually annotated with a facial bounding box and five facial keypoints; eyes, nose, and the outer corners of the mouth. Ultimately, these facial keypoints have created the underlying constraints for the triangulation of the forehead in this project.
The TFW dataset is annotated with two mouth keypoints. In this thesis, these have been averaged to obtain a single mouth keypoint annotation. Moreover, all

Figure 4.4: Samples from the TFW dataset.

the annotations have been considered visible as nothing else is stated and have been assigned the visibility flag $v = 2$ according to the standard COCO keypoint formation.

4.2 Video Datasets for Tracking

This section presents the datasets used to evaluate the tracking algorithm: the 300 Videos in the Wild [50, 51, 52] dataset and the RGBT234 [53] dataset. The datasets are summarized in Table 4.2. To this point, there is no public available in-the-wild RGB video dataset annotated with all keypoint of interest in this thesis. Therefore, tracking of the chest has been left for future work. There is also a lack of publicly available thermal keypoint annotated video dataset; hence, a multimodal dataset has been considered.

Table 4.2: Summary of the datasets used for evaluation of the tracking algorithms.

Dataset	Modality	Number of Sequences	Keypoint Annotated	Moving Camera	Occulsion
300-VW [50, 51, 52]	RGB	114	✓		✓
RGBT234 [53]	RGB+Thermal	234		✓	✓

4.2.1 300 Videos in the Wild

The iBug 300 Videos in the Wild (300-VW) dataset [50, 51, 52] has been used to evaluate the tracking of ROIs in the visual domain. 300-VW is an RGB video dataset for facial landmark tracking annotated with the 68 facial keypoint presented in Figure 4.2. The dataset consists of video sequences acquired under uncontrolled

settings and includes various poses, facial expressions, illumination settings, and occlusion. Moreover, the data is divided into three categories based on difficulty. The categories are *well-lit*, *mild occlusion*, and *challenging* and the videos of category *challenging* have been used for evaluation in this thesis.

Figure 4.5: A sample frame from the iBug 300-VW dataset.

4.2.2 RGBT234 Dataset

The RGBT234 dataset [53] has been used to evaluate the tracking of ROIs in the thermal domain. RGBT234 is a bimodal multi-person video dataset of aligned RGB and thermal video pairs collected under uncontrolled conditions. A frame pair from the dataset can be seen in Figure 4.6. The dataset consists of a wide range of videos in terms of settings, and the most representative sequence have been selected for the task at hand. However, the RGBT234 dataset is not keypoint annotated and has hence been annotated using the bimodality in this project. The thermal video has been annotated by applying the RGB keypoint estimator to the corresponding RGB video. This approach is inspired by Chen et al. [54].

Figure 4.6: An RGB and thermal frame pair from the RGBT234 dataset.

4.3 Motion Blur Augmentation

As a pre-processing step, the data have been augmented with synthetic blur to simulate random motion blur. This section presents the process of generating random motion blur kernels implemented in this project. The approach follows previous work by Boracchi and Foi [40].
The blur kernels have been created using a Markov process followed by sub-pixel linear interpolation. A Markov process is a stochastic process where the next state is determined based on only the current state of the system. The motion trajectories have been continuously sampled on a 2-dimensional grid, where the following position is determined based on the velocity and previous position. The algorithm is described in further detail in Algorithm 1. Three perturbations govern the process: a Gaussian, an impulsive, and an inertial perturbation. The Gaussian perturbation corresponds to a smaller Gaussian deviation, while the inertial term is a larger deviation. In a UAV application, the Gaussian term could be caused by hovering and the inertial perturbation due to wind gusts. The impulsive perturbation is a counter term that counteracts the other perturbations [38, 40]. If the perturbations equal zero, then the motion will be linear.

Algorithm 1 Random trajectory generator
Parameters:
M-number of iterations,
L_{max}-max length of trajectory,
I-inertia,
p_s-probability of an impulsive perturbation,
p_b-probability of an inertial perturbation,
p_g-probability of a Gaussian perturbation,
ϕ-initial angle,
x-the trajectory vector.

$v_0 \leftarrow \cos\phi + i\sin\phi$
$v \leftarrow v_0 \cdot L_{max}/(M-1)$
for t=1 to M-1 **do**
  if randn $< p_b \cdot p_s$ **then** ▷ randn$\sim \mathcal{N}(\mu, \sigma^2)$
    nextDirection $\leftarrow 2v \cdot e^{i(\pi(\text{randn}-0.5))}$
  else
    nextDirection $\leftarrow 0$
  end if
  $dv \leftarrow$ nextDirection$+p_s(p_g(\text{randn}+i\text{randn}) \cdot I \cdot x[t] \cdot L_{max}/(M-1))$
  $v \leftarrow v + dv$
  $v \leftarrow (v/|v|) \cdot L_{max}/(M-1)$
  $x[t+1] \leftarrow x[t] + v$
end for

The obtained motion trajectories have been converted into point spread functions (PSFs) by sub-pixel linear interpolation. Sub-pixel linear interpolation is a sampling method for transforming from sub-pixel resolution to pixel resolution by linear inter-

polation. The PSFs have been generated by sampling the trajectories on a pixel grid and performing linear interpolation along each axis. A selection of obtained PSFs are presented in Figure 4.7 for exemplification. The data was blurred by convolving the PSFs onto the images using openCV [13] and the blurred output created with the PSFs of Figure 4.7 are presented in Figure 4.8.

