The Automation of Camera Trap Distance Sampling with Machine Learning for the Estimation of Population Density and Abundance

Author:
TOM RAYNES

DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF BRISTOL

A thesis submitted to the University of Bristol in accordance with the requirements of the degree of MSc COMPUTER SCIENCE

SEPTEMBER 2025

Abstract

[Abstract]

Acknowledgements

[Acknowledgements]

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: .....

Table of Contents

List of Tables

List of Figures

1 Introduction

Manual distance sampling bottleneck

Aim to automate distance sampling of a large dataset and use estimated distances to achieve accurate estimates for population activity

WCF dataset

How abundance and density is calculated using distances

2 Background

2.1 WCF Dataset

2.2 Methods and Models

2.2.1 Deep Learning

2.2.2 Mega Detector

2.2.3 Segment Anything

2.2.4 Depth Anything

2.2.5 Dense Prediction Transformers

2.2.6 Calibrating Distances

2.3 Estimating Activity

Distance estimation[1]

3 Experimantal

3.1 Calibration Frame Preparation

For each camera location of the dataset, a set of calibration frames (generally fifteen, location dependant) and corresponding binary masks were prepared. The frames were extracted from location-specific reference videos supplied with the dataset consisting of a person standing, facing the camera, at known distance intervals and holding a sheet of A4 paper labelled with the corresponding distance.

3.1.1 Frame Extraction

In an effort to streamline the frame extraction process, an extractor program was created (Figure 3.1). This tool enabled reference videos to be dragged and dropped into a GUI window allowing for the easy navigation between individual frames to extract the optimal one for each distance. With this software, the next/previous frames are accessed with the 'left'/'right' arrow keys, the frame index position is moved ahead/behind 20 places with the 'd'/'a' keys and the frame found at a specific timestamp is accessed with the 's' key followed by entering a time (in seconds) into an input bar. The active frame (at the current index) can be saved to disk with the 'enter' key followed by entering a filename in an input bar.

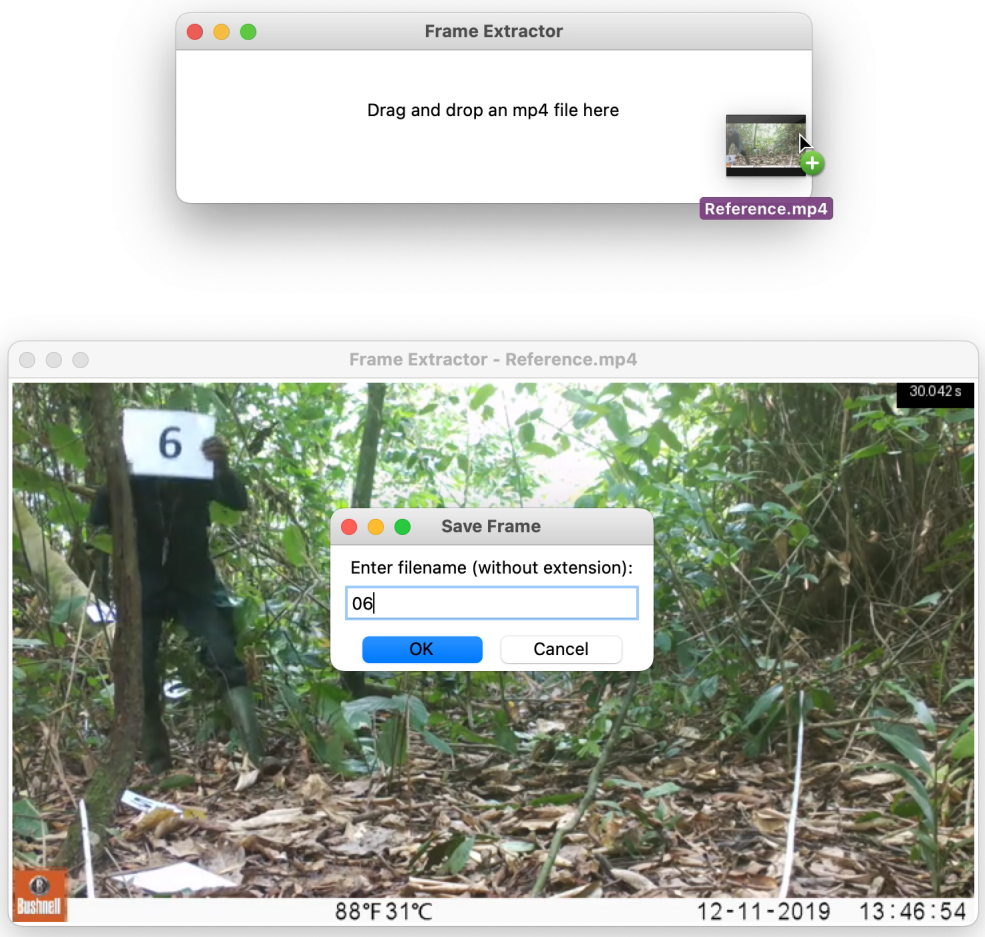

Figure 3.1: Frame extractor program showing the drag/drop (top) and save frame (bottom) features

3.1.2 Frame Mask Creation

For each of the extracted calibration frames, a corresponding binary mask was created. This task can be completed manually using image manipulation software (e.g. Photoshop, GIMP, etc...) where the segmentation boundary is manually traced then filled. This approach, however, is rather labour-intensive, constituting somewhat of a bottleneck, and therefore it is desirable to automate the process.

To assess the feasibility of automation, a two-stage detection/segmentation pipeline was created. Here, raw calibration frames are first processed with YOLOv5 detector[2] to generate bounding boxes (prompted to detect humans) enclosing the frame landmarks. The bounding boxes are then passed to Segment Anything Model which predicts segmentation masks for the landmarks.

The efficacy of this automated method of calibration frame mask generation was tested on several raw calibration frames. Two examples of the raw calibration frames (Figure 3.2) along with their corresponding manual (Figure 3.3) and automated (Figure 3.4) masks are shown below.

Figure 3.2: Raw calibration frames

Figure 3.3: Manual calibration frame masks

Figure 3.4: Automated calibration frame masks

3.2 Detection Frame Preparation

Detection frame preparation is the process whereby the raw camera trap video of the dataset is transformed into an array of representative images depicting the chimpanzees whose distances to the camera will be estimated. Two distinct frame sampling techniques were used in this study which will be referred to as 'manual' and 'automated' sampling.

3.2.1 Sampling

Manual Sample

Accompanying the raw camera trap video, another component of the dataset is a list of human-annotated distance-to-camera estimates for all observed chimpanzees at a recorded date and time (along with additional metadata). These distance data must be used a benchmark by which the accuracy of any modelled estimates are assessed and therefore the specific frames corresponding to these manual annotations must be sampled.

Before sampling, these data were first cleaned by running an automated script to flag any identifiable abnormalities in the data such as duplicate entries and inconsistencies in the date/time formatting. These were then manually corrected.

In order to identify the correct frames to extract, each of the manual annotations was assigned a timestamp corresponding to a time in seconds in which it was recorded in its associated camera trap video. These timestamps then constituted discreet identifiers of sampled frames.

Although the start-date/times of the videos themselves were not explicitly pre-labelled within the dataset, assigning timestamps was still possible since the date/time at zero seconds into a given video can be inferred as equal to that of the earliest of all recorded annotations associated with the video. This was a valid assumption to make as the overwhelming majority of videos begin exactly at the point in which a chimpanzee enters the frame, thus corresponding to the first annotated observation. In the few rare cases where this heuristic did not apply, all relevant timestamps were manually corrected upon being flagged.

Automated Sample

The automated sampling process is significantly more straightforward. Here, detection frames are sampled at an interval of two seconds over all camera trap video in the dataset. Resultant model-estimated distances from this sample are intended to be used purely as distance data for the automated modelling of population abundance.

In contrast to the manual sample, this automated sample is not directly associated with any human-annotated distance estimates. It does, however, remain representative of the dataset, capturing images of the same individual chimpanzees over the same time period. This sample also differs due to the sampling of many 'empty' frames where the image captures no individual. The impact of these empty frames on subsequent abundance estimates will be minimal. The empty frames may collectively incur a small number of false positive detections; however, the total will be negligible and the probability of a given empty frame resulting in a false positive detection is no more likely than that of any other frame.

All in all, this approach exemplifies a simple frame sampling method that can be used as part of an automated distance sampling pipeline.

3.2.2 Frame Extraction

The identified sample frames were extracted from the raw camera trap video using an automated script. Examples of extracted detection frames capturing chimpanzees (Figure 3.5) as well as empty detection frames (Figure 3.6) are shown below.

