UNIVERSITY OF
BATH

Multi-Modal Learning For Robust Key Point Estimation: Application On Human Pose Estimation

Abdelhadi Mar'i

Msc in Machine learning and Autonomuous Systems
The University of Bath
2022-2023

UNIVERSITY OF BATH

Multi-Modal Learning For Robust Key Point Estimation: Application On Human Pose Estimation

Submitted by: Abdelhadi Mar'i

Copyright

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of Master of Science in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Abstract

Human Pose Estimation has been rapidly developing and has received a big spotlight due to the significant revolution happening in Deep learning. With the introduction of complex methodology of the Convolutional Neural Network, it has opened the field of progress for HPE. There are now different types of HPE systems, such as 2D or 3D, single view and multi-view, single-person, and multi-person models. Human Pose Estimation has reached an accurate rate in predicting human poses and joint location and is now being developed to yield a high accuracy in common context scenarios. However, human pose estimation is still considered a very challenging task, as it is still difficult to gather accurate pose estimation results in common context scenarios that involve rapid, explosive movement, and occluded joints as it makes it intrinsically difficult for a model to locate accurate positions of the human joints. As well as other factors such as multi-person frames or objects obscuring the full view of a person from the monocular camera's viewpoint. The introduction of depth data into computer vision models has proven to be successful in making several object detection models more robust and accurate. This dissertation has explored and evaluated the effect of depth data on 2D human pose estimation, and displays how depth information can assist human pose estimation models in handling these challenges.

Contents

List of Figures

List of Tables

Acknowledgements

My commendation goes to my supervisor, Jordan Taylor, for his unwavering support and guidance throughout this project.

Chapter 1

1.1 Introduction

Human pose estimation plays a crucial role in a wide range of applications, including action recognition, sports analysis, and human-computer interaction. The accurate and robust estimation of human poses from images or videos is essential for extracting meaningful insights and understanding human activities. However, training human pose estimation models is inherently challenging due to several factors.

Figure 1.1: HPE applied to analyse a player's position before and movement while taking a shot. Taken from "Automated tracking of body positions using match footage" paper [4].

The struggle to capture fine-grained details and handle occlusions, which affect the accuracy of your pose estimation has been faced by many human pose estimation models, it was also reported in the DeepPose official research paper [27]. Explicit annotations may also restrict the model's ability to learn subtle variations and inherent structures, hindering its performance with unseen scenarios or challenging poses. In addition, multi-person estimation is a challenge faced by many models that results in multiple false predictions, that was also expressed in the convolutional pose machine (CPM) model that was introduced in [28]. To address these challenges, this dissertation explores the potential and ability of the introduction of depth data into a human pose estimation model. The introduction of depth data in computer vision models by training the models with RGB-D channels has given several object detection models [30, 22, 15, 21] the required advantage to overcome the challenges they previously faced. By allowing the model to get a 3D understanding of the environment within the frame, it allows it to draw better patterns between the locations of different joints and their surrounding environment than a 2D understanding. By advancing our understanding of human pose estimation models and the introduction of depth data into computer vision models, we look to introduce a more robust and accurate version of a human pose estimation model that can better handle and work with the challenges faced by human pose estimation.

1.2 Problem description

The performance of Human Pose Estimation methods has recently increased significantly, and they are successful in yielding high accuracy in key point detection on several data sets. However, human pose estimation models face multiple challenges. One prominent issue is the models' abilities to be able to deal with complex poses, where a person might be sitting down or doing a rapid movement. Such scenarios lead to blurry frames, and occluded joints, where the model would be unable to detect the location of the joint, and is limited to the 2D data that it is receiving as seen in [27, 28]. Traditional supervised learning methods heavily rely on labelled data, which may not be sufficient to capture the full range of poses in scenarios that include robust and rapid movement. Furthermore, pose estimation models, such as [16, 29, 19, 18, 3, 20], have faced complications in predicting joints in multi-person frames. The models would fail to predict the key points when the joints of different people are in close proximity to one another, and the models have no method of recovery or method to deal with such a situation.

Mutli-modal learning methods have emerged as a promising solution to address the limitations of data in various computer vision tasks. By leveraging valuable dimensional spatial information, models can learn from unlabelled data and exploit inherent structures or patterns

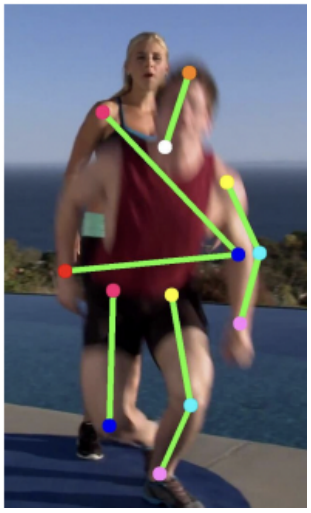

Figure 1.2: A HPE model that results in noisy and false predictions due to obscurity of joints due to multiple people being in the frame [5].

within the data. Object detection learning with depth has shown potential in improving model predictions in very complex scenarios, such as camouflaged object detection for example [30]. The results demonstrated significant improvements in object detection accuracy compared to a previous model that was trained only on RGB data without any depth information.

1.3 Aims

This project aims to deliver a more robust and improved human pose estimation that can better address the challenges faced in human pose estimation by training it with additional depth data. The main human pose estimation challenges that will be targeted within this project would be complex poses, occlusions within the data and joint obscurity. The project aims to explore the effect that depth data will have on the overall performance of the model and its effect on different key points estimation in human pose estimation and to understand the influence that depth information has created on feature extraction within a human pose estimation model. The project will be carried out in correlation with the tasks scheduled in Appendix A.1.

1.4 Contributions

Human Pose Estimation (HPE) is a quickly advancing field within computer vision that has seen significant progress with the adoption of deep learning which led to deep neural network architectures, specifically Convolutional Neural Networks (CNNs), that achieved remarkable advancements in accurately detecting and localizing human body key-points. These deep learning models can automatically learn and extract meaningful features from large amounts of training data, enabling them to capture complex patterns and variations in human poses. Recent HPE models have managed to yield high accuracy in key point detection, where an advanced model such as Residual Steps Network (RSN) was able to yield the highest mean average precision (mAP) of 78.6% in the COCO 2019 keypoint challenge, by getting a 3% higher mAP than the winner of the COCO2017 key points challenge [18]. Whereas, the last key points challenge made by COCO was the COCO 2017 key points challenge, where OpenPose [18] won the challenge with an mAP of 75.6%. However, many challenges in HPE models still exist, whereas the models struggle to handle joint obscurity, image blurs, and multi-person frames where people are in close vicinity of each other. This dissertation will investigate the effect of introducing another modality (depth) into the data of a human pose estimation model. It will investigate how multi-modal learning with depth can be adapted and the effect it has to yield a model that has a high accuracy in key point detection in different contexts and complicated scenarios.

Chapter 2

Literature and Technology Survey

2.1 Related Work

Several accurate and real-time models have been developed for the purpose of human pose estimation in recent years, provided the significant development in deep learning. Models like [26, 18, 3, 20, 21], can provide real-time or close to real-time accurate inference of human pose key points estimation. Many of the models differ in their approach. Some models follow a top-down approach [17, 8], where the model first detects each person in an image frame with bounding boxes and runs inference on the join key points of each person separately. Whereas other models run a bottom-up approach [3, 20], where a model tries to detect all the key points in an image frame and then tries to identify the different people in an image.

2.1.1 Open Pose

OpenPose is a bottom-up human pose estimation model, one of the best performing models created for human pose estimation with benchmark results. It achieves one of the top performances on the COCO 2017 (common objects in common scenarios) test data set with a mean average precision (mAP) above 70% , winning the COCO 2017 key points challenge, as reported in the official OpenPose published paper [18]. Openpose is able to provide real-time inference, producing inference on around 25-30 frames per second (FPS). The model created for openpose is called BODY25, which uses VGG-19 as backbone. The model introduces a new solution to the problem of running inference on multi-person frames called part affinity fields (PAF).

The paper addresses the challenge of assembling full-body poses from detected body parts, particularly in situations where multiple people are present. The key problem is to determine the confidence level of the association between each pair of detected body parts, indicating whether they belong to the same person. A method suggested in the paper to assess this association is by detecting an additional midpoint between each pair of body parts on a limb and checking for its presence among candidate part detections. However, this method becomes problematic when people crowd together, as these midpoints can result in false associations. The paper introduces Part Affinity Fields (PAFs) as a solution to these limitations. PAFs maintain both the location and orientation information within the region of support of a limb, addressing the issues mentioned earlier. Essentially, PAFs consist of 2D vector fields for each limb, where each pixel in the limb's area encodes a direction pointing from one part of the limb to the other. This approach allows for more accurate and reliable association of body parts, making it valuable for tasks such as human pose estimation.

Initially, the features are extracted by the network, an initial estimation of part affinity fields and heatmaps is performed, following the five refinement stages. The model is able to detect and find 18 types of keypoints. Then grouping procedure searches the best pair (by affinity) for each keypoint, from the predefined list of keypoint pairs. The pipeline is illustrated in Fig. 1. During inference, input image is resized to match network input size by height, the width is scaled to preserve image aspect ratio, then padded to the multiple of 8 [18].