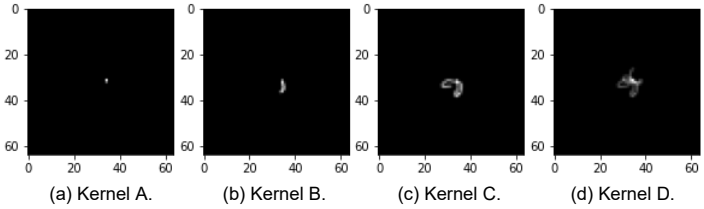


Figure 4.7: Different random motion blur kernels with different standard deviations.

4.4 Triangulation of the Forehead and Chest

The additional ROIs of this thesis that are not annotated have been triangulated from the existing keypoint annotations as a post-processing step. The forehead has been triangulated from the eyes and nose for all use cases. Given an x-axis along eye level and a y-axis intersecting the nose at $(0, \text{-}d_{nose})$, the forehead keypoint has been located at coordinate $(0, 0.5d_{nose})$. The chest keypoint has been triangulated from the hips and the shoulders. If the shoulders and the hips constitute a square, then the chest has been defined in the center of the upper half of the square. Triangulation has the drawback of breaking in the absence of a keypoints the triangulation relies upon. Hence, one should be aware that triangulated ROIs are more sensitive to occlusion than their independent counterparts. Moreover, one should note that the distances for triangulation are not chosen on specific scientific grounds for vital parameter measurement in this project.

4.5 Network Architecture

This section presents the architecture and training procedures of the implemented keypoint estimator. The small KAPAO version has been implemented in this thesis project which relies on a CSPDarkNet53 backbone for feature extraction. The DarkNet53 is a CNN architecture of 53 layers comprising 3×3 and 1×1 kernels. To enhance the learning capabilities of the CNN, CSPDarkNet53 employs a Cross Stage Partial Network (CSPNet) strategy in which the gradient flow is divided and propagated through the networks in different paths. It is achieved by dividing the feature map of the base layer into two and by fusing them through a cross-stage hierarchy [55]. The architecture is further described in Table 4.3.

(a) Convolved with kernel A.

(b) Convolved with kernel B.

(c) Convolved with kernel C.

(d) Convolved with kernel D.

Figure 4.8: The blurred result obtained when applying the blur kernels in Figure 4.7.

Table 4.3: CSPDarknet53 architecture, the backbone to YOLOv5s version 6.0.

Type of Layer	Filter	Size	Repetitions
Conv	64	3×3	1×
Conv	128	3×3	1×
C3	128	1×1	3×
Conv	256	3×3	1×
C3	256	1×1	6×
Conv	512	3×3	1×
C3	512	1×1	9×
Conv	1024	3×3	1×
C3	1024	1×1	3×
SPPF	1024	5×5	1×

Following the approach of McNally et al. [26], the use cases not relying on fine-tuning have been initialized with the weights of YOLOv5s to decrease the training time. According to the default setting of YOLOv5s, also employed by McNally et al., a stochastic gradient descent (SGD) optimizer and an initial learning rate of 0.01 have been used for training. For the use cases not relying on fine-tuning, the training has been performed during 250 epochs using batch size 32. The model fine-tuned on thermal data has been initialized with the weights of the RGB model, and the

learning rate decreased to 0.001. Moreover, the number of epochs has been set to 50[1].
The data has been augmented during the training process to expand the dataset artificially. Once more, we have followed YOLOv5s and McNally et al. [26] and used the augmentation techniques and corresponding probabilities in Table 4.4. The HSV techniques alter the color and mosaic creates new images by combining multiple images.

Table 4.4: Augmentation techniques and corresponding probabilities were implemented for the expansion of the datasets.

Method	Probability
HSV-Hue	0.015
HSV-Saturation	0.7
HSV-Value	0.4
Translate	0.1
Scale	0.9
Flip left-right	0.5
Mosaic	1.0

4.6 Experimental Setup

This section describes the experiments conducted to answer the research question. The experiments conducted for the detection and tracking tasks are described separately below.

4.6.1 Detection

The different scenarios for keypoint detection that have been taken under consideration are presented in Table 4.5. An RGB and thermal baseline have been created for reference purposes, trained on the modified COCO dataset and TFW outdoor data respectively.
As the impact of characteristic UVA motion blur is of interest in this project, an RGB model adapted to motion blur have been created. This has been achieved by adversarial training on a blurred version of the COCO dataset, where 1/3 of the training data is perturbed with synthetic motion blur as described in Section 4.3. The impact of the blur has been evaluated by comparing the RGB baseline and the adopted model's performance in the presence of synthetic blur.
Three additional use cases have been considered to further investigate the properties of the thermal domain and KAPAO. As the TFW outdoor data is a relatively small dataset, an expanded model has been introduced trained on the combined TFW outdoor and indoor data. The temperature ranges of an outdoor versus indoor

[1]The readers unfamiliar with the fundamental concepts of training neural networks are referred to [56] for further reading.

Table 4.5: The use cases for which the experiments have been conducted.

Use Case	Training Data	Evaluation Data
RGB baseline	COCO	COCO
		COCO blurred
		TFW outdoor
RGB blurred	COCO blurred	COCO blurred
Thermal baseline	TFW outdoor	TFW outdoor
RGB fine-tuned on thermal	COCO+TFW outdoor	TFW outdoor
Thermal expanded	TFW outdoor+TFW indoor	TFW outdoor
Thermal with modified pose object	TFW outdoor	TFW outdoor

environment can largely differentiate, and it is not trivial that combining them will increase performance. Another approach taken under consideration is fine-tuning from the visual to the thermal domain to compensate for the limited amount of available thermal keypoint annotated data. The last model adopts an alternative pose object. Rather than implementing the modified COCO skeleton, the face grid illustrated in Figure 4.9 have been adopted. The aim of redefining the pose object is to better understand the modeling of local and global features in the thermal domain. Like the thermal baseline, all the additional thermal models have been evaluated on the TFW outdoor dataset for comparison.