Figure 3.5: Mixed subsample of extracted detection frames capturing chimpanzees (taken from both the manual and automated sample)

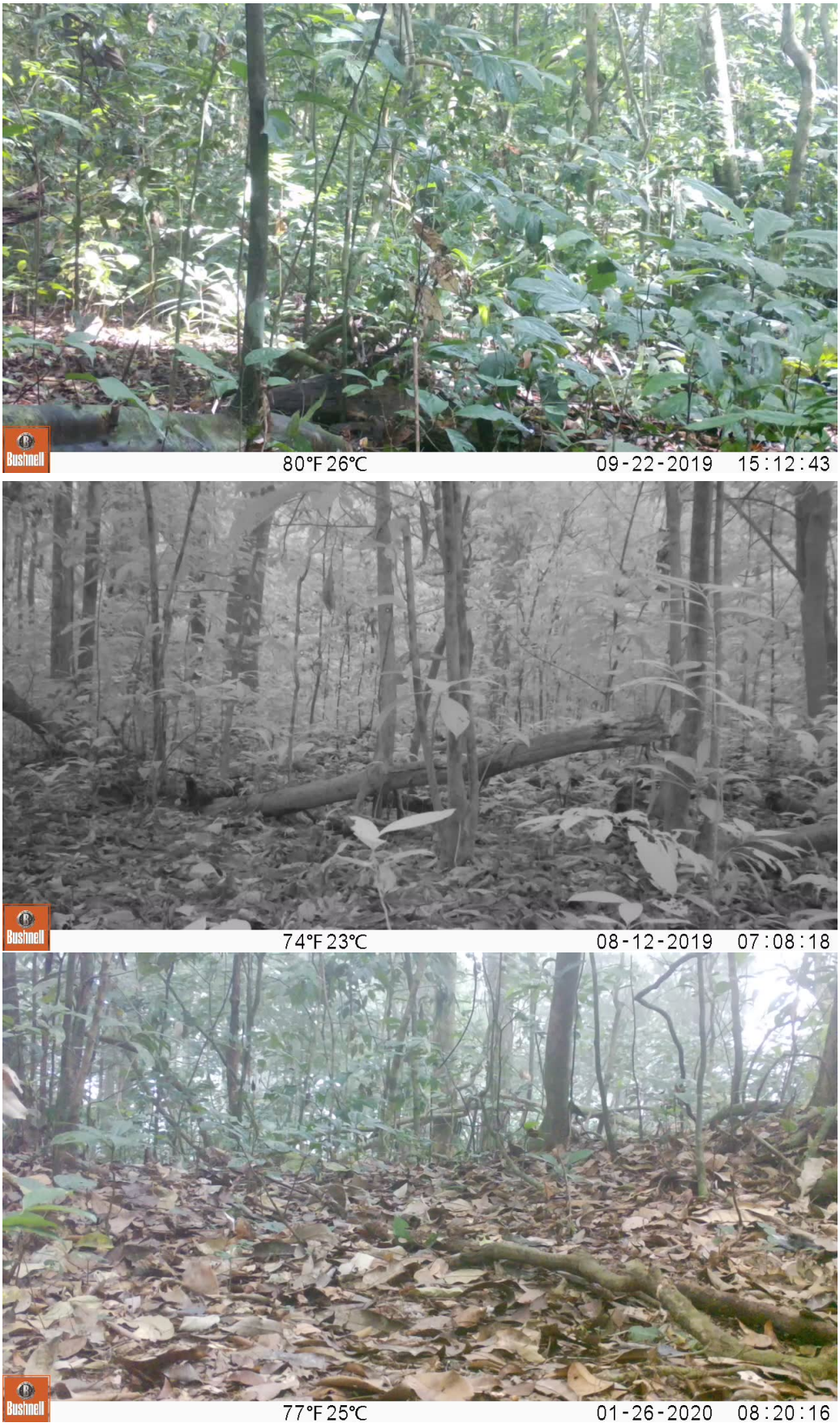

Figure 3.6: Mixed subsample of empty detection frame
(taken exclusively from the automated sample)

3.3 Distance Estimation

3.3.1 Setup

BlueCrystal4 was used to run the distance estimation pipeline across the dataset. Minor adjustments to the load_model() methods of the DPT and DepthAnything classes were made in order to configure ONNX Runtime to utilise all CPU cores allocated by SLURM (28 in this case).

It is also worth noting that for these distance estimation runs, the manual calibration frame masks were used. The reason for this was to maximise the consistency of the masks due to a drop in the quality of the automated masks at longer distances (see Section 4.1).

3.3.2 Configurations

Four distinct configurations of the distance estimation pipeline were applied to the dataset. Each configuration combined one of two detection methods (bounding box, segmentation) with one of two distance estimation models (DPT, Depth Anything). Using each of these configurations, distance estimates for both the manually and automatically sampled detection frames were computed.

Detection Methods

Both detection methods use Mega Detector as a starting point to generate bounding boxes enclosing the animal.

For the plain bounding box detection method, all pixels enclosed in the bounding boxes are used as candidates during the later-stage calculation of detection distance from the depth maps generated by either DPT or DepthAnything. Specifically, once the depth estimation model has predicted a distance for each pixel, the distance estimate for the entire detection is calculated to be equal to that of the pixel corresponding to the 20th percentile depth of all pixels within the bounding box.

The segmentation method, however, introduces an additional step in which the generated bounding boxes are passed to Segment Anything Model which subsequently predicts segmentation masks for the detection. It is these segmented pixels exclusively that are used in the later detection distance calculation. In contrast to the plain BBOX detection method where distance is calculated based on percentile, the segmentation method detection distance is calculated as equal to the distance estimate of the pixel corresponding to the centre-most point of the segmentation mask (i.e., the pixel that is furthest from the segmentation boundary).

Additionally, all runs using the segmentation detection method were parameterised with the 'calibration mask animals' flag. This has the following effect: During the detection calibration alignment where the scale of the furthest calibration frame reference target is transferred to the detection frame, only the pixels inside the animal segmentation boundary (rather than the entire bounding box) are excluded. This is most impactful at very close detection distances where the animal occupies a significantly large area of the frame, allowing for a more accurate alignment to be achieved.

3.3.3 Testing Calibration Modifications

The effects altering the calibration stage of the distance estimation pipeline were also tested. Calibration frame preparation comprises significant manual annotation resulting in a bottleneck, therefore it would be a positive outcome if this stage could be streamlined with a reduction in calibration requirements.

Exclusive Close/Far Reference Calibration

Firstly, a modification was tested over the whole dataset with DPT distance estimation using both bounding box and segmentation detection methods. Here, the program's run() method was altered to ignore all calibration frames except those corresponding to closest and furthest landmarks from the camera.

Variable Calibration

Secondly, the effects of varying the number of calibration frames was tested on a single camera location. A total of fourteen runs were carried out using DPT distance estimation with both bounding box and segmentation detection methods. The first of these used only the 1 and 2 meter calibration frames with each subsequent run then including the next calibration frame in the sequence (i.e., 1–2, 1–3, 1–4 . . . 1–15).

3.4 Activity Estimation

Using the computed distance estimates from the previous section, chimpanzee population density and abundance was estimated. Estimates were calculated for each of the detector/depth model combinations using a script adapted from one supplied with the dataset.

3.4.1 Formatting

For the distance data to be used for density and abundance estimation, it first must be formatted and joined with the relevant location and camera trap video metadata as to satisfy the script's input requirements. The final must then be identical in structure to the manually annotated detection data supplied with the dataset in CSV formatting containing the columns 'video_name', 'time_in_place_(days)', 'effort', 'starting_date', 'date_observation', 'time_observation', 'year', 'month', 'day', 'hour', 'minute', 'second', 'distance', 'Sample.Label', 'Effort', 'Region.Label' and 'Area'.

Manual Sample

When formatting the distance data generated from the manual sample, the goal was to overwrite each of the manually annotated distances (in original supplied dataframe) with the corresponding model estimated distances, leaving all other data unchanged.

For reasons detailed in Section 4.2.1, a decision was made to format the data such as to only overwrite manual distances associated with frames capturing a single chimpanzee. This gave a 'supplemented' dataframe, where distance estimates originating from single-chimp frames were model-estimated while those from multi-chimp frames were estimated manually. Nevertheless, density and abundance estimates based on the supplemented data are still informative and give insight into the effectiveness and accuracy of using model estimated distance data in this context.

Automated Sample

When formatting the distance data generated from the automated sample however, a different method was used. First, a lookup table was created to hold the metadata associated with each specific camera trap video. A script was then run to join each of the modelled distances with the video metadata on video name which gave a formatted dataframe holding entirely model estimated distances.

3.4.2 Using Activity Script

With the distance data and associated metadata now appropriately formatted, density and abundance estimates were then developed using the activity script. For each dataframe, the script was adapted to apply appropriate truncation and binning to the distance data in order to generate a smooth probability density histogram. Preliminary runs of the script were then used to fit several detection functions to this histogram, with the best fitting (lowest AIC value) being selected.

4 Analysis

4.1 Analysis of Calibration Frames

The figures below show both the manual and automated calibration frame masks for three (Figure 4.1) and ten (Figure 4.2) meters (seen in section 3.1.2) superimposed atop their corresponding raw calibration frame.

Figure 4.1: Three meter manual (top) and automated (bottom) calibration frame segmentation masks overlayed over the raw calibration frame

Figure 4.2: Ten meter manual (top) and automated (bottom) calibration frame segmentation masks overlayed over the raw calibration frame

At distances where the calibration landmark (human) is fully captured within the frame (Figure 4.1), the automated approach generally yields masks with a segmentation quality on a comparable level to those produced manually. This pixel-perfect segmentation, however, does not hold for all calibration frames, especially at longer distances. For longer-distance frames (Figure 4.2), the landmarks occupy fewer pixels and, most importantly, become broken up into many segments due to occlusion from the dense foliage found in this environment. These factors result in the loss many of the pixel-level cues that Segment Anything Model uses to derive a segmentation boundary, leading to calibration masks that display edge-leakage and the masking of pixels that are not themselves part of the landmark.

Evident from Figures 4.1 and 4.2, the automated segmentations exclude all pixels corresponding the non-human components of the landmarks (i.e., that distance labeled sheets of paper). This is undesirable for the calibration process since the depth scaling of the calibration frames should be aligned based on only the transect environment features (i.e., the pixels outside the landmark segmentation that are common to all the calibration frames).