2.1.2 Alpha Pose

Another well-performing human pose estimation model is AlphaPose, which also follows a top-bottom framework, and uses ResNet as its backbone network. AlphaPose has achieved a mAP of 74% on the COCO dataset, being the second best performing model in the COCO 2017 key points challenge by getting an mAP of 1.6% less than the winner of the challenge OpenPose [18]. AlphaPose addresses the drawbacks of several common top-down human pose estimation models. Being that in top-down approaches, since the detection stage and the pose estimation stage are separate if the detector were to fail, there would be no method for the pose estimator to

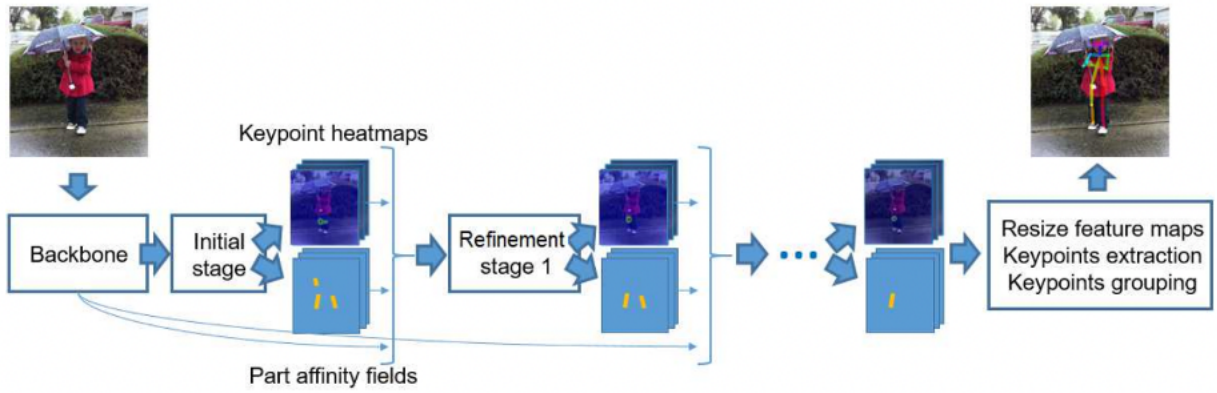

Figure 2.1: The network initially extracts features, followed by an initial estimation of heatmaps and pafs. Subsequently, it undergoes five refinement stages. This network is proficient at detecting 18 different types of keypoints. A grouping procedure is then employed to identify the optimal pair, based on affinity, for each keypoint. These pairs are drawn from a predefined list of keypoint pairs, such as left elbow and left wrist, right hip and right knee, left eye and left ear, totaling 19 pairs in all. During inference, the input image is resized to match the network's input size by adjusting the height, while the width is scaled to maintain the image's aspect ratio, followed by padding as needed. This figure diagram is taken from the official open pose research paper by [18].

recover the human pose, and current researchers adopt strong human detectors for accuracy, which makes the two-step processing slow in inference [26]. The solution proposed by AlphaPose is to lower the detection confidence and Non-Maximum-Suppression threshold to provide more candidates for subsequent pose estimation. Non-Maximum-Suppression is a post-processing technique used to refine the output of a pose estimation model by removing redundant and overlapping pose predictions. The resulting redundant poses from redundant boxes are then eliminated by a parametric pose Non-Maximum-Suppression, which introduces a novel pose distance metric to compare pose similarity.

The Alphapose network model is shown in figure (figure). The model first detects different people in a frame using an off-the-shelf object deter like YoloV3. A cropped frame is created specifically for every person in the frame individually to be passed through pose estimation and tracking networks to return inference on the coordinates of every key point.

Techniques introduced in this paper to solve key point miss detection, such as the symmetric integral regression, are introduced for more accurate key point localization. This solution targets the issues that exist in conventional soft-argmax operation for key point regression. Integral regression, also known as the soft-argmax operation, is introduced as a differentiable method. It transforms heatmap-based approaches into regression-based approaches, allowing for end-to-end training. The operation estimates joint locations by computing a weighted sum of pixel coordinates and pixel likelihoods on a heatmap after normalization.

when the gradient of the loss with respect to pixel likelihood is calculated assymetry is exhibited. This asymmetry arises from the dependence of gradient amplitude on the absolute position of the pixel, not the relative position to the ground truth. As a result, for the same distance error, the gradient varies depending on the position of the keypoint, which can disrupt the translation invariance of Convolutional Neural Networks (CNNs) and impact performance. To address the gradient asymmetry and improve learning efficiency, an Amplitude Symmetric Gradient (ASG) function is proposed. ASG is an approximation to the true gradient and is given as shown in 2.1.

$$\delta_{\text{ASG}} = A_{\text{grad}} \cdot \text{sgn}(x - \hat{\mu}) \cdot \text{sgn}(\hat{\mu} - \mu) \qquad (2.1)$$

The amplitude of gradients A_{grad} is a manually set constant, typically set to 1/8 of the heatmap size. ASG centers the gradient distribution at the predicted joint locations $\hat{\mu}$, which can improve the utilization of heat maps during training and approximate ground-truth locations more directly.

In summary, the integral regression operation is introduced to enable differentiable training in heatmap-based keypoint detection. However, the asymmetry in the gradient amplitudes is identified as a problem, and the

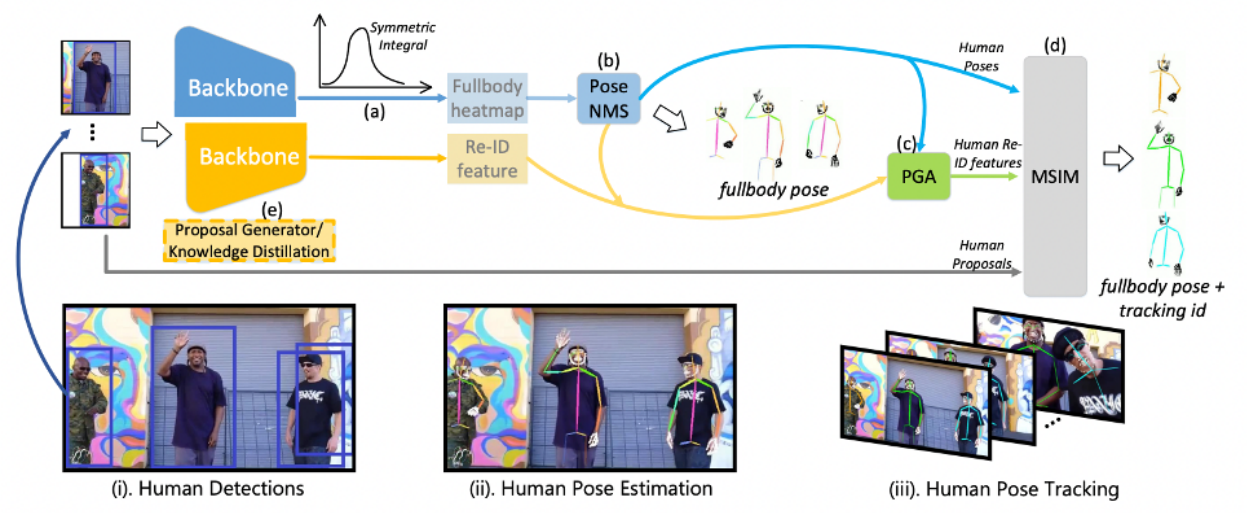

Figure 2.2: Alpha Pose model architecture. Figure taken from the official Alpha pose paper [26].

Amplitude Symmetric Gradient (ASG) function is proposed as a solution to address this issue and improve the learning process. This approach aims to better leverage heatmap information for accurate keypoint localization.

2.1.3 MoveNet

Another model that applied the bottom-up approach is Movenet. MoveNet is geared towards accuracy, the model expects input images of size (256×256) with three channels but employs an extensive depth multiplier of 1.75. These channels can also be seen as feature maps in individual layers. MoveNet adopts a bottom-up approach, with MobileNet V2 serving as its feature extractor. MoveNet is particularly influenced by CenterNet [10]. Unlike standard anchor-based detection models that rely on bounding boxes, CenterNet adopts a unique strategy. It treats the centre point as the sole anchor, seeks and classifies objects through regional proposals, and refrains from inferring objects solely based on Intersection over Union (IoU) values. This innovative approach eliminates the need for Non-Maximum Suppression (NMS) and excels in distinguishing between objects in a single stage, thereby achieving high performance.

As depicted in figure 2.3, MoveNet efficiently processes all four steps concurrently. Firstly, it generates a heatmap centred around the individual, pinpointing their presence. The system then discerns the location with the highest heatmap score, effectively identifying the person's position. Subsequently, an initial set of key points for that individual is established through a regression-based approach. The system confirms a person's identity by validating that the regression aligns with the arrangement of predetermined key points. Moreover, MoveNet incorporates a weighted approach by assigning each pixel a weight inversely proportional to its distance from the regressed key point. This intelligent weighting mechanism ensures that key points belonging to background individuals are effectively excluded from the computation. In the final step, the set of key points is meticulously refined, ultimately determined by the maximum heatmap values within each key point channel.

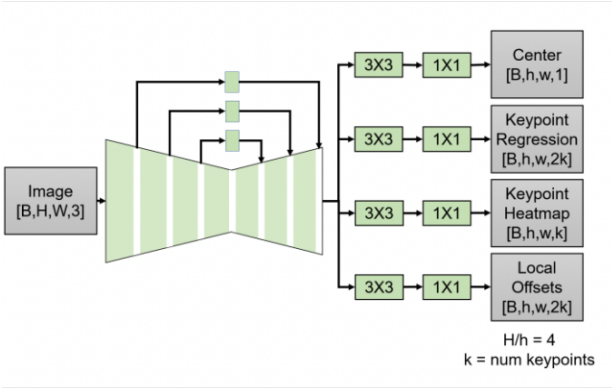

Figure 2.3: MoveNet model architecture [24].

Not much documentation or research information has been shared publicly by the creators of Movenet. In the Table 2.1, average precision and mean average precision data when evaluated by the COCO test dataset in the COCO 2017 key point detection challenge, where Alpha Pose records higher results than Open Pose in its accuracy. Moreover, AlphaPose houses fewer GFLOPS (Giga

floating points per second) making it more computationally efficient when running inference and can be easier to deploy and run using different hardware platforms.