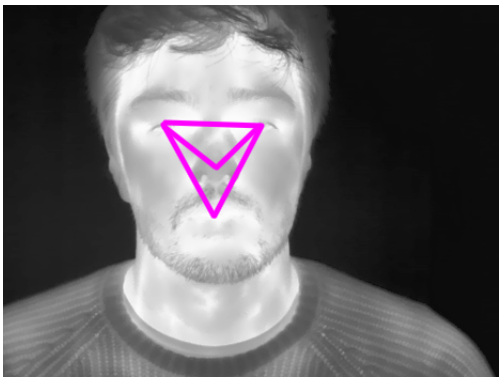


Figure 4.9: Illustration of the facial grid used as a pose object. The image is taken from the TFW indoor dataset.

4.6.2 Tracking

For the tracking of ROIs in the visual and thermal domains, tracking algorithm SORT has been implemented according to the theory of Section 3.7. The best-performing keypoint estimators of the respective domain have been used for detection of ROIs. In addition to the SORT algorithm, a naive tracking algorithm where the tracking IDs are assigned based on frame-to-frame differences has been employed for reference purpose. The experiments conducted referring to tracking are listed in Table 4.6.

Table 4.6: The use cases for the tracking task.

Use Case	Keypoint Estimator	Evaluation Data
Naive algorithm	RGB	300-VW challenging
	RGB	RGBT234 video sequence
	Thermal	RGBT234 video sequence
SORT	RGB	300-VW challenging
	RGB	RGBT234 video sequence
	Thermal	RGBT234 video sequence

As previously mentioned, the thermal video dataset is not keypoint annotated. Hence, the chosen RGBT-234 video has been annotated with the RGB baseline. The RGB keypoint estimator has been applied to the RGB modality and transferred onto the thermal modality to evaluate the performance.

4.7 Evaluation

The metrics described in Section 3.8 have been implemented according to the standard COCO metrics for keypoint detection. The COCO evaluation metric for keypoint detection relies on an object keypoint similarity (OKS) measurement defined as,

$$\text{OKS} = \sum_i \exp\left[\frac{-d_i^2}{2s^2\kappa_i^2}\right] \delta\left(v_i > 0\right) \Big/ \sum_i \delta\left(v_i > 0\right). \tag{4.1}$$

The OKS represents the average keypoint similarity across the keypoint labels where d_i corresponds to the Euclidean distance between a detected keypoint and corresponding ground-true annotation, and v_i the annotated keypoints visibility. Note that the visibility of the detected keypoint is not taken under consideration. The $s\kappa_i$ corresponds to the keypoint standard deviation where s denotes the object scale, and κ_i is a keypoint constant. The object scale is defined as the square root of the segmented object area and the keypoint constant $\kappa_i = 2\sigma_i$ where σ_i corresponds to the keypoint standard deviation for the object scale. By the COCO standard σ_i for human keypoint detection are given by [0.026, 0.025, 0.035, 0.079, 0.072, 0.062, 0.107, 0.087, 0.089] for the nose, eyes, ears, shoulders, elbows, wrists, hips, knees, and ankles, respectively. These values have been derived from the COCO evaluation dataset [49]. In this project, the additional mouth keypoint has been assigned the same σ_i as the nose.
The primary evaluation metric for keypoint detection is mAP at OKS = 0.50 : 0.05 : 0.95 and OKS = 0.50. An OKS threshold determines the AP for keypoint evaluation. Keypoints with an OKS exceeding the threshold are considered TP, and vice versa for FP, from which the AP can be determined. According to the standard COCO metrics for keypoint detection the mAP at OKS=0.50:0.05:0.95 corresponds the mean mAP over OKS=0.50, 0.55, 0.60, ..., 0.95 [49]. From here on, mAP at OKS=0.50:0.05:0.95 will be denoted as mAP and mAP at OKS=0.50 as $\text{mAP}^{.50}$.

In addition to the COCO standard metric for keypoint detection, the PoseTrack evaluation metric has been considered. PoseTrack is a benchmark for human pose estimation and tracking comprising of three tasks, 1) single-frame pose estimation, 2) pose estimation in videos, and 3) pose tracking in the wild. A geodesic point similarity (GSP) has been adopted for the estimation in the thermal domain. The GSP mimics the COCO OKS in Equation 4.1, but rather than relying on COCO-specific normalization instances, a mean geodesic distance is used as a normalization factor [57]. The ground truth eye-eye distance is commonly used as a normalization constant in facial landmark detectors [58]. However, the ground truth eye-nose distance has been used for normalization in this project as it is less sensitive to occlusion and different poses.
For tracking, PoseTrack employs the MOT metric described in Section 3.8. The PoseTrack evaluation server has been implemented to evaluate the tracking of keypoint. The server reports the MOTA score for each keypoint label and as an average over all keypoint labels [57].

5
Results

This chapter presents the results of the conducted experiments presented in Chapter 4. The results referring to detection are presented in Sections 5.1 and 5.2, while the tracking results are shown in Section 5.3.

5.1 Blur Impact

This section presents the impact of the synthetic motion blur and the adversarial training in the visual domain. The performance of the RGB models, previously described in Section 4.6, on different blurred versions of the COCO evaluation dataset is presented in Table 5.1. Note that the blur percentage reports the percentage of evaluation images augmented with synthetic blur. The baseline model achieved a mAP=0.619 and $mAP^{.50}$=0.8 on the original COCO evaluation dataset.

Table 5.1: The performances of the RGB baseline and the model trained on synthetic blur on the different blurred versions of the COCO evaluation dataset. The *Blur Percentage* reports the percentage of images augmented with artificial blur. The scores have been calculated according to the COCO standard metric for keypoint evaluation, where the additional mouth keypoint has been assigned the same keypoint standard deviation as the nose.