Additionally, calibration frames associated with the furthest of distances (i.e., approaching fifteen meters), yielded no segmentations at all. This is due to a failure to localise the landmarks in these frames by YOLOv5 detector. This is an unsurprising result given the nature of the environment at the Taï National Park. At these distances, landmark occlusion leads to extremely subtle pixel-level detection cues and therefore great difficulty of identification even for a human. An example of this failure case is shown in Figure 4.3.

Ultimately, these factors allude to the manual method of calibration frame mask creation constituting a better approach. Despite increased time and labour requirements, it is of greater value to achieve a more accurate calibration than to sacrifice mask quality in order to save time during calibration frame preparation.

Figure 4.3: Example of a calibration frame (top) where landmark localisation with the automated masking method failed. The corresponding manual binary mark is shown on the bottom

4.2 Analysis of Distance Estimates

4.2.1 Model / Manual Distance Comparison

In this section, the precision and accuracy of the distance estimates generated using the four configurations of the distance estimation pipeline (outlined in Section 3.3.2) are evaluated. In order for these data to be analysed, the estimates are benchmarked against their corresponding manual distance estimates (supplied with the dataset), thus these estimates use detection frames originating from the manual sample (Section 3.2.1).

Ideally, each and every manual distance should be joined to its corresponding modelled distance estimate; however, due to the absence of any frame-position data associated with the manual annotations, automating this using traditional algorithms is impossible in circumstances where multiple chimps are captured in a single frame. This is because when multiple individuals are detected, there is ambiguity in regard to which distance a given modelled distance should be joined to. Moreover, approaching this task manually is extremely labour-intensive and was therefore outside the scope of this project. As a result, this analysis focuses on distance comparisons associated with frames capturing only a single individual.

Upon joining the modelled distance estimates to their corresponding manual estimates, the overall errors of each of the four pipeline configurations were calculated. Table 4.1 shows the mean average error, root mean squared error, average difference (i.e., mean($model_i$ - $manual_i$)) and the result of a statistical difference test (i.e., where p-value < 0.05). To note, while both the mean average and root mean squared errors give a measure of the absolute error of the configurations, the average difference does not and is affected by cancellation between overestimates and underestimates. Therefore, it is used as a metric to assess whether a given configuration over or under-predicts relative to the manual distance estimation method, where a positive value indicates over-prediction while a negative value indicates under-prediction.

The modelled distance estimates were then grouped by their corresponding manual estimates (i.e., 0.5 m, 1.0 m, ... 15 m). Figure 4.4 shows the averages of the modelled estimates for each group while Figure 4.5 shows each individual modelled estimate for each group, giving a visualisation of the spread of the estimates. Table 4.2 shows the gradients of the fitted regression lines correlating the averages of all grouped distance estimates with their corresponding manual distance estimate for each configuration. These individual errors for each of these groups were then calculated. Figure 4.6 show the change in mean average error for each group while Figure 4.7 shows the change in root mean squared error for each group.

Table 4.1: Mean average error (MAE), root mean squared error (RMSE), average difference between model and manual estimate ($\Delta_{average}$) and result of Wilcoxon signed rank test for statistical difference. These data describe all distance estimates for each pipeline configuration collectivelly.)

Method	MAE / m	RMSE / m	$\Delta_{average}$ / m	Statistical Difference
DPT, BBOX	1.81	2.66	0.586	YES
DPT, SEG	1.70	2.45	0.836	YES
DA, BBOX	2.03	2.62	1.49	YES
DA, SEG	3.00	3.52	2.80	YES

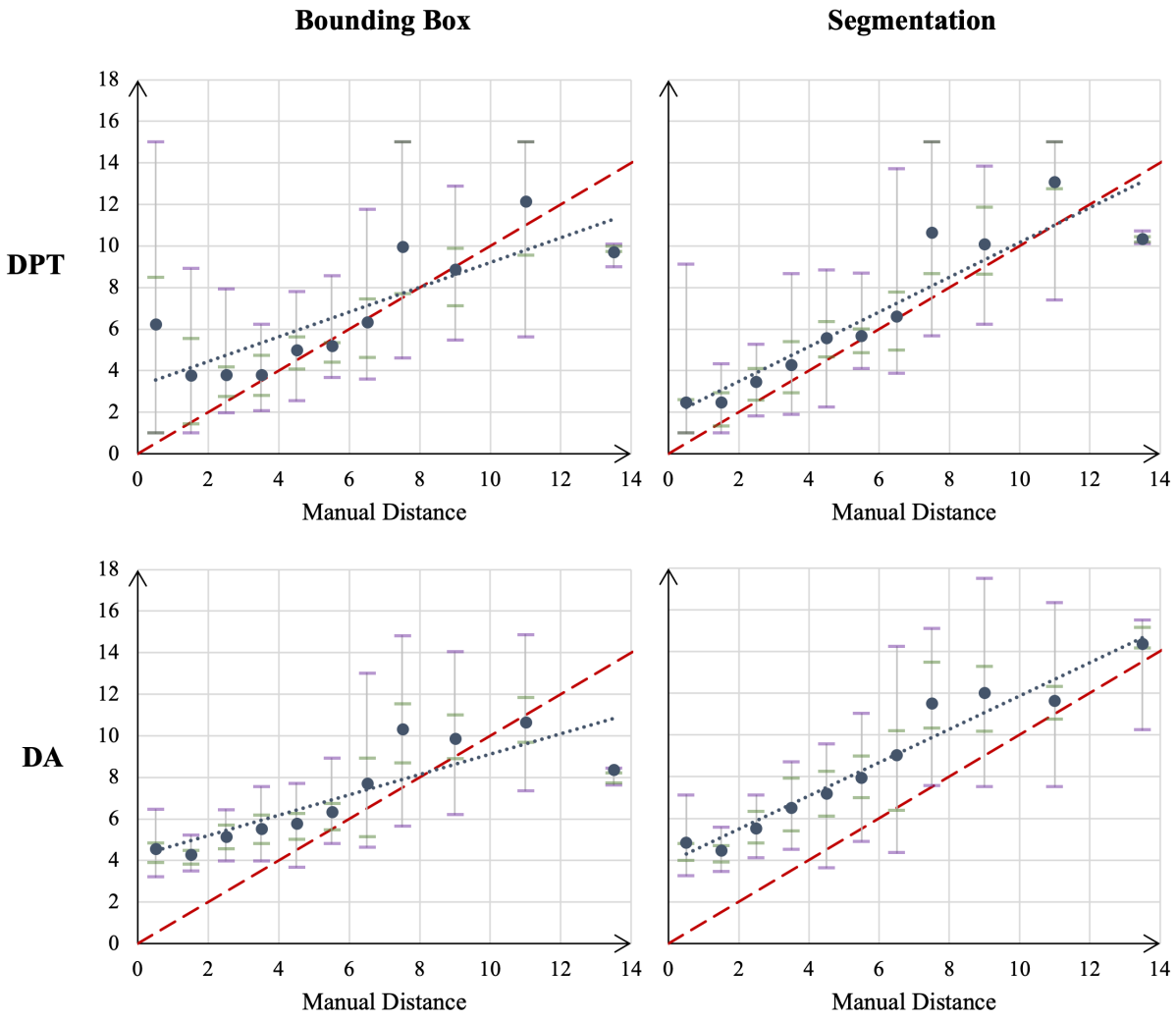

Figure 4.4: Graphs showing mean modelled distance estimates mapped to thier corresponding manual estimates for the configurations: DPT/bounding box (top-left), DPT/segmentation (top-right), DA/bounding box (bottom-left) and DA/segmentation (bottom-right). The blue dotted line shows the fitted regression line. The red dashed line shows the ideal (i.e., model=manual). the error bars show the 25–75 (green) and 5–95 (purple) percentiles. All distances are in units of meters.

Table 4.2: Gradients of the regression lines (blue dotted lines in Figure 4.4) correlating the averages of all binned distance estimates with their corresponding manual distance estimate for each configuration.

Method	Regression Gradient
DPT, BBOX	0.59
DPT, SEG	0.84
DA, BBOX	0.49
DA, SEG	0.80