A comparison has been made between three state-of-the-art human pose estimation models (Pose Net, Movenet, and Blazepose) in [11] for their application in smart mirrors. The paper focused on four localization errors, which are: jitter, miss, inversion, and swap. Despite PoseNet having the best accuracy within their results, compared against the COCO dataset. The paper chose MoveNet as the best of the three models, in terms of both speed and precision, as it offers a reliable and fast form of pose estimation. The paper also goes on to mention how having depth information provided in human pose estimation models will open functionality provided by physical rehabilitation systems as corrections made by the models won't only be limited to the 2D information but will also consider depth information. Suggesting that models with depth information would be able to detect better patterns to estimate 2D coordinates for key points in human pose estimation models.

2.2 Depth in object detection

In [30], the paper studies depth information in camouflaged object detection. The paper utilizes depth maps generated using established monocular depth estimation (MDE) techniques. The paper introduces two solutions to the problem caused by the domain gap between the MDE dataset and the camouflaged object detection, to result in more accurate depth maps that can be directly employed.

The first technique applied was an auxiliary depth estimation branch that aims to predict more accurate depth maps. The paper found that this was particularly crucial when dealing with their scenarios of "generated depth". The second technique introduced is a multi-modal confidence-aware loss function through a generative adversarial network. This loss function helps in appropriately weighing the influence of depth information on camouflaged object detection. The extensive experimentation on various camouflaged object detection datasets highlights that conventional RGB-D segmentation techniques based on "sensor depth" perform inadequately when applied to "generated depth". Our proposed dual solutions work synergistically, effectively exploring the contribution of depth for improved camouflaged object detection.

As there are no RGB-D datasets that currently exist for camouflaged object detection, the paper utilises multi-modal learning, and depth was extracted for the COD training data set with existing monocular depth estimation methods. Since the conventional monocular depth estimation models are trained on natural images, there may not exist any camouflaged objects. The paper also evaluates the results of three different state-of-the-art off-the-shelf monocular depth estimation models (Midas, Monodepth2, and Frozen People). Overall, trained on 10 different datasets, Midas has displayed the best cross-dataset performance and the strongest generalization ability. Whereas, MonoDepth2 is trained only on the KITTI dataset [[7]], which is why it performed the worst in their scenario.

The depth contribution exploration network takes three input channels (RGB) of an image as shown in 2.4 as input. The model first runs an RGB image-based object detection, then an auxiliary depth estimation, and then runs an RGB-D object detection. Furthermore, the suggested multi-modal confidence-aware loss function provides a clear assessment of depth contribution by utilizing prediction confidence (Crgb and Crgbd) as an indicator. This confidence is derived from the newly introduced probabilistic model, which leverages a generative adversarial network.

2.3 Data Sets

The choice of dataset is very effective towards a human pose estimation model's performance; thus, the choice of what dataset a model is trained on is very critical. There are several human pose annotated datasets that are publicly available. The choice of dataset relies on several factors such as the number of key points, the definition of frames, and diversity of data to avoid variance and over fitting.

2.3.1 Leeds Sports Dataset

The Leeds Sports Pose dataset by [12] contains 2000 pose annotated images of mostly sports people gathered from Flickr using the tags shown in Figure 5 above. The images have been scaled such that the most prominent person is roughly 150 pixels in length. Each image has been annotated with 14 joint locations. Left and right joints are consistently labelled from a person-centric viewpoint. The paper by [12] demonstrates how the pose appearance in the data set was able to improve accuracy in pose estimation. The Leeds Sports Pose dataset specifically focuses on sports activities, providing a variety of sports-related poses and movements. The dataset offers fine-grained annotations for multiple bodily joints and includes images with occlusions which can help train a robust model to handle obstructed body joints. However, the Leeds Sports Pose dataset is relatively smaller in image numbers when compared to datasets

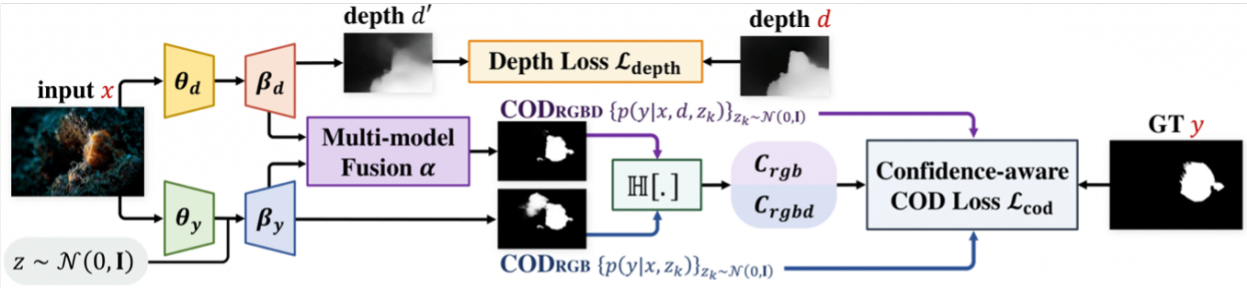

Figure 2.4: Model architecture of the multi-model camouflaged object detection model. Taken from official paper [30].

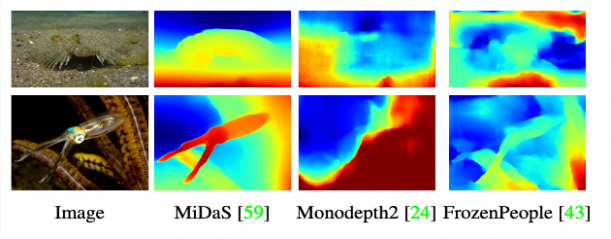

Figure 2.5: Visual comparison between state-of-the-art Monocular depth estimation models [30].

Figure 2.6: LSP dataset image examples with sport labels.Figure taken from [12].

like MPII. It also does not cover a wide range of sports and poses as comprehensively as other datasets.

2.3.2 MPII

The MPII Human Pose dataset by [1], includes around 25K images containing over 40K people with annotated body joints. The images were systematically collected using an established taxonomy of everyday human activities. Over- all, the dataset covers 410 human activities, and each image is provided with an activity label as shown in Figure 6 above. All images were extracted from YouTube videos. For the research, rich annotations

Figure 2.7: MPII image examples with scenario labels. Figure taken from the official paper released for the dataset [1].

of occluded joints and limbs were attained. The MPII dataset contains a significant number of images with diverse human poses with accurate annotations for 16 key point locations on the human body, making it suitable for training pose estimation models. The MPII dataset focuses on general human poses and activities. The MPII dataset contains images primarily from a controlled indoor environment and often features people in specific poses (e.g., standing, sitting). It may not represent the full diversity of human poses and appearances encountered in real-world scenarios, such as sports, dance, or unconventional activities. In addition, the dataset contains relatively few examples of occluded body parts or complex poses, which may not adequately challenge advanced human pose estimation models that aim to handle diverse and challenging scenarios.

2.3.3 Sports Videos in the Wild (SVW)

The SVW dataset by [25], is comprised of 4200 videos captured solely with smartphones by users of Coach's Eye smartphone app, a leading app for sports training developed by Tech-Smith corporation. SVW includes 30

Figure 2.8: SVW image examples with scenario labels.Figure taken from [25].

categories of sports and 44 different actions. Due to the imperfect practice of amateur players and unprofessional capturing by amateur users, SVW is very challenging for automated analysis. However, only includes sport tags (as shown in figure 2.8 above) and does not include any joint or key point annotations, it could allow us to use the data for a semi-supervised or self-supervised approach with the combination of data from a different annotated dataset.

2.3.4 COCO 2017

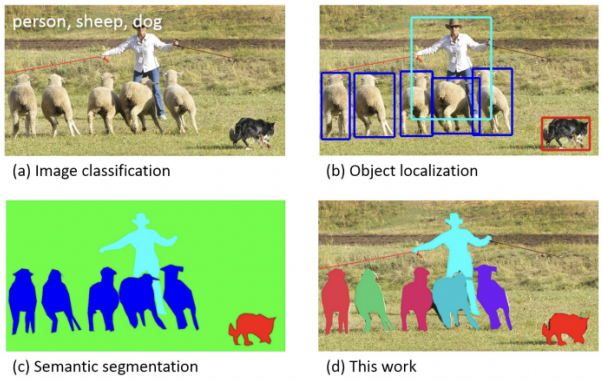

Figure 2.9: COCO dataset 2017 different annotations and labels examples. Figure taken from the official COCO paper[13].

Microsoft has introduced its COCO (common objects in common contexts) dataset with the goal of advancing the state of the art in object recognition. This is achieved by gathering images of complex everyday scenes containing common objects in their natural context. Objects are labelled using per-instance segmentations to aid in precise object localization as can be seen in figure 3.1. The dataset contains photos of 91 object types, with a total of 2.5 million labelled instances in 328k images. This diversity can be beneficial for training robust human pose estimation models as it includes various environments, lighting conditions, and poses. Although COCO is primarily for object detection, it also provides annotations for keypoint detection. Specifically, it annotates 17 key points per human instance, making it useful for human pose estimation tasks. However, while COCO annotates key points for humans, not all human instances in the dataset have complete key point annotations. Some instances may have missing key points, which can be challenging for training models effectively.

While primarily designed for instance segmentation and object detection, Mask R-CNN by [9] has also been adapted for human pose estimation using COCO as a training dataset. Open Pose has also leveraged COCO annotations for training its model. It has been used in various applications, from gesture recognition to action recognition.

In summary, using the COCO dataset for human pose estimation has both advantages and limitations. It provides a vast and diverse source of data, but it may not be as specialized or detailed as datasets explicitly created for pose estimation. Researchers and developers often choose to fine-tune models trained on COCO with additional data or use it as a starting point for training more specialized pose estimation models.