Use Case	Blur Percentage	mAP@[OKS=0.50:0.05:0.95]	mAP@[OKS=0.50]
RGB baseline	0	0.62	0.80
	33	0.53	0.77
	100	0.34	0.62
RGB blurred	33	0.53	0.79
	100	0.36	0.67

5.2 Detection in the Thermal Domain

This section is dedicated to the results regarding detection of the ROIs in the thermal domain. The models' performance on the thermal TFW outdoor dataset is presented in Table 5.2. As shown in Table 5.2, the model fine-tuned from visual to thermal domain yields the highest scores while the RGB baseline achieved the lowest mAP. However, in terms of $mAP^{.50}$, the RGB baseline obtained a significantly high

score compared with the acquired mAP. One should note that the reported scores in Table 5.2 have been determined using the geodesic normalization described in Section 4.7.

Table 5.2: The detection scores for the different models were determined on the TFW outdoor evaluation dataset. The scores have been calculated with the eye-nose distance as a normalization factor.

User Case	mAP@[OKS=0.50:0.05:0.95]	mAP@[OKS=0.50]
RGB baseline	0.21	0.73
Thermal baseline	0.52	0.54
RGB fine tuned on thermal	0.86	0.89
Thermal expanded	0.64	0.66
Thermal with modified pose object	0.73	0.77

5.3 Tracking of ROIs

The results obtained by the naive algorithm employed for reference and the SORT algorithm are presented in Tables 5.3 and 5.4. The tracking algorithms have been evaluated according to the PoseTrack standard on the subset of the 300-VW dataset labeled *challenging*, and a selected sequence from the RGBT234. In addition to the ROIs for vital parameter measurement, the MOTA scores for the eyes used for triangulation of the forehead have been reported. According to equation 3.4, the MOTA score ranges from negative infinity to one. A high performing tracking algorithm should achieve a MOTA score greater than zero implying that the number of true observations exceeds the total number of misses, false alarms, and ID mismatches. One can observe that the algorithms achieve a nearly perfect score on the 300-VW dataset, compared to the scores on the RGBT234 dataset, which are significantly lower. Moreover, the result shows that the implementation of the SORT algorithm increased the performance.

Table 5.3: The obtained MOTA scores for the naive tracking algorithm. The MOTA is reported per keypoint, and the *Total* corresponds to the average score over the keypoint labels.

Use Case	Dataset	Nose	Left Eye	Right Eye	Mouth	Forehead	Total
RGB baseline	300-VW	0.981	0.985	0.990	0.976	0.972	0.976
RGB baseline	RGBT234	-0.956	-0.955	-0.955	-0.957	-1.000	-0.972
Fine-tuned on TFW	RGBT234	-0.674	-0.674	-0.674	-0.694	-0.998	-0.743

Table 5.4: The obtained MOTA scores for the SORT algorithm. The MOTA is reported per keypoint, and the *Total* corresponds to the average score over the keypoint labels.

Use Case	Dataset	Nose	Left Eye	Right Eye	Mouth	Forehead	Total
RGB baseline	300-VW	0.989	0.991	0.991	0.988	0.989	0.989
RGB baseline	RGBT234	-0.837	-0.838	-0.839	-0.885	-0.893	-0.858
Fine-tuned on TFW	RGBT234	-0.441	-0.434	-0.444	-0.501	-0.550	-0.474

6
Discussion

In this chapter, the results in Chapter 5 are discussed in relation to the research question. In addition, the adaption of this thesis for a potential UAV triage application is discussed as well as the ethical concern focusing on dataset biases.

6.1 Detection in the Visual Domain

The RGB baseline can be compared with the state-of-the-art KAPAO model to gain some understanding of the performance. The RGB baseline obtains a mAP=0.619 and $mAP^{.50}$=0.800 which is similar accuracy to the original small KAPAO model which achieves a mAP=0.638 and $mAP^{.50}$=0.884. Training of RGB baseline is computationally demanding and requires a considerable amount of time. Hence the number of training epochs has not been further investigated in this project, and one could probably achieve higher AP scores by prolonging the training. Moreover, only the small KAPAO version has been the object of investigation in this project, as inference speed has been prioritized. One should note that implementing the larger version would likely increase the accuracy as the original large KAPAO model achieves a mAP=0.703.

As expected, the results in Table 5.1 demonstrate a decrease in performance in the presence of motion blur. The result shows that adversarial training can improve performance, although the difference is marginal compared to the result by Vasiljevic et al. [33]. Vasiljevic et al. regain moste of the lost accuracy by fine-tuning their pre-trained model on synthetic blur. One should note that there are two significant differences compared to the work of Vasiljevic et al.: 1) they are considering conventional objects rather than keypoint objects, and 2) the implementation of blur kernels. Firstly, keypoint objects are smaller than conventional objects, and small objects are more sensitive to blur [59]. Hence the object size could explain the results. Secondly, Vasiljevic et al. generate 100 different blur kernels, which are applied randomly rather than applying truly randomized blur kernels as in this project. This can explain why generalization among kernels cannot be observed to the same extent.

As adversarial training appears to not be very effective in this case, it might be interesting to investigate possible image prepossessing alternatives for deblurring. In a UAV application, that could be optical stabilizing measures. Another possible option could be to remove blur artifacts by deconvolution of the kernels with a deblurring generative adversarial network (GAN).