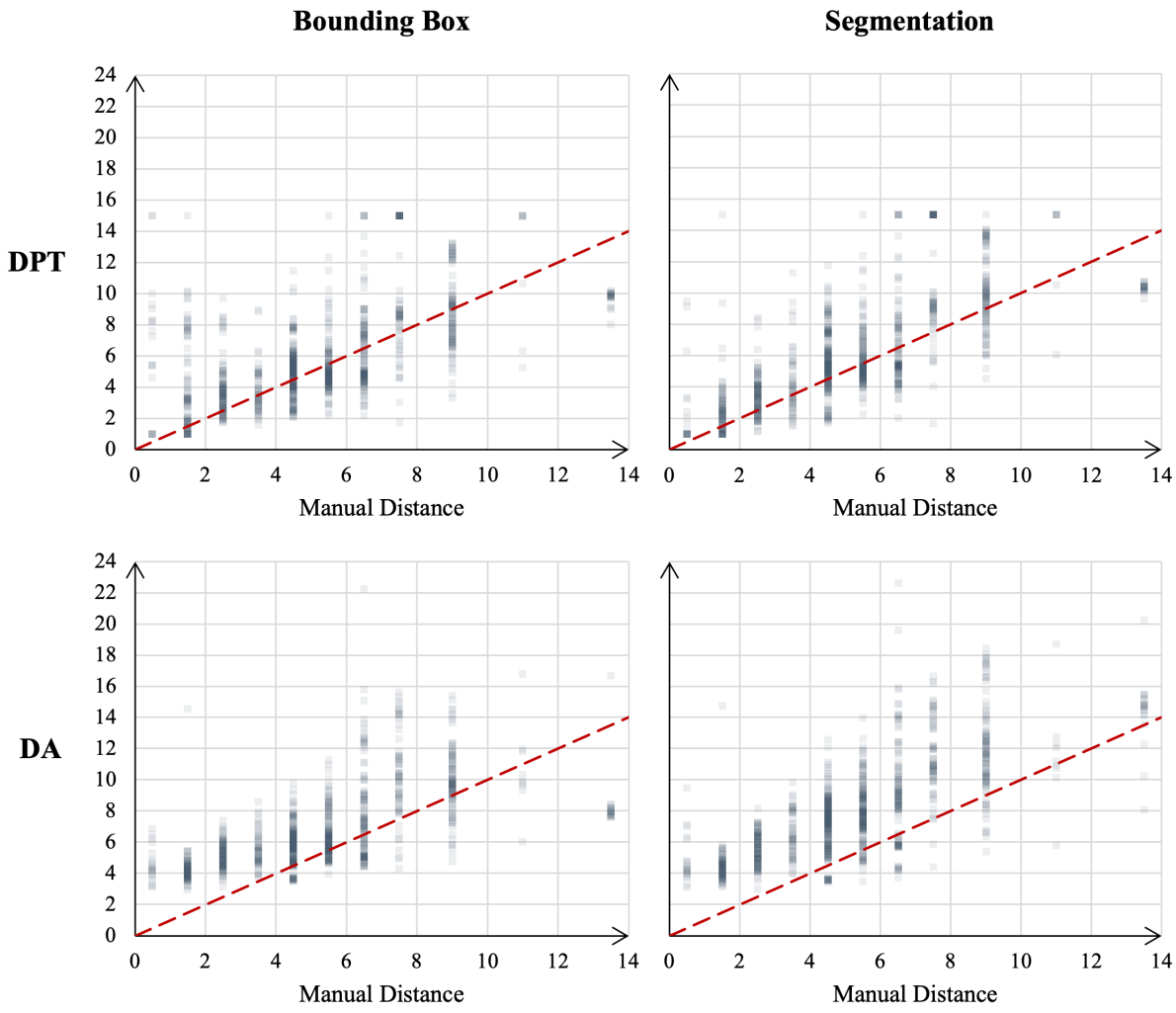

Figure 4.5: Graphs showing all modelled distance estimates mapped to thier corresponding manual estimates for the configurations: DPT/bounding box (top-left), DPT/segmentation (top-right), DA/bounding box (bottom-left) and DA/segmentation (bottom-right). The red dashed line shows the ideal (i.e., model=manual). All distances are in units of meters.

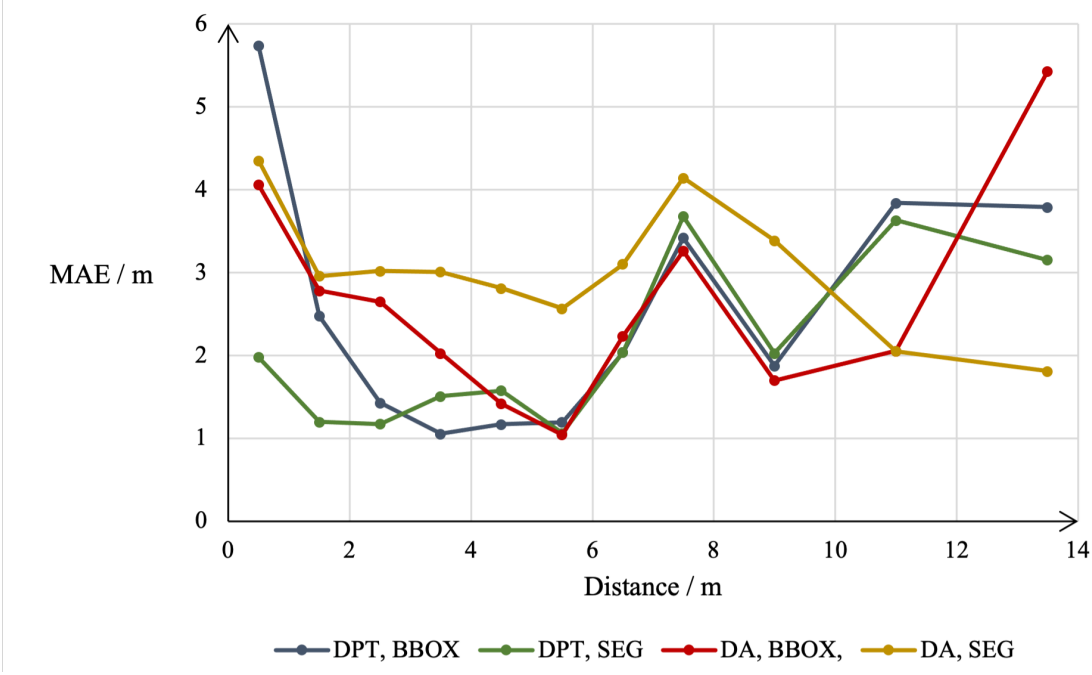

Figure 4.6: Mean average error for distance estimates grouped by their corresponding manual estimates for each pipeline configuration.

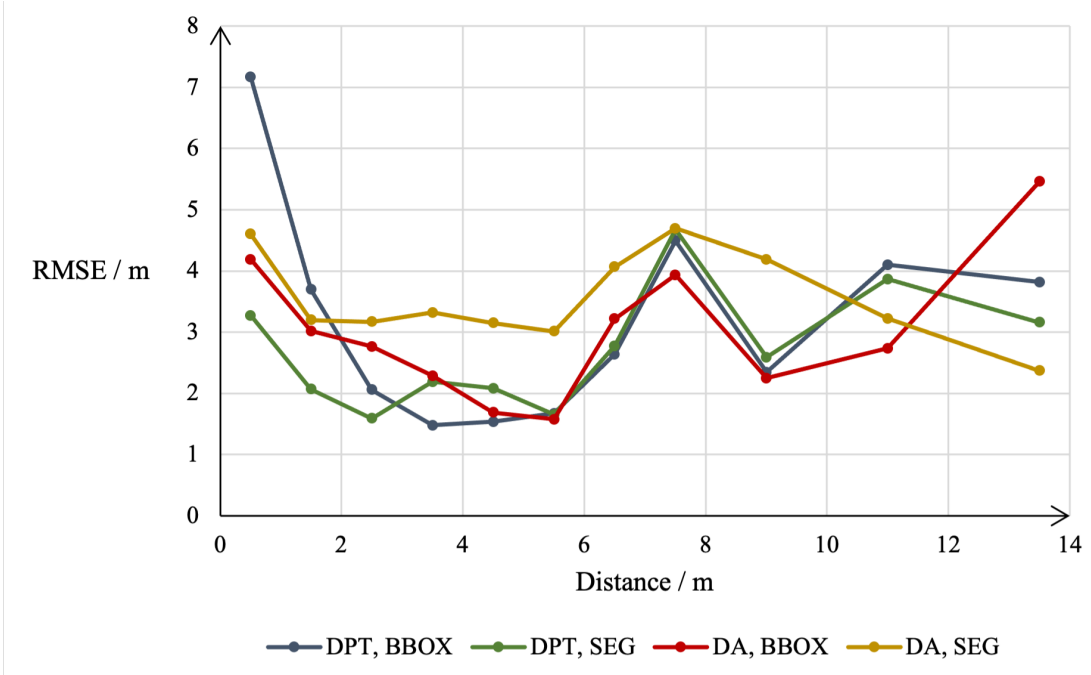

Figure 4.7: Root mean squared error for distance estimates grouped by their corresponding manual estimates for each pipeline configuration.

Detection Method Effects

Contrasting the distance estimates of the two detection method using DPT distance estimation, an increase in both accuracy and precision is observed for the segmentation method compared to the bounding box method at close distances (i.e., < 2 meters). For modelled distances corresponding to manual estimates of 0.5 meters, the bounding box method gives a MAE of 5.74 meters, a RMSE of 7.17 meters and an interquartile range of 7.50 meters while the segmentation method gives a MAE of 1.98 meters, a RMSE of 3.27 meters and an interquartile range of 1.61 meters. This is explained by a better detection frame calibration alignment. Here, only the detection frame pixels defined by the chimpanzee segmentation are masked as opposed to those within the entire bounding box. As a result, more pixels corresponding to the transect background which are common to both the calibration and detection frames are available for depth scale alignment, leading to a superior calibration of the detection frame depth scale.

This effect is exemplified in Figure 4.8, where the bounding box method results in an effectively failed calibration while the segmentation method succeeds.