2.4 Metrics for model evaluation

To evaluate and analyse the performance of a human pose estimation, suitable metrics will need to be used to measure the desired outcomes of a human pose estimation model. The metrics need to align with the purpose and aim of the model and allow to make a clear evaluation of the performance and model.

2.4.1 Intersection Over Union (IoU)

Intersection Over Union is a metric that is commonly used in object detection models and can also be applied to human pose estimation model. Intersection Over Union evaluates the overlap between a predicted region, like bounding boxes for example, and the ground truth region. It quantifies the accuracy of the prediction (spatial similarity) by measuring the ratio of the intersection area to the union area between the predicted region and the ground truth region. By evaluating our human pose estimation model using IoU, we can assess its accuracy, robustness to scale, and spatial alignment with ground truth regions. It provides a quantitative measure of performance that can be compared across different models or variations of our model. It is also sometimes used in human pose estimation to determine viable predictions for calculating mean average precision.

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{2.2}$$

2.4.2 Mean average precision (mAP)

mAP is a commonly used metric in object detection tasks and can also be applied as a metric to evaluate a pose estimation model. It considers the precision and recall values at different thresholds to compute an overall average precision. MAP provides a comprehensive evaluation of a model's performance. The mAP is obtained by calculating the mean of the average precision of each key point class, which is computed by a matching process of the spatial proximity and semantic correspondence between the model's predicted key points and the ground truth key points. As shown in 2.3, where **N** is the total number of predictions, and $\mathbf{AP_i}$ is the average precsion of the prediction

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i \tag{2.3}$$

In the COCO 2017 human pose estimation model challenge evaluation guidelines, mAP was calculated by averaging the AP over 80 object classes and all 10 IoU thresholds from 0.5 to 0.95 with step size of 0.05 [13].

2.4.3 Inference Speed

Inference speed refers to the time it takes for a model to process an input and generate predictions. It is a measure of the model's computational efficiency and real-time performance. In the context of human pose estimation, inference speed indicates how quickly the model can estimate poses for input images or video frames. By evaluating the inference speed, we can assess if the model can process frames quickly enough to meet the real-time constraints of the application. It also allows us to identify models that can achieve the desired accuracy while being optimized for specific hardware or deployment environments. Moreover, Inference speed evaluation allows for a trade-off analysis between speed and accuracy. By measuring a pose estimation model's inference speed, we can determine the optimal balance between computational efficiency and pose estimation accuracy for a specific application. In pose estimation models, we would be running asynchronous calculations of inference speed, as we would not be waiting for a result of the calculation before running the calculation for the following frame.

Chapter 3

Design

3.1 Initial Ideas

3.1.1 Sensors For Depth Information Collection

Figure 3.1: ITOP dataset frames examples, the environment remains the same. Only the poses, and person in images change. No blurry frames, obstructed joints, or more than one person in the frame. Samples taken from the official EVAL dataset[6].

An different approach initially introduced, prior to exploring multi modal fusion, for us to create a highly accurate human pose estimation model using depth data. Which was to investigate the ability to create a significantly high-accurate 2D human pose estimation model through an RGB-D channel, wherein depth is retrieved using the IR sensors on the Xbox360 Kinect camera. Following the proceedings in [31], where they trained a model to estimate 3D human pose from point cloud. The model was trained on the ITOP dataset. The ITOP dataset, which is made publicly available by [14], is adapted from the EVAL dataset by [6], and contains 100k images of 20 different people, a single person within each frame. The images are captured using the IR sensors on Microsoft's Kinect camera at 30fps. The data images provided from each pose are given from two viewpoints, a side vantage point, and a top vantage point relative to the object. Coordinates of the joint key points are given for both image views. Coordinates of height, width, and depth are given.

This method would have provided very robust results and inference, given that loss in depth calculation would have been non-existent, as a sensor would be collecting the readings for depth. This would allow the model to marginalize heavily on the effect of depth information, and the distance between different joints to gain high accuracy in pose estimation, and also provide accurate estimations to complex poses.

However, several problems could've risen from the application of such a method. Firstly, training the model on data captured using a combination of an RGB camera and IR sensors, as shown in figure 3.2, would mean that all data to test and even run inference on the model would need to be captured using an RGB camera with depth data, needs to be captured using IR sensors for best results, and that would significantly influence the application of the model and it's use.

Figure 3.2: A depiction of Microsoft's Xbox360 Kinect camera.

It would make the model unfeasible to deploy on several devices such as phones, laptops, and desktops. These devices would typically only contain an RGB camera,

which makes our model only able to be deployed and used with specific hardware. It would also mean that there can't be a line of comparison that we can draw to evaluate it against existing 2D human pose estimation models. However, the biggest issue that came to light was the dataset. Even though the dataset is large and uses different people in its frames, all samples are homogeneous. Wherein, all frames are taken in the same indoor environment, and the only changed variable is the person and their position in the frame. There are no obstructed joints, and all joints are clear. There is not a single blurry frame either. That would mean that the model would be highly susceptible to incorrect predictions when it's used to run inference in common contexts, as it will display no contextual understanding of blurry frames, obstructed joints, or a case where another person might show in a frame.

Chapter 4

Implementation and Testing

4.1 Methodology

4.1.1 Introduction

As previously mentioned, 2D human pose estimation is quite a complicated task in computer vision. There are several state-of-the-art 2D human pose estimation models out there that give promising results in human pose estimation, however, still have their weak points when running inference on specific frames. Frames where some joints are obscured or are indistinct, or where there are multiple people in a single frame. As previously mentioned, this paper aims to experiment with a new technique that leverages the vast development of deep learning to be applied in human pose estimation. Offering a good approach to further enhance existing state-of-the-art human pose estimation models. Following the proceedings in [21], where the CCTV images used in video surveillance have been shown to house occlusion problems due to loss of topological information caused by projecting 3D real world in 2D image. The occlusion that causes inaccurate object detection can be resolved by using depth information. The paper later shows how the introduction of depth information into the model has allowed for robust key-point detection. The paper claims that their proposed method solves the occlusion problem by performing object detection by adding depth information based on this object detection method. We have also seen in section 2.2 how the introduction of depth data through multi-modal infusion, has allowed solving the problem of camouflaged object detection.

In this paper, we will investigate the effect of introducing depth data into 2D human pose estimation by training a model on RGB-D data. We have seen object detection models trained on RGB-D data prove to be successful in the context of camouflaged object detection 2.2. We will investigate whether it will be successful in improving the accuracy and robustness of a human pose estimation model, and what effect would it have on the inference generated by the model. A modified version of the model will be created to accept and be trained on RGB-D data. The results of the modified model will be compared to and evaluated against the results of the original version of the model that was trained on RGB data of the same dataset. This will allow us to explore and visualize the direct effect that RGB-D information has had on the features and inference of the model.

Figure 4.1: Comparison of key points estimation before and after including depth data [21].

4.1.2 Model

The model of choice to carry out this experiment with is MoveNet. MoveNet is known for its efficiency due to its MobileNet V2 backbone, this efficiency makes it easier to train and more suitable for real-time applications, especially when integrating RGB-D data. Other models, such as OpenPose and AlphaPose, on the other hand, use deeper and more computationally intensive networks (e.g., ResNet), which can be challenging to train and adapt for real-time performance when incorporating depth information as discussed in 2.1.

4.1.3 Data fusion

The choice of dataset that would suit this project needs to be a diverse and large human pose estimation dataset where the image frames are taken in different contexts and environments, and the pixel information needs to be given along 4 channels, RGB-D. However, the only human pose datasets that contain RGB-D data are all

taken in controlled environments, such as ITOP data set [14] which contains only single-person frames in indoor environments.

A different approach was taken by [15] and [30], through multi-modal learning. In [30], there was no dataset that contained image data and depth available for camouflaged object detection. Multi-modal learning was sought out to combine image data from an RGB dataset and the depth of every frame was extracted using a monocular depth estimation model. Moreover, in [15] they sought an approach that was based on a mixture of a CNN that incorporated several modalities including appearance, depth, and motion. Both models were able to output a very visible improvement once multi-modal learning was applied to their models, making them more robust and able to run more accurate inferences.

A feasible approach, which we've seen successfully implemented in [30] from 2.2, is through multi-modal learning. We will implement a modification of an existing RGB human pose dataset. We will export the depth D in each pixel of every sample within the dataset, using a state-of-the-art monocular depth estimation model, and append the D depth to the array of RGB channels as a fourth channel.

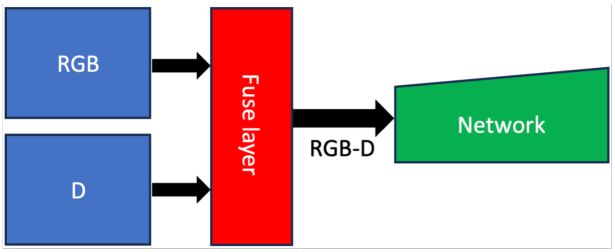

Figure 4.2: high-level demonstration model of data fusion within the model.

We will proceed with the method of early fusion as shown in figure 4.2. The method of fusing data from multiple sources, where each modality represents a distinct type of data, where we combine data at the input level, is called early fusion. A paper submitted by [2], has shown the benefits of early fusion, by demonstrating that immediate fusion of audio and visual inputs in an C-LSTM later results in higher performing networks that are more robust to the addition of white noise in both and visual inputs. They have experimented with different with different layer fusion, and the more robust results were retrieved from the early fusion model, where the data was fused at the input layer.