6.2 Detection in the Thermal Domain

The results in Table 5.2 conclude that the RGB model fine-tuned on thermal data achieves the highest scores of mAP=0.856 and $mAP^{.50}$=0.890 on the TFW outdoor dataset. The performance of the fine-tuned model provides evidence for that it is possible to generalize from the visual to the thermal domain. However, the performance of the RGB baseline is inconclusive. The RGB baseline achieves the lowest mAP=0.205, but the $mAP^{.50}$=0.728 which is considerably higher. Deviation in annotation standards between the datasets could explain the differing AP score of the RGB baseline.

Observing the two additional thermal models, the result in Table 5.2 shows that the performance increased by including the indoor data for training and implementing the face grid. The relatively poor performance of the thermal baseline could reflect upon the size of the dataset but also the thermal feature space. Firstly, the increase in performance due to the expansion of training data is not trivial as the outdoor and indoor settings differ considerably in the thermal domain. It would be interesting to investigate if this is a model-specific behavior, possibly enabled by the pose object, or not. Secondly, McNally et al. [26] state that the keypoint objects are intended for keypoint with local features and pose objects for keypoints with global features. As facial keypoints are categorized as keypoints with local features, it is interesting that the modification of the pose objects has such a big impact on the performance. It would be interesting to investigate this behavior further as it speaks for the innovative height of the pose object.

The TFW dataset is a relatively newly published dataset; hence, there is limited published work regarding the dataset. The authors of the dataset [48] provide two baselines, a YOLOv5 and a YOLOv5Face model, both trained on their dataset. Kuzdeuov et al. use a normalized mean error for evaluation. However, it is not stated how they have handled false detections, making it impossible to compare the results.

6.3 Dataset Biases

To understand the limitations of a deep learning model, one has to be aware of dataset biases. Amplification of biases within the training dataset is a well-documented problem within computer vision [60]. For example, a deep-learning model can be biased in terms of gender, ethnicity, and age, which are ethical concerns.

Studies have been conducted to investigate biases within the COCO dataset. Zhao et al. [61] conclude that the COCO dataset is biased toward skin color and gender at image and instance levels. According to their study, the COCO dataset contains 7.5 times more light-skinned subjects than dark-skinned subjects and two times more males than females. Moreover, Zhao et al. conclude that there are visual differences and that darker-skinned subjects more frequently appear in outdoor settings whiles light-skinned subjects more regularly appear in indoor environments. Although COCO is a large-scale dataset, it possesses biases that could cause ethical concerns in a real-world application.

Biases within the thermal domain are more complex compared to biases in the visual domain. For example, one should be aware that the heat signatures of a red house made of wood and a red brick house may differ due to the different thermal properties of the materials. The houses can appear the same on a cloudy day, but their heat signatures will differ in direct sunlight. A thermal dataset can, for example, be biased in terms of lighting conditions, geographic locations, weather conditions, time of day or year of capture, and due to material properties. Hence, the size and diversity of the thermal data are essential. Biases within the TFW dataset have not been investigated, but one can assume it possesses biases due to its small size.
Analyzing bias propagation becomes especially important in a potential health care application to avoid discrimination. In addition, biases regarding health status have to be taken under consideration in a medical application. The data used in this project are limited to healthy subjects, and the models are consequently biased towards healthy subjects. Scenarios likely to occur in a mass casualty incident such as bruising, blood, and loss of limbs have not been investigated. The same applies in the thermal domain, as skin temperature is correlated to health status. In a potential mass casualty incident application, domain-specific features such as those mentioned must be investigated to ensure unbiased models.

6.4 Tracking of ROIs

The tracking of ROIs has been evaluated on two different datasets in terms of image quality and distance to the observed subjects. The difference in distance likely explains the difference in MOTA scores as the distance directly relates to the pixel area of the ROIs. Due to data shortage, further quantifying the distance dependency has been left to future work.

6.4.1 Evaluation on the 300-VW Dataset

The results in Tables 5.3 and 5.4 show that both the naive and the SORT algorithm achieve high accuracy on the 300-VW dataset, although only evaluated on the videos of category *challenging*. The forehead and mouth achieve the lowest MOTA scores, likely due to triangulation and lip movements. Hence, one can conclude that it is favorable to use the nose for triangulation rather than triangulate from the mouth. The result shows that implementing the SORT algorithm improves the tracking score, although the scores were nearly perfect to start with.
An aspect to consider is the relative size and motions of the ROIs. The 300-VW dataset mainly consists of videos where the ROIs cover larger pixel areas, which is favorable from a tracking perspective. Nevertheless, the remaining question is what is a suitable distance for remote vital parameter measurement. Related research on remote vital parameter measurement by Yang et al. [14] report a method that performs well on a $0.6 - 1.2m$ distance in indoor settings, which according to Yang et al. is a considerably large distance compared to previous works. If the distance is realizable in a UAV application is debatable; however, the range falls within distances featured in the 300-VW dataset.

6.4.2 Evaluation on the RGBT234 Dataset

The evaluation on the RGBT345 dataset does not achieve as high MOTA scores as the results obtained on the 300-VW dataset (see Table 5.3 and 5.4). Implementing the SORT algorithm improves the scores for all use cases. However, the large distances and the small relative motions of the ROIs are presumably limiting the performance of the algorithm. Based on visual observations, detection failures and identification failings likely cause the low MOTA scores. This behavior does not seem specific to the thermal domain, as the same could be observed in the RGB modality.
The performance of a tracking algorithm is ultimately determined by the performance of the detector. Small objects with weak appearance and features decrease the performance of object detectors [59] which is likely what we observed here. As the ROIs cover small pixel areas, it is presumable that the detector is sensitive to larger distances and that the combination of larger distances and poor image quality becomes especially problematic. One could assumably increase the performance by shortening the detection distances, but there are other approaches to improve the performance on small objects. A prominent solution is YOLO-Z, introduced by Benjumea et al. [62]. YOLO-Z is a YOLOv5 model modified on an architectural level, improving the detection abilities of small objects without any significant speed trade-off.