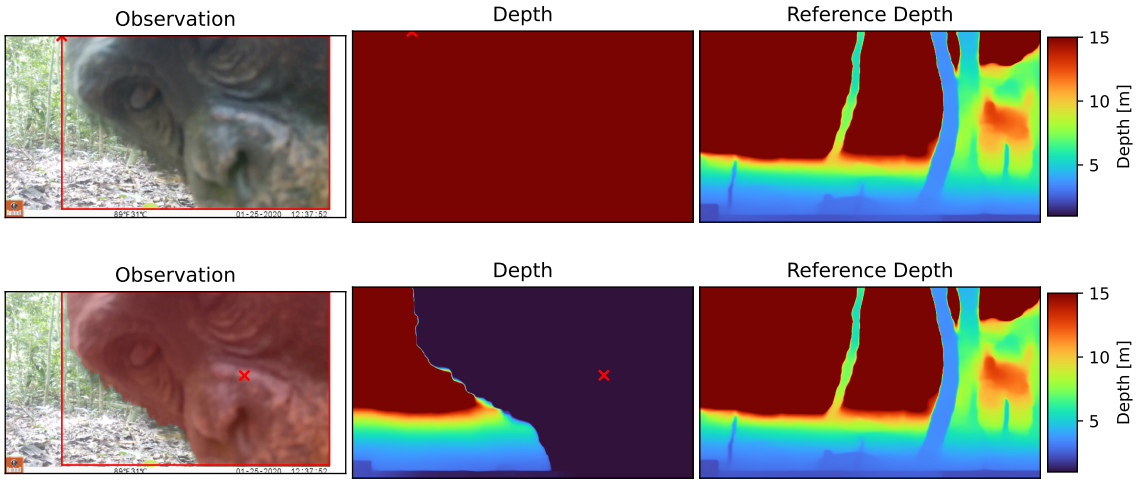

Figure 4.8: Example of DPT depth maps generated using bounding box (top) and segmentation (bottom) detection methods at a detection distance of 0.5 meters (manual estimate). In this example, the bounding box method gives a distance estimate of 15.0 meters while the segmentation method gives a distance estimate of 1.0 meters

In contrast, this detection effect does not apply when Depth Anything is used to estimate distance. No significant difference is observed in the accuracy and/or precision of the distance estimates obtained using the two detection methods. For modelled distances corresponding to manual estimates of 0.5 meters, the bounding box method gives a MAE of 4.06 meters, a RMSE of 4.19 meters and an interquartile range of 0.95 meters while the segmentation method gives a MAE of 4.35 meters, a RMSE of 4.60 meters and an interquartile range of 0.80 meters. This is an expected result given that Depth Anything is a metric depth model. Unlike with DPT, the scale of the generated depth maps are not aligned to that of the reference frames meaning that the detection method does not influence this scale. Therefore, it can be inferred that the difference in distance estimates given by these detection methods is a result of the different pixel sampling techniques used to calculate the final distance estimate (i.e., bounding box depth from the pixel corresponding to the 20th percentile depth and segmentation depth from the depth of the centre-most pixel) rather than differences on depth scaling.

Figure 4.9 shows the corresponding Depth Anything depth maps using both detection methods for the same example shown in Figure 4.8. It can be seen that both detection methods lead to identical depth maps, with the final detection distance estimates being very close.

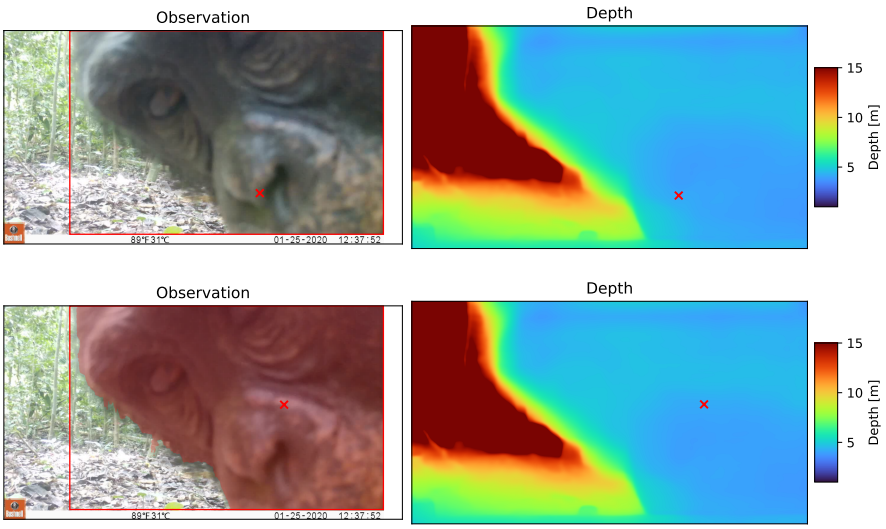

Figure 4.9: Example of Depth Anything depth maps generated using bounding box (top) and segmentation (bottom) detection methods at a detection distance of 0.5 meters (manual estimate). In this example, the bounding box method gives a distance estimate of 4.11 meters while the segmentation method gives a distance estimate of 4.17 meters.

Upon inspection of Table 4.2, it can be seen that for both DPT and Depth Anything models, the gradients of the regression lines (correlating the averages of the binned distance estimates with the corresponding manual estimates) of the segmentation detection methods are higher relative to their corresponding bounding box regression gradient and also better correlated to the ideal. A possible explanation for this trend is a minimised contribution of occluding pixels during the final detection distance calculation at medium to long distances when using the segmentation method. With this particular dataset, it is from these medium distances where detections start to become occluded by foliage. When the bounding box method is used, the final distance can be biased towards a lower estimate in circumstances where a significantly large portion of the bounding box is composed of occluding pixels that correspond to a significantly shorter distance. Conversely, a high quality segmentation often avoids this. This effect is exemplified in Figure 4.10.

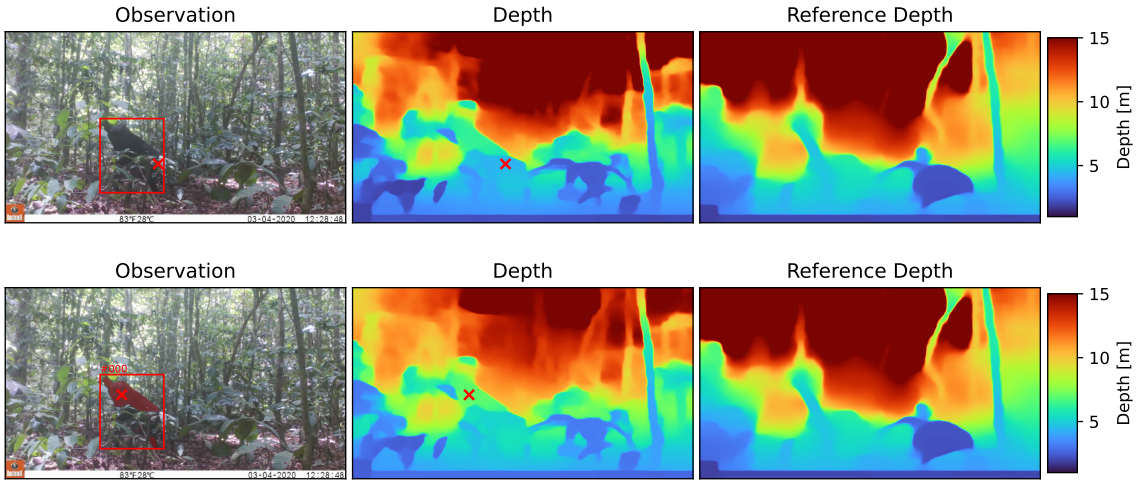

Figure 4.10: Example of DPT depth maps generated using bounding box (top) and segmentation (bottom) detection methods at a detection distance of 6.5 meters (manual estimate) where the detected individual is partially occluded by foliage. Here, the bounding box method gave a distance estimate of 4.40 meters while the segmentation method gave a distance estimate of 6.96 meters.

If, however, the quality of the segmentation is poor, the distances of occluded detections may still be underestimated. Figure 4.11 highlights a scenario where an individual was correctly detected but, due to poor detection frame quality, the occluding foliage was instead segmented rather than the individual.

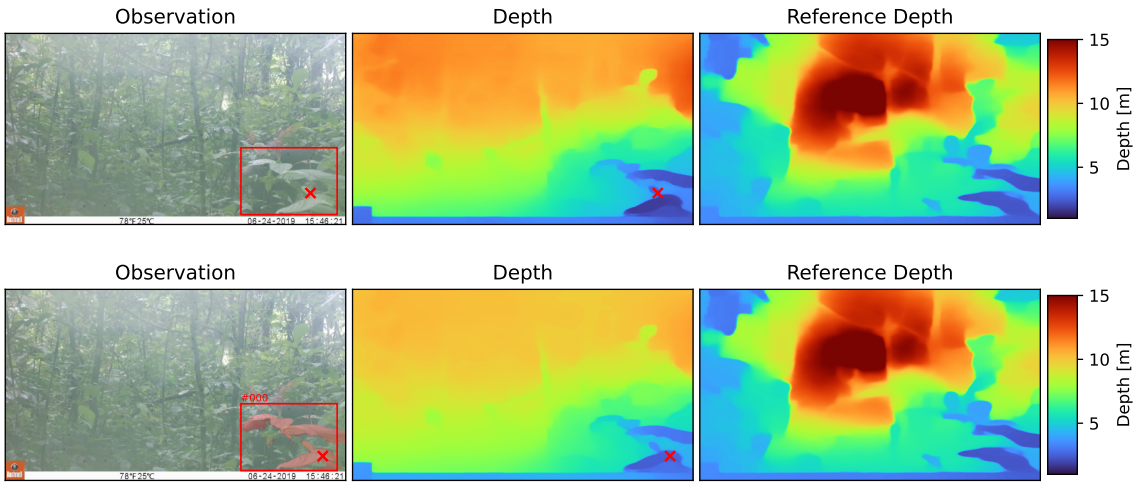

Figure 4.11: Example of DPT depth maps generated using bounding box (top) and segmentation (bottom) detection methods at a detection distance of 4.5 meters (manual estimate) where the detected individual is partially occluded by foliage. Here, moisture on the camera lense has resulted in a slightly hazy image, leading to a failed segmentation of the detected individual. The bounding box method gave an estimated distance of 2.64 meters while the segmentation method gave a distance estimate of 2.56 meters.