4.1.4 Data Set

The data set chosen to train the model is the COCO2017 data set 2.3.4. The COCO2017 data set is one of the largest and most diverse data sets available for human pose estimation [13]. It contains a wide range of images with diverse backgrounds, poses, and lighting conditions. This diversity makes it suitable for training models that need to perform well on real-world data. COCO2017 images capture scenes from everyday life, making them highly relevant for applications like human-computer interaction, surveillance, sports analysis, and more. The diversity of this data set is highly likely to allow a distinctive result between the model trained on RGB-D and the model trained on RGB. It should allow for a more robust result, and it should help us analyse its effect on each feature extracted by the model when running inference. However, since Microsoft has not made the annotations for its test data set publicly available, we have used the training data set that is comprised of approximately 118,000 images for training and validation, and used the validation data set to serve as a test data set and help us test our models.

4.1.5 Monocular Depth Estimation Model

Depth has been extracted from the RGB images in the dataset using the monocular depth estimation model Midas. MiDaS estimates relative inverse depth based on a single input image. The repository offers a variety of models tailored for diverse applications, spanning from a compact, high-speed model to an extensive model delivering the utmost accuracy. These models have undergone training using multi-objective optimization techniques on ten distinct datasets, guaranteeing robust performance across a broad spectrum of input scenarios. Midas has proved to be more robust and accurate than other state-of-the-art monocular depth estimation models for use in complex object detection as displayed in 2.2.

4.2 Loss Functions

As discussed thoroughly in 2.1.3 and shown in figure 2.3, the architecture of MoveNet closely resembles CenterNet [10] and leverages MobileNetV2 as the foundation for its feature extractor. Additionally, a Feature Pyramid Network (FPN) has been integrated. Notably, with an output stride set to 4, this design is capable of efficiently processing high-resolution feature map outputs. The model generates 4 outputs, including a heatmap indicating the centre of a person, a field for regressing key points, a heatmap highlighting the person's key points, and a field containing 2D offset information for each key point. Thus, we calculate the loss of each output and

Figure 4.3: output examples from running inference of monocular depth estimation using Midas. Taken from the official Midas research paper [23].

use it for training in the model to adjust the weights.

4.2.1 Offset Loss

the offset loss function calculates the L1 loss (mean absolute error) between predicted and ground truth key point offsets for each key point, considering a mask that identifies which key points are relevant for each sample. This loss guides the model during training to learn how to predict the positional offsets of key points accurately, which is crucial for tasks like human pose estimation. The loss is normalized by dividing by the number of key points to make it scale-invariant.A mask is applied to the L1 loss for each key point. This mask indicates whether a particular key point is present in the image or not. The mask is used to exclude key points that are not visible or relevant for a given sample. The total loss is divided by the number of keypoints (number of joints) to calculate the average loss per keypoint. This normalization ensures that the loss is not affected by the number of keypoints.

$$\text{Loss} = \frac{1}{N}\sum_{j=0}^{N}\left(|\mathbf{x}_{\text{gt}} - \mathbf{x}_{\text{pred}}| + |\mathbf{y}_{\text{gt}} - \mathbf{x}_{\text{pred}}|\right) \quad (4.1)$$

4.2.2 Regression Loss

The regressions loss helps assess how well the model predicts the positional offsets (x and y coordinates) of key points compared to the ground truth. It tells us how accurate the model's key point predictions are in terms of their spatial positions. A lower value of this loss indicates that the model's predictions are closer to the ground truth, signifying better performance in key point localization. In summary, regression loss was used to train the model model to predict the correct x and y coordinates for key points, which is essential for human pose estimation. The loss is normalized to be independent of the number of key points. Regression loss is calculated as shown in 4.2, where N is the total number of joints, j is the joint number, and L1 is the absolute difference between the predicted and ground truth points.

$$\text{Loss} = \frac{1}{N}\sum_{j=0}^{N}\left(L_1(\text{mask}_j(\mathbf{x}_{\text{pred}} - \mathbf{x}_{\text{gt}})) + L_1(\text{mask}_j(\mathbf{y}_{\text{pred}} - \mathbf{y}_{\text{gt}}))\right) \quad (4.2)$$

4.2.3 Centre Loss and Heatmap Loss

The centre loss and heatmap loss are a calculation of their prediction's MSE, and a custom loss function was created. The loss function provides a way to adjust the contribution of each prediction-target pair in the overall loss calculation. By using a weight mask that depends on the target values, we allowed the model to focus more on certain samples or regions of interest in the data. This is particularly useful when dealing with imbalanced datasets or when specific samples require special attention in the training process. The loss value itself represents how well the model's predictions match the ground truth, with greater emphasis placed on samples or elements that are weighted more heavily. A lower loss value indicates better alignment between predictions and targets, reflecting improved model performance in capturing the desired patterns in the data. In summary, the MSE loss function created is a customized loss function that helped the model pay varying levels of attention to different samples based on their importance, which lead to more effective training and better adaptation to specific challenges in the data. We calculate the MSE as shown in 4.3.

$$\text{Loss} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(\text{pred}_{ij} - \text{ground\_truth}_{ij}\right)^2 \cdot \left(8 \cdot \text{ground\_truth}_{ij} + 1\right) \quad (4.3)$$

Where:

- M is the number of rows in the tensors.
- N is the number of columns in the tensors.
- $\mathbf{pred}_{ij}$ represents the element at row i and column j in the predicted tensor.
- $\mathbf{ground\_truth}_{ij}$ represents the element at row i and column j in the ground_truth tensor.

4.2.4 Bone Loss

The bone loss function assesses how well the model predicts bone-related information (e.g., relationships between key points or structural elements) compared to the ground truth. It tells us how accurate the model's predictions are with respect to these structures. A lower loss value indicates that the model's predictions align more closely with the ground truth in terms of bone structures. Therefore, this loss helps train the model to recognize and predict the correct structural relationships between key points or elements, which is crucial for human pose estimation. In summary, the bone Loss function evaluates the accuracy of the model's predictions regarding structural relationships (bones) and guides the training process to improve the model's ability to capture these important patterns in the data. We calculate the loss based on the Frobenius norm (also known as the L2 norm or Euclidean norm) between pairs of bone-related elements in the predicted and target tensors. Specifically, we compute the Frobenius norm for each pair of bone elements and average these norms to get the final loss as shown in 4.4.

$$\text{Loss} = \frac{1}{|\text{N}_{\text{bones}}| \cdot \text{N}_{\text{images}}} \sum_{\text{b}_\text{i} \in \text{b}_{\text{total}}} \|\text{prediction} - \text{ground\_truth}\|_F \tag{4.4}$$

Chapter 5

Results

5.1 Metric Results

5.1.1 Models' Losses

Table 5.1 contains the losses calculated for both models when we ran them on the test dataset. The thesis drawn from the results is that the RGB model has accumulated a slightly smaller Total Loss than the RGB-D model.

Offset Loss

For both versions (RGB and RGB-D) the offset loss is relatively low, indicating that the model is performing well in estimating the keypoints' spatial positions. Both models have similar performance in terms of offset loss, with the RGB-D model having a slightly higher offset loss. However, the difference is relatively small.

Centre Loss

Centre loss assesses how accurately the model predicts the centers of objects, in this case being the person in the frame. Similar to offset loss, both versions exhibit low centre loss, suggesting that the model is effective at estimating the centers of keypoints. The RGB-D model under performs the RGB model, with a higher centre loss. This suggests that incorporating depth information has negatively affected the model's ability to predict the center of objects or keypoints. Which is something to keep a note of, as Movenet uses the centrepoint as an anchor to generate predictions for the keypoints.

Heat Map Loss

Both models have relatively similar heatmap losses, indicating comparable performance in this aspect. The slight advantage of the RGB-D model suggests that it produces slightly better heatmap representations. The heatmaps generated by Movenet is of the location of joints and ligaments. This advantage implies that the RGB-D version surpasses the RGB version detecting areas of keypoints.

Regression Loss

Regression loss measures how well the model estimates keypoint coordinates directly. For both versions, regression loss is relatively low, signifying that the model's direct keypoint estimation is accurate. The RGB-D model has a lower regression loss, indicating improved accuracy in predicting keypoint location compared to the RGB model.

Bone Loss

Bone loss quantifies the error in predicting connections (bones) between key points. Both versions show low bone loss, suggesting that the model accurately captures the relationships between key points.Both models have very similar bone losses, suggesting that the incorporation of depth information did not significantly impact the model's performance in this aspect.

Conclusion

In summary, while the RGB-D model incurs slightly higher total losses in some aspects compared to the RGB model. The effect of depth information on the model's performance is nuanced, with trade-offs in different metrics. However, from what can be concluded from the AP (Average precision), and offset losses, is that the RGB-D model shows an improvement in key point detection over the RGB model.

5.1.2 Performance and complexity results

Table 5.2 displays the metric results used to evaluate the models based on precision, complexity, and inference speed. Inference speed was measured when inference was run using an NVidia RTX3090 GPU on the test data set. The RGB-D model's higher mAP score demonstrates its superiority in terms of key point detection accuracy. The depth information extracted from Midas displays its contribution to better understanding the 3D structure of

Model	Offset loss	Centre loss	Heatmap loss	Regression loss	Bone loss	Total Loss
RGB	0.4172	9.536	7.917	2.735	0.3668	21.069
RGB-D	0.4193	10.629	7.734	2.325	0.3475	21.454

Table 5.1: Loss in features extracted of both model and the mean average precision on test dataset

Model	mAP	Model Params	FPS
RGB	74.4%	1912118	130
RGB-D	78.9%	1912406	32

Table 5.2: Complexity metrics as mentioned in 2.4 of both models

key points, leading to improved 2D localization. However, the trade-off for this improved accuracy is a substantial reduction in inference speed, where the RGB model is able to run inference up to 4 times faster than the RGB-D model, which is a significant difference. The RGB-D model operates at a much lower frame rate, which may be acceptable in applications where accuracy is paramount but less suitable for real-time or high-speed scenarios where a 30FPS processing speed would be considered low. Both models have similar complexities, suggesting that the addition of depth information did not lead to a significant increase in model size.