6.5 Adaption to UAV Applications

This project focuses on characteristic UAV motion blur. However, for a potential UAV implementation, additional aspects concerning the UAV domain have to be considered. The bird's-eye view and the pitch angle of the UAV are particular features to be taken into account.
Another aspect of bringing attention to is hardware limitations. Challenges for deploying a computer vision application on a UAV platform are: 1) the energy consumption, the application has to consume minimal power to minimize the impact on the UAV flight time, 2) the memory consumption and computational power as the UAV payload capability are limited, and 3) the data processing, the input data has to be processed with low latency to be applicable for real-time applications [20]. The detection and tracking of ROIs will only be a part of a large UAV triage application, making these aspects even more relevant.

7
Conclusions

The objective of this thesis has been to investigate the possibilities for a UAV to detect and track ROIs for remote measurement of vital parameters in the visual and thermal domains. The ROIs have been the forehead, nose, mouth, and chest, and the UAV characteristic taken under consideration is motion blur due to random camera motion. In this project, we have taken an object detection approach to keypoint estimation and tracking. The state-of-the-art keypoint detector KAPAO and the tracking algorithm SORT have been implemented and evaluated in several experimental setups. Various metrics have been used to assess the performance of the keypoint estimators and tracking algorithms. The keypoint estimators have been evaluated using mAP and different keypoint object similarity measurements. For the evaluation of the tracking performance, MOTA has been used.
For detecting ROIs in the thermal domain, the model created by transferring knowledge from the visual to the thermal domain by fine-tuning showed the highest performance. Furthermore, as the second best performing model in the thermal domain, the expansion of the pose object improved the performance significantly. This result demonstrates the innovative use of the spatial information.
Adversarial training on motion blur had a minimal impact on the performance in the presence of blur. The result problematizes the sizes of the ROIs and the motion characteristics of low-altitude UAV flights. Since no generalization among random blur kernels could be observed, the results support the use of optical stabilization in a possible UAV triage application.
Regarding tracking of ROIs, the result concludes that the SORT algorithm improved the performance for all use cases. In addition, one can observe that both the SORT algorithm and the naive approach achieved an almost perfect score on the 300-VW dataset. This is a significant result as the dataset represent the distances of previous research on remote measure of vital parameters. The result shows that the pixel area is relevant, and it can be concluded that the distance and image quality impact the performance. This poses a potential limitation in a UAV triage application.

7.1 Future Work

In this project, we have been using public available keypoint annotated datasets to avoid manual annotation of data. The UAV motion characteristics have been simulated, and further adopting the models for UAV triage application has been left for future work. Challenges to be met have previously been discussed in Sections 6.3 and 6.5. With respect to the UAV triage characteristic, possible directions for future

work could be to 1) collect and manually annotate UAV data from a mass casualty incident or 2) synthesize data and simulate corresponding scenarios. An idea could be to create synthetic data by segmenting and adding humans to UAV footage.
The research question is related to limitations associated with remote measurement of vital parameters for triage. It would be beneficial to further investigate to what extent it is possible to remotely measure vital parameters to define constraints regarding this project in terms of accuracy, distances, and angles. Such limitations would put this project in context and make it possible to tune the evaluation metrics accordingly.
This project aims to detect and track the ROIs in the visual and thermal domains. It would be interesting to investigate the properties of the thermal domain further. A possible direction of future studies could be to investigate the thermal dataset biases and implications of using apparent temperature differences for imaging. As there is a limited public available keypoint annotated thermal data, another direction of future studies could be to explore the possibility of transforming RGB images into thermal images using, for example, GANs.
The tracking algorithm SORT has been implemented in this thesis project. However, other possible tracking algorithms might be of interest. DeepSORT [63] is a successor to SORT which additionally employs a re-identification network. The re-identification network is a CNN trained to identify object similarity to reduce identity switches. The original DeepSORT re-identification network is trained on human objects, and a possibility could be to create a re-identification dataset for the ROIs. However, that has been left for future work due to limitations.
Before deploying the models in a real-world application, dataset biases and bias propagation has to be addressed. Aspects to be taken under consideration have been previously mentioned in Section 6.3, which have to be investigated to be addressed accordingly. A possible idea could be to expand the dataset artificially using GANs for image synthesis to increase the diversity of the model and eliminate biased behavior.
Several ethical aspects have to be investigated before deploying the model in a real-world triage application. In a medical application, privacy concerns become particularly important and should hence be analyzed further. Storage of the patents personal data, patient privacy and consent, as well as reconstruction of training data, for example, has to be addressed before a potential deployment.

Bibliography

[1] J. Bazyar, M. Farrokhi, and H. Khankeh, "Triage systems in mass casualty incidents and disasters: A review study with a worldwide approach," *Open Access Macedonian Journal of Medical Sciences (OAMJMS)*, vol. 7, no. 3, p. 482, 2019. DOI: http://dx.doi.org/10.3889/oamjms.2019.119.

[2] J. Rantakokko, M. G. Lozano, G. Tolt, L. Thors, and A. Bucht, "Evakuering av skadade med obemannade farkoster," tech. rep., Totalförsvarets forskningsinstitut (FOI), 2022. ISSN: 1650-1942.

[3] K. Khabarlak and L. Koriashkina, "Fast facial landmark detection and applications: A survey," *arXiv:2101.10808*, 2021.

[4] Etikprövningsmyndigheten, "Värnar människan i forskning." URL:https://etikprovningsmyndigheten.se, accessed: October 25 2022.

[5] Swedish Defence Research Agency (FOI), "About FOI." URL: https://www.foi.se/en/foi/about-foi.html, accessed: Mars 2 2022.