Blurry detections frames

djou09

Depth Model Effects

As seen in Table 4.2, the gradients of the regression lines correlating the averages of the binned DPT distance estimates with respect to their manual estimates are greater and also closer the ideal (gradient = 1) than that of Depth Anything. This shows that, overall, DPT is better capturing the true scale of depth in the detection frames than Depth Anything. At short to medium distances (i.e., 2–7 meters), Figure 4.4 shows that distance estimate averages closely follow a linear relationship, indicating that within this distance region, both depth models are differentiating relative depth well. However, Depth Anything generally over-predicts estimated depth in this region while the DPT estimates are much closer to the ideal. This trend also holds for the extreme-close distance region for Depth Anything (with both detection methods) and also DPT using segmentation (DPT with bounding box over-predicts as previously discussed).

While DPT gives a better measure of depth scale at close distances, the depth maps show that Depth Anything is superior at differentiating the depth of fine-detail in the detection frames. This is shown in Figure 4.12, where the depth maps generated by all configurations with a common detection frame are shown side-by-side.

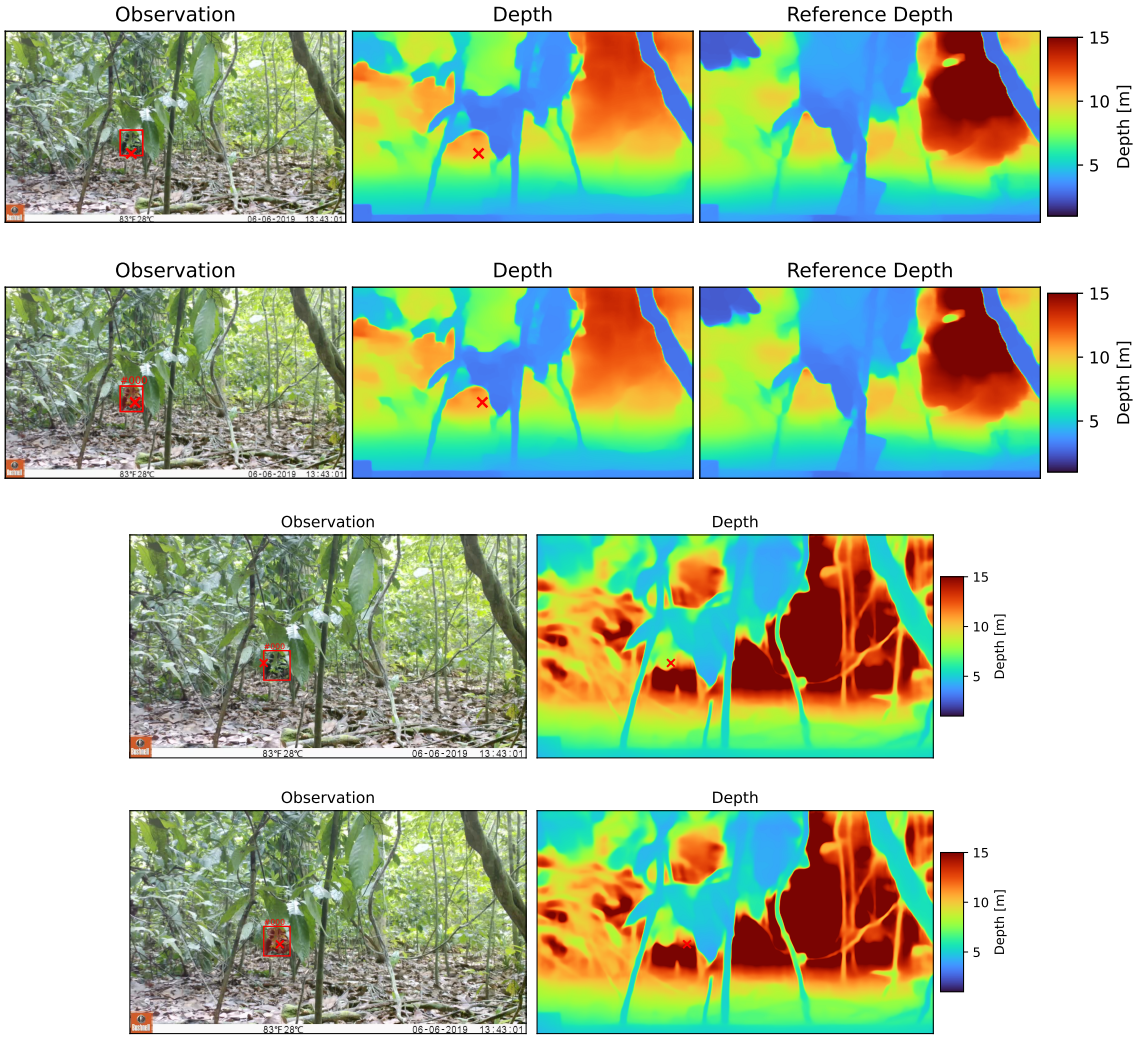

Figure 4.12: Example of depth maps generated using DPT/BBOX (row one), DPT/SEG (row two), DA/BBOX (row three) and DA/SEG (row four) detection methods at a detection distance of 13.5 meters (manual estimate). The following distance estimates were given: DPT/BBOX = 10.0 meters, DPT/SEG = 10.4 meters, DA/BBOX = 7.83 meters, DA/SEG = 13.6 meters.

The graphs in Figures 4.4 and 4.5 show that for both DPT configurations as well as Depth Anything with bounding box, distance estimates corresponding to manual estimates of 13.5 meters are significantly under-predicted. This is not the case, however, for estimates given by Depth Anything using segmentation detection. This is an interesting effect which is perfectly illustrated in Figure 4.12. In this example both of the DPT depth maps fail to capture any significant differentiation in depth within the small detection region. As a result, the final distance estimates yield roughly equal values of approximately 10 meters. In contrast, the Depth Anything depth map (identical for both detection methods) captures much more detail in this region, therefore enabling the different pixel sampling techniques used in the detection distance calculation to yield different results. In the case of the bounding box method, the contribution of close-distance occluding pixels to the percentile calculation skews the distance estimate, resulting in a value of 7.83 meters. For the segmentation method, however, this is no such influence since detection distance is determined by the depth of centre-most point of the segmentation, resulting in a value of 13.6 meters which is aligned closely to the manual estimate of 13.5 meters. Results such as these are achieved in spite of imperfect segmentation since the most important factor in these cases, where there is high contrast surrounding the desired pixel sampling region, is negating the influence of these surrounding pixels.

Summary

Errors table

All configurations over-predict

Sweet spot 2–7 meters

Error spike at 7.5 meters

Literature

4.2.2 Effects of Varying Calibration

Exclusive Close/Far Reference Calibration

Table 4.3 shows the overall errors of DPT distance estimates (for bounding box and segmentation) calibrated using only the closest and furthest reference frames at each transect. Figure 4.13 shows both the individual modelled distance estimates as well as the averages of the modelled distance estimates mapped to their corresponding manual estimates.

Table 4.3: Mean average error (MAE), root mean squared error (RMSE) and average difference between model and manual estimate ($\Delta_{average}$) for the bounding box and segmentation detetion methods using only the closest and furthest calibration frames as references.)

Method	MAE / m	RMSE / m	$\Delta_{average}$ / m
Bounding Box	2.42	3.13	-1.03
Segmentation	2.16	2.83	-0.755