The choice between the RGB and RGB-D models depends on the specific requirements of your application. If you prioritize accuracy and can accommodate slower processing speeds, the RGB-D model is preferable due to its higher mAP score. However, if you need real-time processing or faster inference times and can accept a slightly lower mAP score, the RGB model is a better fit. Ultimately, the decision should align with the specific use case and performance trade-offs that better suit the application's needs.

5.2 Features Inference Analysis

As previously mentioned, the MoveNet model extracts four features, three of which it uses to help it extract the fourth feature, the key points predictions. One can see quite a visible difference between the RGB and the RGB-D models from the sample images in figure 5.2, figure 5.1, and figure 5.3. Generally, both models perform very well when running inference for key point estimation in a frame, however, the results tell us that when depth is introduced, the model has a slight improvement, and after running the test data set, we can see where that advantage is coming into play. The bigger difference starts to show when there are parts of an image that are blurry, have obstructed key points, or several key points that are close to each other. The image frames in figure5.1, figure5.2, and figure5.3 were chosen specifically to display the scenarios of frames that contain rapid movement, and hidden or obstructed key points.

5.2.1 Heat Maps

The RGB-D model generates a more precise key points heat map, which we have also seen from 5.1. This can easily be noticed, precisely in scenarios where the person in the frame is doing a rapid or complicated movement. This also comes into play when the model is running inference on frames where there are other objects near the person or key points. If we were to look at figure 5.2, we can see in the heatmap of the RGB-D model that the silhouette of the arms of the person trying to catch the frisbee is precise and clear, whereas, in the RGB model the silhouette is quite blurry, wherein, a person can't visually make out where the arms of the person are. This can also be seen in figure 5.1, where the image is quite blurry due to the person in the frame doing a rapid movement, in this case, throwing a frisbee. The same statements can be drawn, where the silhouette of the person in the RGB-D model detects and displays the person's key points more accurately, whereas the RGB model returns a blurrier key points heatmap. However, in figure 5.3 we can see the key points heatmaps of both models like very similar, to an extent where you can barely spot a difference, and that is due to the person in the frame being in a static pose.

Both models seem to display more accurate and precise key point heatmaps for the lower part of the body (ankles, knees, and hips) than for the upper part. That was also indicated in figure 5.4, where the models displayed higher correct predictions on the key points in the lower part of the body than the key points in the higher part of the body. That could be due to several reasons, one being that the lower part of the body is usually more static in images, as compared to upper parts of the body like wrists and elbows which are usually more dynamic in images taken in common contexts. Also, the keypoints in the lower part of the body are bigger than the keypoints in the upper part, making them more visible and easier to spot.

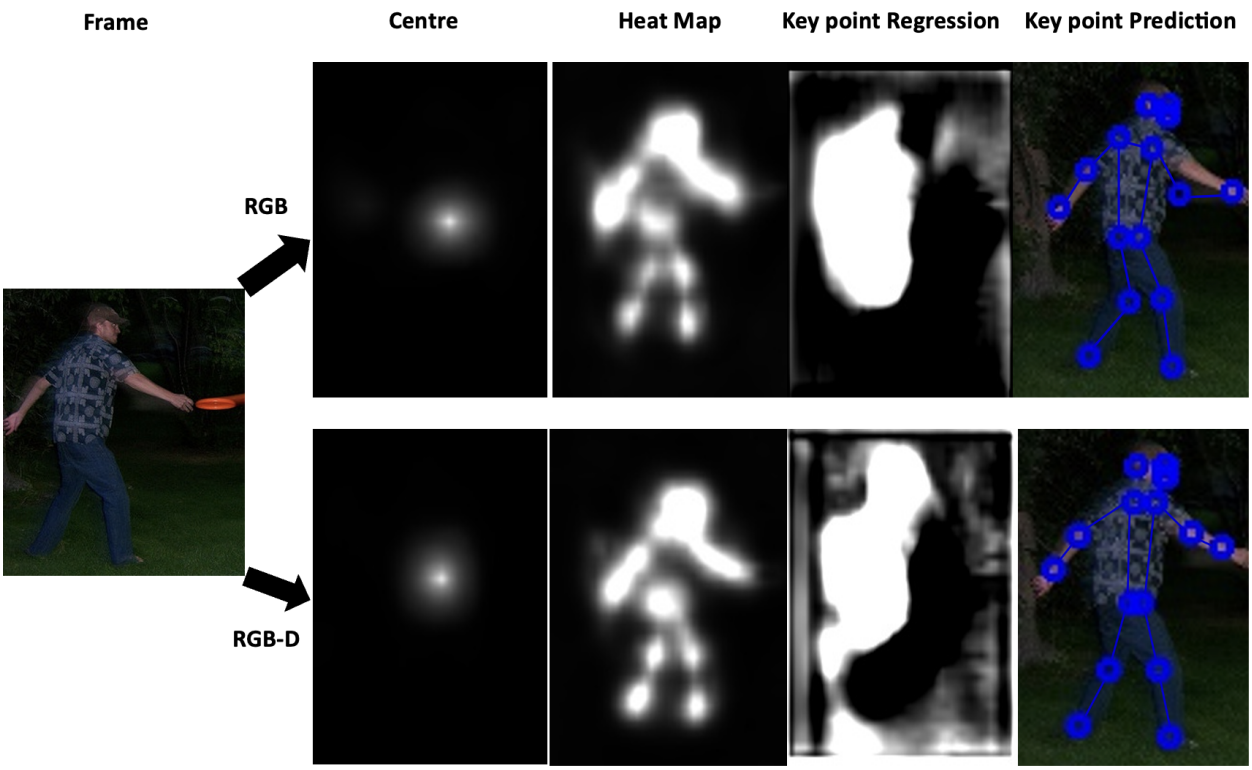

Figure 5.1: Features extracted from image of a person throwing a Frisbee in a dark environment. The Frame appears to be a little blurry around the right wrist.

5.2.2 Regression Maps

The regression maps help the model get an idea of the scale and orientation of the person in the image. It assists in understanding the size and orientation of the object with respect to the centre point. There doesn't seem to be any visible difference between both models in regressions maps generation. That was also hinted at in the regression losses in the table in 5.1. Both models generate correct regressions of the area of the object in the frame, and the sizes of the person in the frame are up to scale with their generated regression.

5.2.3 Key points

A difference can be seen in the images in the key point prediction between both models. The RGB-D model generates better and more accurate key point predictions in scenarios where the person in the frame is executing a rapid movement and has obstructed key points. In figure 5.3, the person in the frame is sitting down on a bench, where his hips are not visible, and are obstructed by his arms and knees. The RGB model struggles to return accurate key points, and that can be seen by falsely predicting the location of the person's right wrist and left knee, however, doesn't miss the predictions by a high margin and returns accurate predictions of all the other key points. On the other hand, the RGB-D model can return more accurate key point predictions. It can also be seen in figure 5.2, where the person has their left arm obstructing their facial key points (nose, eyes, and ears), and the RGB-D model is still able to return correct key point predictions for obstructed joints, and key points that are obstructing other key points. The RGB-D model also returns accurate predictions for the wrists, which are the dynamic key points in the frame. On the other hand, the same can't be said for the RGB model, where it misses the dynamic key points as it miss-predicts the location of both wrists.

5.3 Key points predictions

5.3.1 Key Observations

Generally, the RGB-D model seems to have a slight advantage in terms of the number of correct predictions compared to the RGB only model. This suggests that incorporating depth information (RGB-D) has a positive impact on key point estimation, as it provides additional

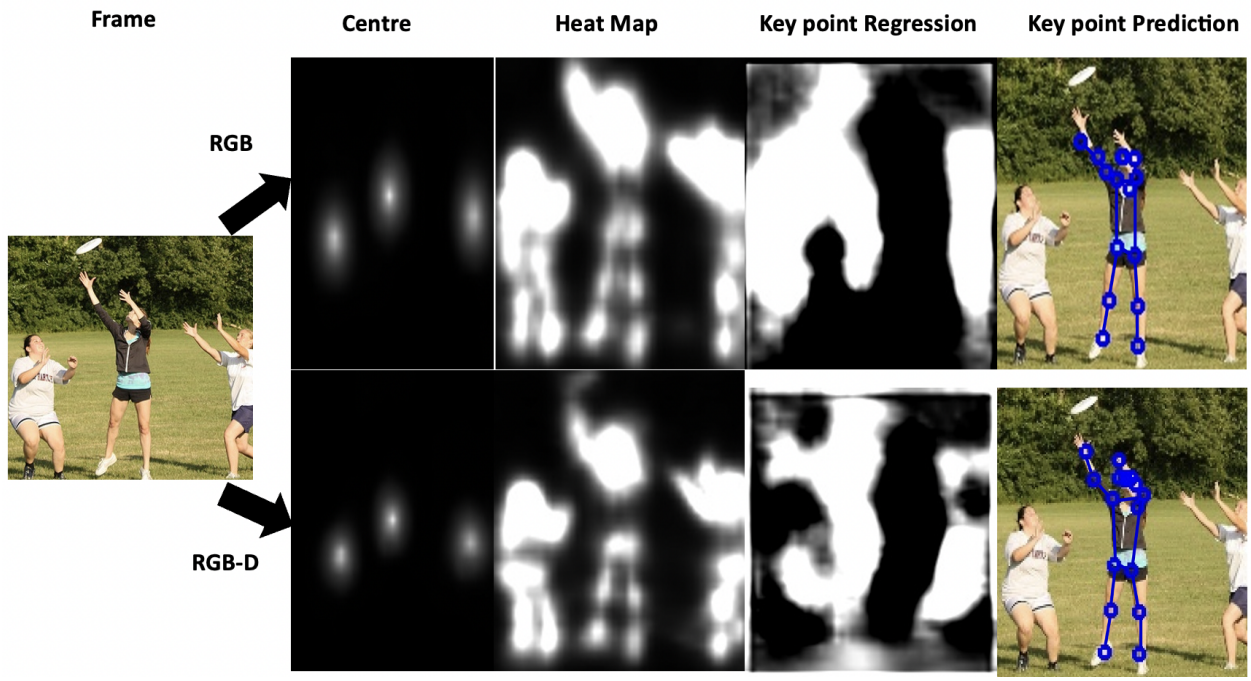

Figure 5.2: Features extracted from image of a person catching a Frisbee while jumping. Image taken in an outdoors environment with multiple people in the frame. The person's facial key points are occluded by their arm.

information that helps in understanding the 3D structure of the human body.