[6] A. Khorram-Manesh, J. Nordling, E. Carlström, K. Goniewicz, R. Faccincani, and F. M. Burkle, "A translational triage research development tool: Standardizing prehospital triage decision-making systems in mass casualty incidents," *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine (SJTREM)*, vol. 29, no. 1, pp. 1–13, 2021. DOI: http://dx.doi.org/10.1186/s13049-021-00932-z.

[7] E. Lee, E. Chen, and C.-Y. Lee, "Meta-rPPG: Remote heart rate estimation using a transductive meta-learner," in *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 392–409, Springer, 2020. DOI: http://dx.doi.org/10.1007/978-3-030-58583-9_24.

[8] F. J. Rodriguez-Lozano, F. León-García, M. Ruiz de Adana, J. M. Palomares, and J. Olivares, "Non-invasive forehead segmentation in thermographic imaging," *Sensors*, vol. 19, no. 19, p. 4096, 2019. DOI: https://doi.org/10.3390/s19194096.

[9] V. Hartmann, H. Liu, F. Chen, W. Hong, S. Hughes, and D. Zheng, "Toward accurate extraction of respiratory frequency from the photoplethysmogram: Effect of measurement site," *Frontiers in Physiology*, vol. 10, p. 732, 2019. DOI: http://dx.doi.org/10.3389/fphys.2019.00732.

[10] D. Müller, A. Ehlen, and B. Valeske, "Convolutional neural networks for semantic segmentation as a tool for multiclass face analysis in thermal infrared," *Journal of Nondestructive Evaluation*, vol. 40, no. 1, pp. 1–10, 2021. DOI: http://dx.doi.org/10.1007/s10921-020-00740-y.

[11] D. Djeldjli, F. Bousefsaf, C. Maaoui, F. Bereksi-Reguig, and A. Pruski, "Remote estimation of pulse wave features related to arterial stiffness and blood pressure using a camera," *Biomedical Signal Processing and Control*, vol. 64, p. 102242, 2021. DOI: http://dx.doi.org/10.1016/j.bspc.2020.102242.

[12] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[13] G. Bradski and A. Kaehler, "OpenCV," *Dr. Dobb's Journal of Software Tools*, vol. 3, p. 2, 2000.

[14] F. Yang, S. He, S. Sadanand, A. Yusuf, and M. Bolic, "Contactless measurement of vital signs using thermal and RGB cameras: A study of COVID 19-related health monitoring," *Sensors*, vol. 22, no. 2, p. 627, 2022. DOI: http://dx.doi.org/10.3390/s22020627.

[15] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot multi-level face localisation in the wild," in *Proceedings of IEEEConference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5203–5212, 2020. DOI: http://dx.doi.org/10.1109/CVPR42600.2020.00525.

[16] S. Skansi, "Feedforward neural networks," in *Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence*, pp. 79–105, Cham: Springer, 2018. DOI: https://doi.org/10.1007/978-3-319-73004-2_4.

[17] H. H. Aghdam and E. J. Heravi, "Convolutional neural networks," in *Guide to Convolutional Neural Networks: A Practical Application to Traffic-Sign Detection and Classification*, pp. 85–130, Cham: Springer, 2017. DOI: https://doi.org/10.1007/978-3-319-57550-6_3.

[18] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv:1511.08458*, 2015.

[19] Y. Pang and J. Cao, *Deep Learning in Object Detection*, pp. 19–57. Singapore: Springer Singapore, 2019. DOI: https://doi.org/10.1007/978-981-10-5152-4_2.

[20] S. Vaddi, *Efficient object detection model for real-time UAV applications*. PhD thesis, Iowa State University, 2019. DOI: http://dx.doi.org/10.5539/cis.v14n1p45.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 779–788, 2016. DOI: http://dx.doi.org/10.1109/CVPR.2016.91.

[22] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv:1804.02767*, 2018.

[23] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 7263–7271, 2017. DOI: http://dx.doi.org/10.1109/CVPR.2017.690.

[24] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *arXiv:2004.10934*, 2020.

[25] U. Nepal and H. Eslamiat, "Comparing YOLOv3, YOLOv4 and YOLOv5 for autonomous landing spot detection in faulty uavs," *Sensors*, vol. 22, no. 2, p. 464, 2022. DOI: http://dx.doi.org/10.3390/s22020464.

[26] W. McNally, K. Vats, A. Wong, and J. McPhee, "Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation," *arXiv:2111.08557*, 2021.

[27] C. Zheng, W. Wu, T. Yang, S. Zhu, C. Chen, R. Liu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A eurvey," *arXiv:2012.13392*, 2020.

[28] National Aeronautics and Space Administration, Science Mission Directorate, "Infrared waves," 2010. URL: http://science.nasa.gov/ems/07_infraredwaves, accessed: May 12 2022.

[29] M. Vollmer, "Infrared thermal imaging," in *Computer Vision: A Reference Guide*, pp. 666–670, Springer, 2021. DOI: http://dx.doi.org/10.1007/978-3-030-63416-2_844.

[30] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020. DOI: http://dx.doi.org/10.1109/JPROC.2020.3004555.

[31] M. Sayed and G. Brostow, "Improved handling of motion blur in online object detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1706–1716, 2021. DOI: http://dx.doi.org/10.1109/CVPR46437.2021.00175.

[32] A. Chaubey, N. Agrawal, K. Barnwal, K. K. Guliani, and P. Mehta, "Universal adversarial perturbations: A survey," *arXiv:2005.08087*, 2020.

[33] I. Vasiljevic, A. Chakrabarti, and G. Shakhnarovich, "Examining the impact of blur on recognition by convolutional networks," *arXiv:1611.05760*, 2016.