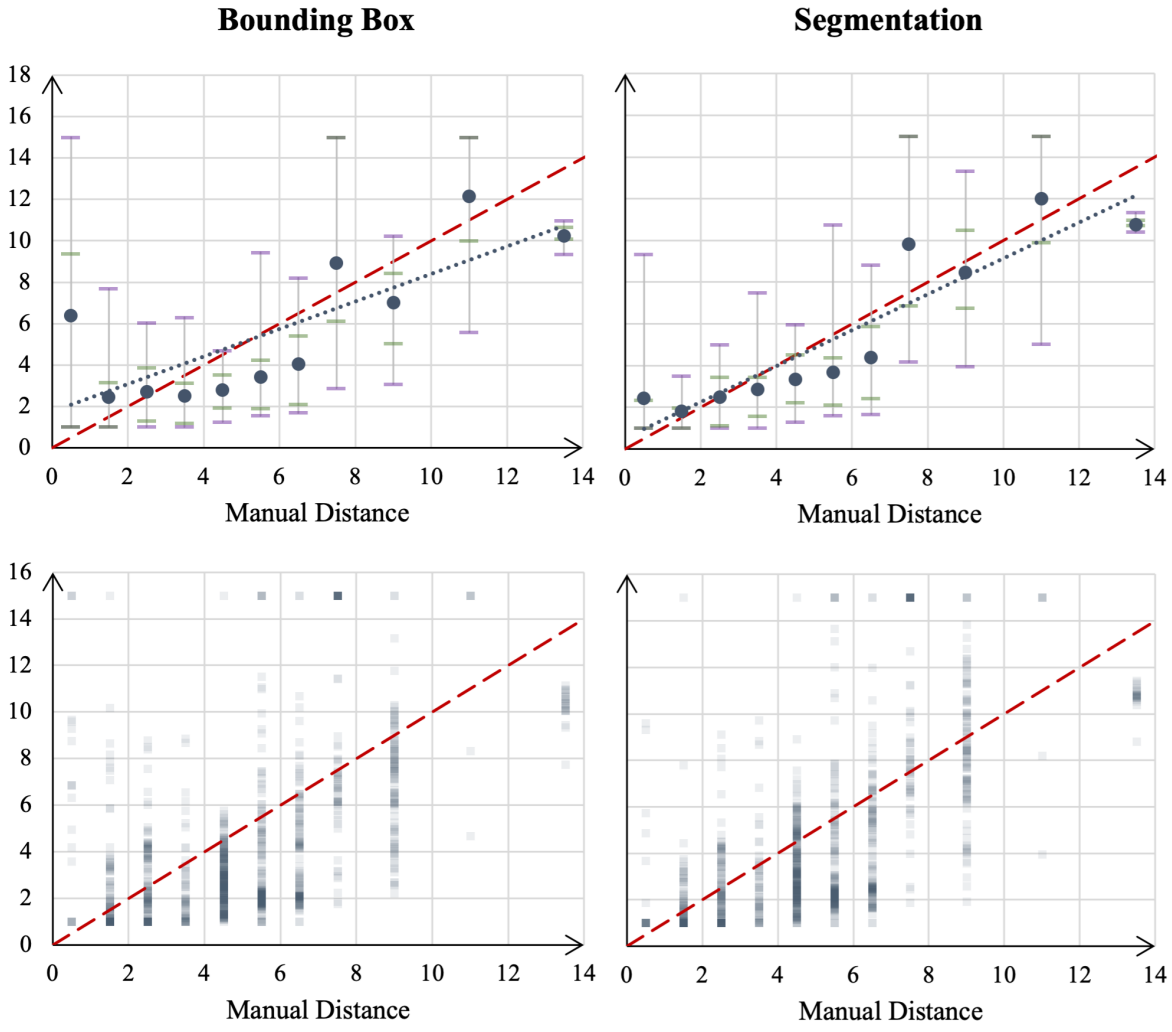

Figure 4.13: Graphs showing mean DPT distance estimates (top) and all modelled distance estimates (bottom) mapped to thier corresponding manual estimates for bounding box (left) and segmentation (right) detection methods using only the closest and furthest calibration frames as references. The blue dotted line shows the fitted regression line. The red dashed line shows the ideal (i.e., model=manual). the error bars show the 25–75 (green) and 5–95 (purple) percentiles. All distances are in units of meters.

On inspection of Table 4.3 and Figure 4.13, it is evident that this reduced calibration has impacted the results significantly. The bounding box results give an overall MAE of 2.42 meters and RMSE of 3.13 meters which correspond to increases of 0.61 meters and 0.47 meters respectively when compared to the results obtained using all reference frames while the segmentation results give an overall MAE of 2.16 meter and RMSE of 2.83 meters which correspond to increases of 0.46 meters and 0.38 meters respectively. The spread of the distance estimates have also increased for frames corresponding to all manual distances. Moreover, the negative $\Delta_{average}$ values show that this calibration, on average, leads to underestimation of distances in this dataset

The increase in error of this calibration method is most likely a result of the calibration function being much more susceptible to any outliers in the DPT estimated depth of the reference frames in conjunction with any noise/imperfections in the corresponding calibration frame masks. When many references are used, the impact of sparse outliers are minimised due to calibration function derivation from predominantly representative data.

More interestingly, the observed shift towards underestimates may be explained by a possible non-linearity in the relationship between raw DPT generated depth and ground-truth depth. Specifically, DPT outputs a disparity for each pixel in the processed image. The calibration frames are used to derive a linear function that maps these raw disparity values to a calibrated disparity[1] that is equal to inverse distance (i.e., distance = 1/calibrated disparity). However, if the true relationship between raw disparity and inverse distance is non-linear, (e.g., polynomial), a calibration using only the closest and furthest references would result in a biasing of distance estimates in the mid-distance region. This effect is visualised in Figure 4.14.

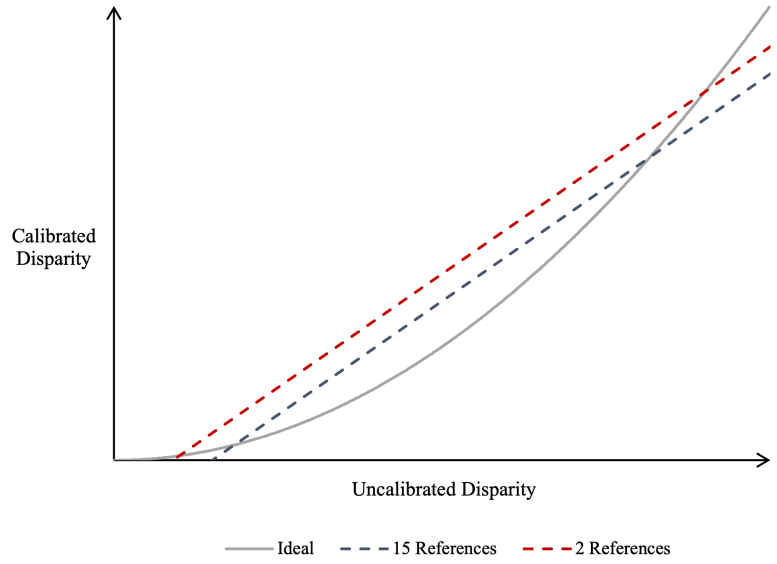

Figure 4.14: Simple visualisation showing how the number of reference frames can influence the fitting of a calibration function. The grey solid line represents the true mapping of the uncalibrated DPT outputted disparity to the calibrated disparity (i.e, 1 / distance). The blue dashed line represents a linear disparity mapping fitted using a series of (i.e., fifteen) reference frames while the red dashed line represents one fitted using only two (i.e., closest and furthest) reference frames

For example, suppose calibration using all references and calibration using only the closest and furthest references give the calibration functions

$$D_{cal} = 1.5 \cdot D_{raw} - 0.2 \qquad \text{and} \qquad D_{cal} = 1.5 \cdot D_{raw} - 0.05$$

respectively, where D_{raw} is the raw uncalibrated disparity and D_{cal} is the calibrated disparity equal to inverse distance. If a pixel has a raw disparity of $D_{raw} = 0.3$, applying the 'all references' calibration gives a calibrated

disparity of D_{cal} = 0.25, resulting in an estimated distance of 4 meters while applying the 'two references' calibration gives a calibrated disparity of D_{cal} = 0.4, resulting in an estimated distance of 2.5 meters.

This hypothesis is supported when the DPT reference depth maps calibrated using both reference frame samples are compared. Figure 4.15 shows examples of how the close/far frame calibration results in a shift in depth towards closer distances.

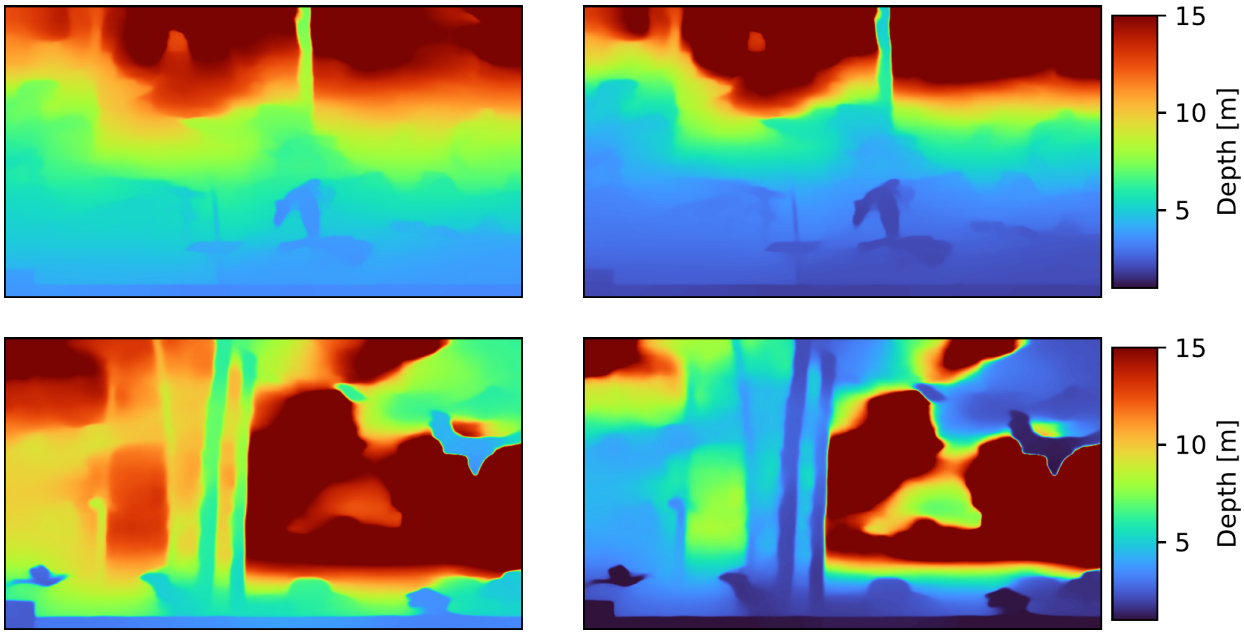

Figure 4.15: Two examples of DPT reference depth maps calibrated using all reference frames (left) and reference frames at one and fifteen meters (right). The images on the right show a general biasing of depth towards closer distances for pixels in the close to medium depth regions.

It is also seen that depth contrast of the far-region pixels appears lost in some cases where the close/far calibration is used (Figure 4.16). This implies that the raw disparity values of these pixels are calibrated in such a way as to correspond to values greater than the maximum depth (i.e., disparity < 1/15, distance > 15) and are therefore clipped to this maximum value.

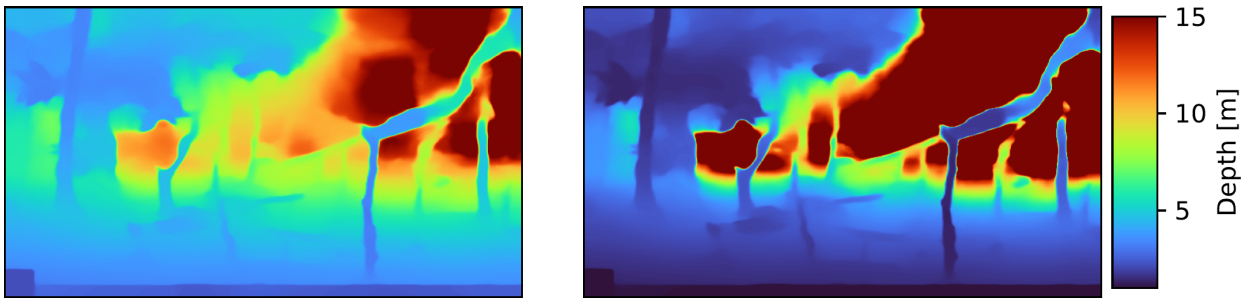

Figure 4.16: Example of DPT reference depth maps calibrated using all reference frames (left) and reference frames at one and fifteen meters (right). The calibration seen on the rights shows a loss of depth contrast in the far-distance region. Here, all pixels corresponding to distances of approximately ten meters and over are overestimated to the maximum depth of fifteen meters. The calibration seen on the left better captures the depth contrast of the far-region.

To rationalise this behaviour, where close pixels are biased towards closer distances and far pixels are biased towards further distances, it can be imagined that the scale of the true mapping of raw to calibrated disparity becomes significantly increases at large raw disparity values. This is visualised in Figure 4.17.

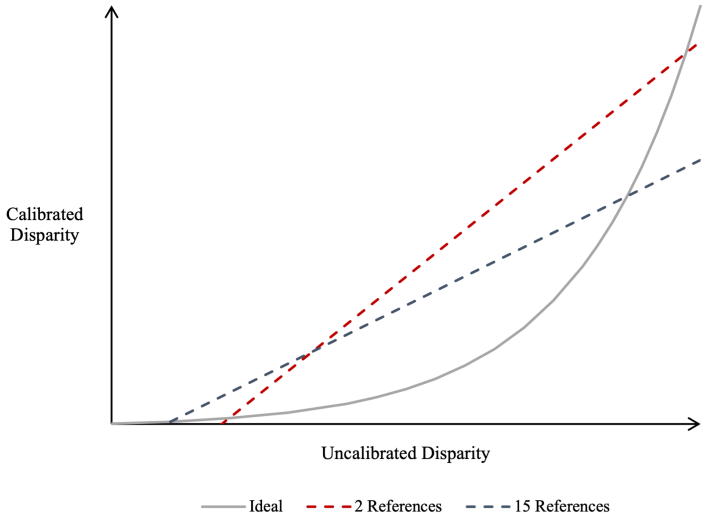

Figure 4.17: Simple visualisation showing how sufficient non-linearity in the raw disparity to calibrated disparity mapping can result in different scales of fitted calibration function approximations when different numbers of reference frames are used. The grey solid line represents the true mapping of the uncalibrated DPT outputted disparity to the calibrated disparity (i.e, 1 / distance). The blue dashed line represents a linear disparity mapping fitted using a series of (i.e., fifteen) reference frames while the red dashed line represents one fitted using only two (i.e., closest and furthest) reference frames

The effect of this is that pixels with a raw uncalibrated disparity greater than a certain value will give a higher calibrated disparity (inverse distance) and result in a lower distance estimate while those below this value will give a lower calibrated disparity and result in a higher distance estimate. Ultimately, this highlights the importance of a well fitting calibration function and suggests that better distance estimates could be achieved by accounting for non-linearity in the disparity relationship.

Variable Calibration Effects

To understand the variable calibration results, the mean average error of the distance estimates with respect to their corresponding manual estimates was calculated for each calibration of DPT with both bounding box and segmentation detection methods (Figure 4.18).

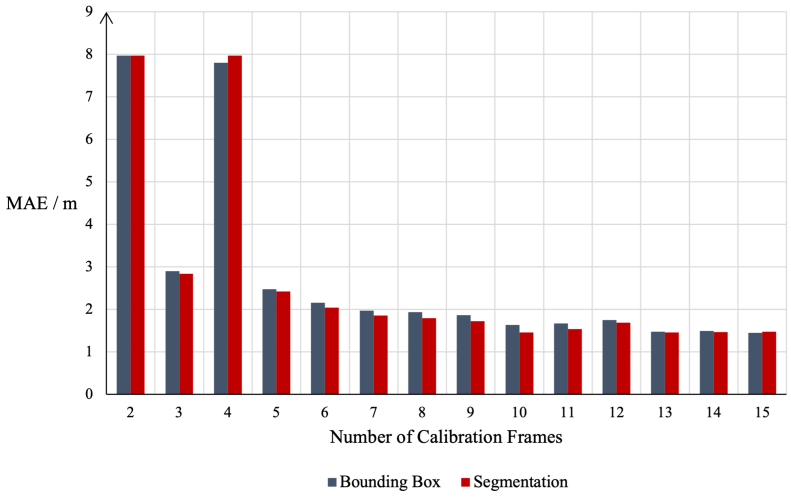

Figure 4.18: Graph showing the mean average error (MAE) of all distance estimates using DPT calibrated with a varying number of refrence frames with respect to their corresponding manual distance estimate.

The graph in Figure 4.18 shows minimal difference in error between the corresponding bounding box and segmentation runs with a general trend towards lower error as the number of reference frames is incremented, owing to the availability of more points of reference by which the calibration function is derived. The error also appears to stabilise onwards from ten references. This reflects the distribution of distances within the dataset, where the majority of manual estimates fall into the mid-distance range. At this particular camera location, the furthest manually estimated distance of any given observation in the associated detection frames was nine meters, aligning with the observed error stabilisation from onwards of approximately ten meters, where additional reference frames beyond this point correspond to distances greater than the maximum observation distance.

To visualise the effect of varying the number of reference frames on calibration, the calibrated camera location depth maps are shown for each of the runs below (Figure 4.19).

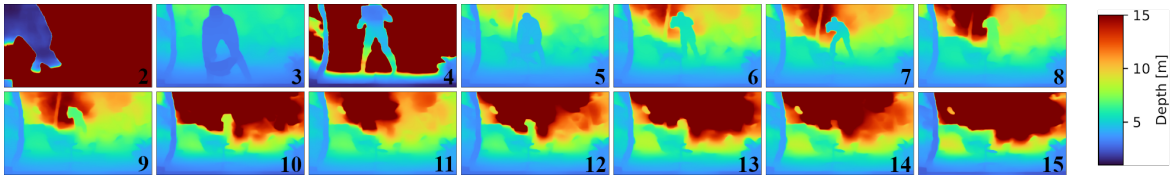

Figure 4.19: DPT depth maps of a single camera location generatated using a variable number of calibration frames, each labelled with the corresponding number of frames (also the landmark distance of the furthest frame) in the bottom-right corner

On inspection of Figure 4.19, calibrations two and four stand out as having vastly overestimated background depth. This shows that, in these cases, the scale of the calibration function is amplified so much as to scale the disparity of these background pixels to below the maximum depth disparity (i.e., lower than 1/15) resulting in these pixels being capped at the maximum depth of 15. This is reflected in Figure 4.19, where large spikes are observed in the errors of the associated distance estimates.

It is not immediately clear why this occurs with just these two calibrations. Calibration three, despite not capturing the overall depth scale, does not show this effect and correlates well with the remaining calibration distance estimate errors and depth maps. A reason may be due to instability of the calibration function when few references are used resulting from a greater influence from errors in DPT output and calibration frame mask quality; however, this remains unclear.

It is also worth noting that, in the close distance calibrations (roughly corresponding to the top rows in Figure 4.19), a reference landmark is depicted as a segmented depth in the depth maps. This is explained by understanding that the calibrated depth map of a given camera location is derived from the calibrated reference frame in which that landmark is at the furthest distance form the camera[1]. This gives a calibration map where the landmark occupies the smallest number of pixels so to maximise the background context. When few references are used, however, the furthest landmark is close and occupies a prominent number of pixels.

4.3 Analysis of Activity Estimates

False Positives and Misses

4.3.1 Manual Sample Activity Analysis

Single chimp frame distances supplemented with manual distances

4.3.2 Automated Sample Activity Analysis

5 Evaluation of Methodology

Improvement over fully manual approach?

Bottlenecks, calibration frame preparation, available compute

Time saved using frame extractor, preserves exact pixels

Limitations of the environment

Calibration and detection frame image resolution

Depth estimation point consistency (the animal itself has depth)

Error in reference videos (person has depth, not always standing up straight)

Better calibration (polynomial/not just linear, more calibration frames)

6 Conclusion

6.1 Further Work

References

[1] T. Haucke, H. S. Kühl, J. Hoyer, and V. Steinhage, "Overcoming the distance estimation bottleneck in estimating animal abundance with camera traps," *Ecological Informatics*, vol. 68, p. 101536, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954121003277

[2] R. Khanam and M. Hussain, "What is yolov5: A deep look into the internal features of the popular object detector," 07 2024.