There is some variance in the performance of each key point, with some key points having higher correct predictions than others. This is expected and can be influenced by factors such as the visibility of key points in the images, their relative complexity, and their spatial relationships with other key points.

The models seem to perform similarly on key points located on both sides of the body (e.g., left and right eye, left and right wrist), indicating that the models are not biased toward one side.

Some key points appear to be more challenging for both models. For example, the key points associated with the ears (L-Ear and R-Ear) have lower correct predictions compared to others. This could be due to factors such as occlusion, pose variations, or the difficulty of detecting ear key points in certain images.

5.3.2 Evaluation of Results

The RGB-D model performs slightly better overall, which suggests that incorporating depth information is beneficial for human pose estimation. The depth information helps in better understanding the 3D structure of the human body, resulting in more accurate key point predictions. The RGB-only model still performs reasonably well, especially given that it doesn't have access to depth information. However, it lags slightly behind the RGB-D model in terms of key point accuracy. While both models perform well, there is always room for improvement. Further fine-tuning, architectural adjustments, or the incorporation of additional data or pre-processing techniques may lead to even better performance. The choice between the RGB and RGB-D models should depend on the specific application and the importance of depth information. If depth information is critical for accurate pose estimation, then the RGB-D model is preferable.

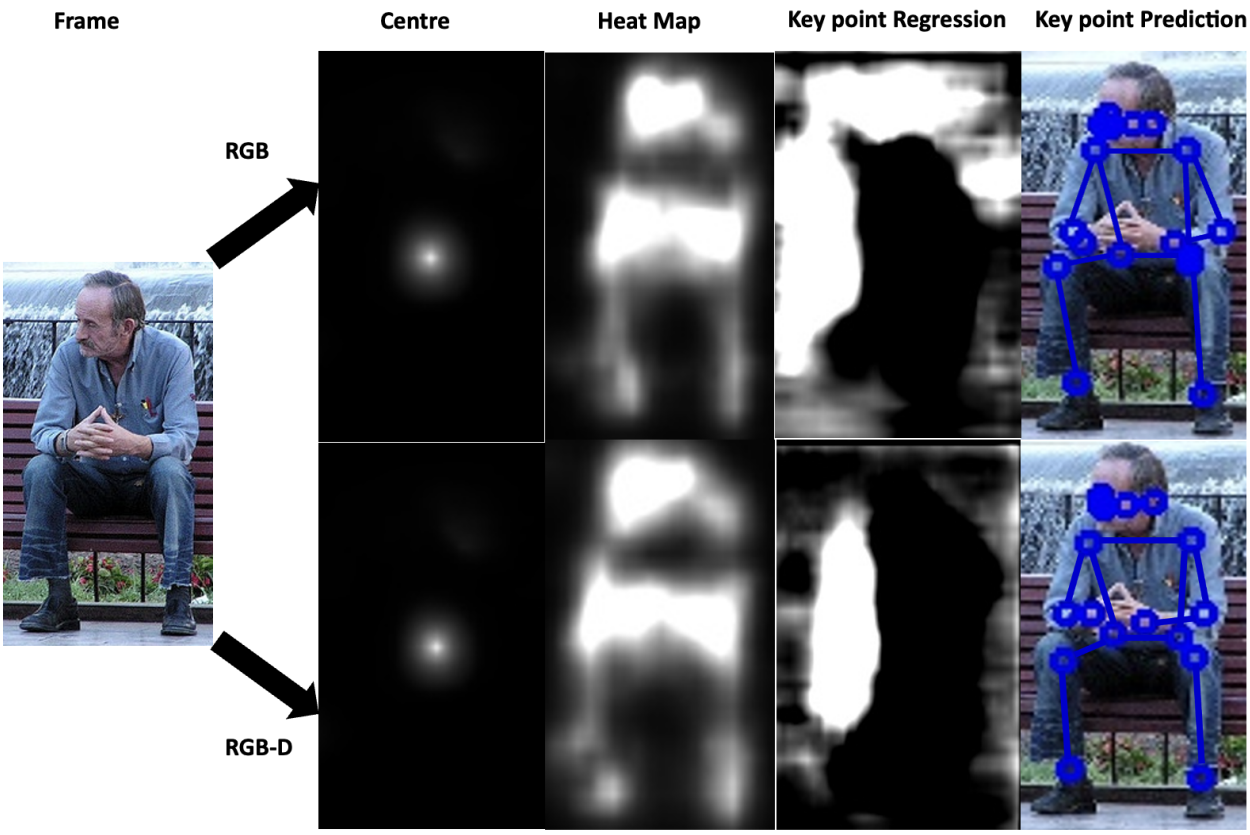

Figure 5.3: Features extracted from image of a person sitting down on a bench.

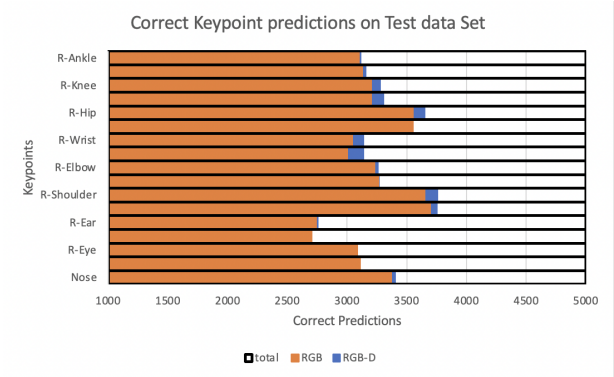

Figure 5.4: Graph from appendices A.2 displaying results of correct Key point predictions for both models on each key point.

Chapter 6

Conclusions

In this project, we trained two variants of the MoveNet human pose estimation model using different data sources: one on RGB-D data and the other solely on RGB data. Depth data was extracted through multi-modal means, where the Midas monocular depth estimation model was used. We were successfully able to demonstrate that depth data can aid a human pose estimation model in creating more accurate predictions, as the RGB-D model exhibited improved key point prediction accuracy, particularly in challenging scenarios involving blurry frames or obstructed key points which are considered as some of the big challenges that are faced in human pose estimation. However, it came at the heavy cost of significantly reduced processing speed due to the incorporation of depth information obtained from the Midas monocular depth estimation model.

The RGB-D model demonstrated a slight but noteworthy improvement in key point prediction accuracy compared to the RGB model. This is especially valuable in real-world scenarios where detecting obscured or ambiguous key points is essential. The RGB-D model exhibited enhanced robustness when faced with challenging conditions, such as blurriness or occlusions. We were also able to visualise the effect that depth has had on the features extracted by the MoveNet model, where we noticed how the key points heatmaps produced by the RGB-D model were much more defined and clearer. This makes the RGB-D model more suitable for applications where accurate pose estimation in complex scenes is crucial.

While the RGB-D model excelled in accuracy and robustness, it suffered from significantly slower processing speeds. The RGB model was able to process 100 frames more per second than the RGB-D, which is a significant difference that shows us that depth introduction through multi-modal learning with Midas, has made the model's processing speed up to 4 times slower. This makes the RGB-D model less suitable for real-time or high-speed applications which limits its application and use significantly.

6.1 Aims and Objectives

The project successfully achieves the aim of designing a robust robust and a more accurate human pose estimation model that is able to better deal with the challenges of human pose estimation. We've seen our modified able to better handle blurry frames, occlusion, and multi-person frames. However, we've seen the cost of such advantage, as it was at the cost of a significantly lower processing speed, which adds weight onto the question of the model's ability to generate real-time predictions. It also limits the model in its application scenarios.

6.2 Further Improvements

There is much room for improvement and many different methods can be applied that could potentially improve the performance of the RGB-D version of the MoveNet model even further. To start with, a lighter version of the monocular depth estimation model Midas could be used. The version used in this project is Midas v3, which is the largest version of the Midas model, it has the highest accuracy, but the slowest inference speeds out of all 3 models. Depending on the criticality of inference speed, a lighter model could be used at the price of high accuracy. Another consideration that would impact inference speed heavily, would be to consider a possible hybrid approach that dynamically switches between RGB and RGB-D models based on the scene complexity or the availability of high-quality depth information. This way, the model can maintain real-time performance when possible while benefiting from improved accuracy in challenging situations.

Another approach that would make the model much more robust and efficient, would be to train the model on data that has a combination of RGB-only and depth data. That way the model can have some depth ground truth points to be able to calculate and account for the loss for depth estimation and to manipulate its weights and bias in the network. Wherein, in the model trained

for this project, the model has relied on the optimal hypermeters retrieved from the original training of the Midas model to retrieve depth. A method that was also introduced in [30].

Bibliography

[1] Andriluka, M., Pishchulin, L., Gehler, P. and Schiele, B., 2014. 2d human pose estimation: New benchmark and state of the art analysis. *Ieee conference on computer vision and pattern recognition (cvpr)*.

[2] Barnum, G., Talukder, S. and Yue, Y., 2020. On the benefits of early fusion in multimodal representation learning. *arxiv preprint arxiv:2011.07191*.

[3] Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S. and Zhang, L., 2020. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. pp.5386–5395.

[4] Croft, H., 2022. *Addressing the problem of 'big data'in sports: A framework for performance analysts*. Ph.D. thesis. Auckland University of Technology.

[5] Fieraru, M., Khoreva, A., Pishchulin, L. and Schiele, B., 2018. Learning to refine human pose estimation. *Proceedings of the ieee conference on computer vision and pattern recognition workshops*. pp.205–214.

[6] Ganapathi, V., Plagemann, C., Koller, D. and Thrun, S., 2012. Real-time human pose tracking from range data. *Computer vision–eccv 2012: 12th european conference on computer vision, florence, italy, october 7-13, 2012, proceedings, part vi 12*. Springer, pp.738–751.

[7] Geiger, A., Lenz, P., Stiller, C. and Urtasun, R., 2013. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11), pp.1231–1237.

[8] Girshick, R., 2015. Fast r-cnn. *Proceedings of the ieee international conference on computer vision*. pp.1440–1448.

[9] He, K., Gkioxari, G., Dollár, P. and Girshick, R., 2017. Mask r-cnn. *Proceedings of the ieee international conference on computer vision*. pp.2961–2969.

[10] Ibbett, R.N., Edwards, D.A., Hopkins, T., Cadogan, C. and Train, D., 1985. Centrenet–a high performance local area network. *The computer journal*, 28(3), pp.231–242.

[11] Jo, B. and Kim, S., 2022. Comparative analysis of openpose, posenet, and movenet models for pose estimation in mobile devices. *Traitement du signal*, 39(1), p.119.

[12] Johnson, S. and Everingham, M., 2010. Clustered pose and nonlinear appearance models for human pose estimation. *bmvc*. Aberystwyth, UK, vol. 2, p.5.

[13] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., 2014. Microsoft coco: Common objects in context. *Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014, proceedings, part v 13*. Springer, pp.740–755.

[14] Marin-Jimenez, M.J., Romero-Ramirez, F.J., Munoz-Salinas, R. and Medina-Carnicer, R., 2018. 3d human pose estimation from depth maps using a deep combination of poses. *Journal of visual communication and image representation*, 55, pp.627–639.

[15] Mees, O., Eitel, A. and Burgard, W., 2016. Choosing smartly: Adaptive multimodal fusion for object detection in changing environments. *2016 ieee/rsj international conference on intelligent robots and systems (iros)*. IEEE, pp.151–156.

[16] Newell, A., Yang, K. and Deng, J., 2016. Stacked hourglass networks for human pose estimation. *Computer vision–eccv 2016: 14th european conference, amsterdam, the netherlands, october 11-14, 2016, proceedings, part viii 14*. Springer, pp.483–499.

[17] Ning, G., Liu, P., Fan, X. and Zhang, C., 2018. A top-down approach to articulated human pose estimation and tracking. *Proceedings of the european conference on computer vision (eccv) workshops*. pp.0–0.

[18] Osokin, D., 2018. Real-time 2d multi-person pose

estimation on cpu: Lightweight openpose. *arxiv preprint arxiv:1811.12004*.

[19] Ouyang, W., Chu, X. and Wang, X., 2014. Multi-source deep learning for human pose estimation. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.2329–2336.

[20] Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J. and Murphy, K., 2018. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. *Proceedings of the european conference on computer vision (eccv)*. pp.269–286.

[21] Park, S., Ji, M. and Chun, J., 2018. 2d human pose estimation based on object detection using rgb-d information.

[22] Piao, Y., Rong, Z., Zhang, M., Ren, W. and Lu, H., 2020. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. pp.9060–9069.

[23] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K. and Koltun, V., 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *Ieee transactions on pattern analysis and machine intelligence*, 44(3), pp.1623–1637.

[24] Ronny Votel, N.L., 2021. Next-generation pose detection with movenet and tensorflow.js. *Tensor flow blog*.

[25] Safdarnejad, S.M., Liu, X., Udpa, L., Andrus, B., Wood, J. and Craven, D., 2015. Sports videos in the wild (svw): A video dataset for sports analysis. *2015 11th ieee international conference and workshops on automatic face and gesture recognition (fg)*. IEEE, vol. 1, pp.1–7.

[26] Sun, K., Xiao, B., Liu, D. and Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. pp.5693–5703.

[27] Toshev, A. and Szegedy, C., 2014. Deeppose: Human pose estimation via deep neural networks. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.1653–1660.

[28] Wei, S.E., Ramakrishna, V., Kanade, T. and Sheikh, Y., 2016. Convolutional pose machines. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.4724–4732.

[29] Wei, S.E., Ramakrishna, V., Kanade, T. and Sheikh, Y., 2016. Convolutional pose machines. *Proceedings of the ieee conference on computer vision and pattern recognition*. pp.4724–4732.

[30] Xiang, M., Zhang, J., Lv, Y., Li, A., Zhong, Y. and Dai, Y., 2021. Exploring depth contribution for camouflaged object detection. *arxiv preprint arxiv:2106.13217*.

[31] Zhou, Y., Dong, H. and El Saddik, A., 2020. Learning to estimate 3d human pose from point cloud. *Ieee sensors journal*, 20(20), pp.12334–12342.

Appendix A

Design Diagrams

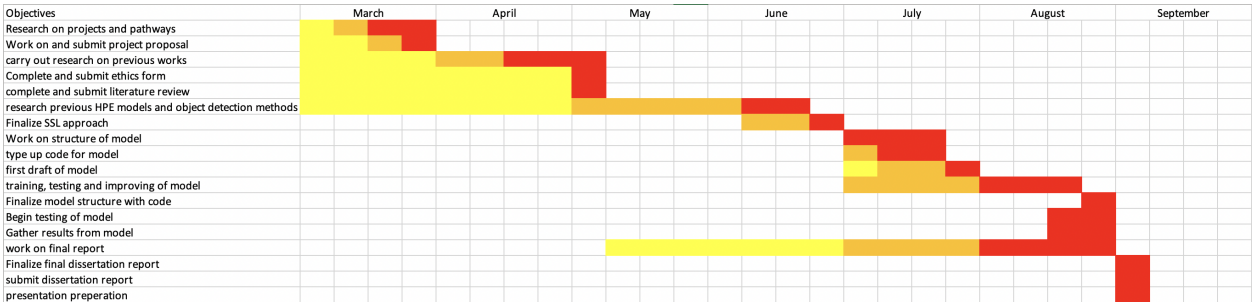

Figure A.1: Gantt chart displaying plan of project progression

A.1 Gantt Chart

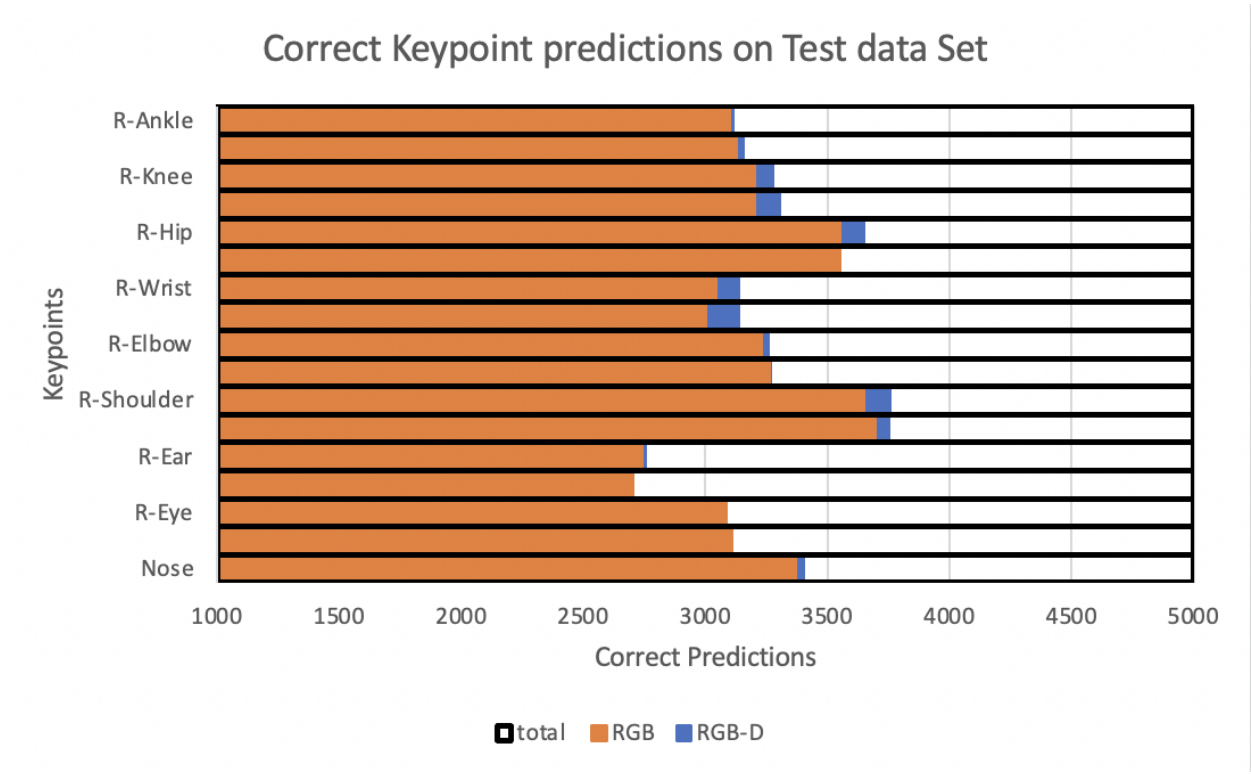

Figure A.2: Graph displaying results of correct Key point predictions for both models on each key point

A.2 Key points predictions results graph