[34] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," *arXiv:1805.06605*, 2018.

[35] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," *IEEE Access*, vol. 9, pp. 155161–155196, 2021. DOI: http://dx.doi.org/10.1109/ACCESS.2021.3127960.

[36] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *Proceedings of 2016 8th International Conference on Quality of Multimedia Experience (QoMEX 2016)*, pp. 1–6, 2016. DOI: http://dx.doi.org/10.1109/QoMEX.2016.7498955.

[37] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Understanding and evaluating blind deconvolution algorithms," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1964–1971, 2009. DOI: http://dx.doi.org/10.1109/CVPR.2009.5206815.

[38] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind motion deblurring using conditional adversarial networks," in *Proceedings of IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8183–8192, 2018. DOI: http://dx.doi.org/10.1109/CVPR.2018.00854.

[39] O. Murat, "Blur kernel estimation by using bees algorithm," *International Journal of System Modeling and Simulation (IJSMS)*, vol. 1, no. 2, pp. 8–11, 2016.

[40] G. Boracchi and A. Foi, "Modeling the performance of image restoration from motion blur," *IEEE Transactions on Image Processing (TIP)*, vol. 21, no. 8, pp. 3502–3517, 2012. DOI: http://dx.doi.org/10.1109/TIP.2012.2192126.

[41] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran, "Detect-and-track: Efficient pose estimation in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 350–359, 2018. DOI: http://dx.doi.org/10.1109/CVPR.2018.00044.

[42] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Uproft, "Simple online and realtime tracking," in *IEEE international conference on image processing (ICIP)*, pp. 3464–3468, IEEE, 2016. DOI: http://dx.doi.org/10.1109/ICIP.2016.7533003.

[43] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955. DOI: http://dx.doi.org/10.1002/nav.3800020109.

[44] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the CLEAR MOT metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, pp. 1–10, 2008. DOI: http://dx.doi.org/10.1155/2008/246309.

[45] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler, "MOT16: A benchmark for multi-object tracking," *arXiv:1603.00831*, 2016.

[46] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proceedings of European Conference on Computer Vision (ECCV)*, pp. 740–755, Springer, 2014. DOI: http://dx.doi.org/10.1007/978-3-319-10602-1_48.

[47] S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, and P. Luo, "Whole-body human pose estimation in the wild," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. DOI: http://dx.doi.org/10.1007/978-3-030-58545-7_12.

[48] A. Kuzdeuov, D. Aubakirova, D. Koishigarina, and H. A. Varol, "TFW: Annotated thermal faces in the wild dataset," *IEEE Transactions on Information Forensics and Security*, 2022. DOI: http://dx.doi.org/10.1109/TIFS.2022.3177949.

[49] COCO: Common Objects in Context, "Keypoint evaluation." URL: https://cocodataset.org/#keypoints-eval, accessed: Mars 2 2022.

[50] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape, "Offline deformable face tracking in arbitrary videos," in *Proceedings of IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 1–9, 2015. DOI: http://dx.doi.org/10.1109/ICCVW.2015.126.

[51] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic, "The first facial landmark tracking in-the-wild challenge: Benchmark and results," in *Proceedings of IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 50–58, 2015. DOI: http://dx.doi.org/10.1109/ICCVW.2015.132.

[52] G. Tzimiropoulos, "Project-out cascaded regression with an application to face alignment," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3659–3667, 2015. DOI: http://dx.doi.org/10.1109/CVPR.2015.7298989.

[53] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognition*, vol. 96, p. 106977, 2019. DOI: http://dx.doi.org/10.1016/j.patcog.2019.106977.

[54] I.-C. Chen, C.-J. Wang, C.-K. Wen, and S.-J. Tzou, "Multi-person pose estimation using thermal images," *IEEE Access*, vol. 8, pp. 174964–174971, 2020.

[55] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 390–391, 2020. DOI: http://dx.doi.org/10.1109/CVPRW50498.2020.00203.

[56] K. Gurney, *An introduction to neural networks*. CRC press, 2018.

[57] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, "Posetrack: A benchmark for human pose estimation and tracking," in *Proceedings of IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5167–5176, 2018. DOI: http://dx.doi.org/10.1109/CVPR.2018.00542.

[58] J. Wan, Z. Lai, L. Shen, J. Zhou, C. Gao, G. Xiao, and X. Hou, "Robust facial landmark detection by cross-order cross-semantic deep network," *Neural Networks*, vol. 136, pp. 233–243, 2021. DOI: http://dx.doi.org/10.1016/j.neunet.2020.11.001.

[59] Y. Zhu, C. Li, Y. Liu, X. Wang, J. Tang, B. Luo, and Z. Huang, "Tiny object tracking: A large-scale dataset and a baseline," *arXiv:2202.05659*, 2022.

[60] M. Hall, L. van der Maaten, L. Gustafson, and A. Adcock, "A systematic study of bias amplification," *arXiv:2201.11706*, 2022.

[61] D. Zhao, A. Wang, and O. Russakovsky, "Understanding and evaluating racial biases in image captioning," in *Proceedings of 2021 IEEE/CVF International Conference on Computer Vision*, pp. 14830–14840, 2021. DOI: http://dx.doi.org/10.1109/ICCV48922.2021.01456.

[62] A. Benjumea, I. Teeti, F. Cuzzolin, and A. Bradley, "YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles," *arXiv:2112.11798*, 2021.

[63] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE international conference on image processing (ICIP)*, pp. 3645–3649, IEEE, 2017. DOI: http://dx.doi.org/10.1109/ICIP.2017.8296962.

DEPARTMENT OF PHYSICS
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY