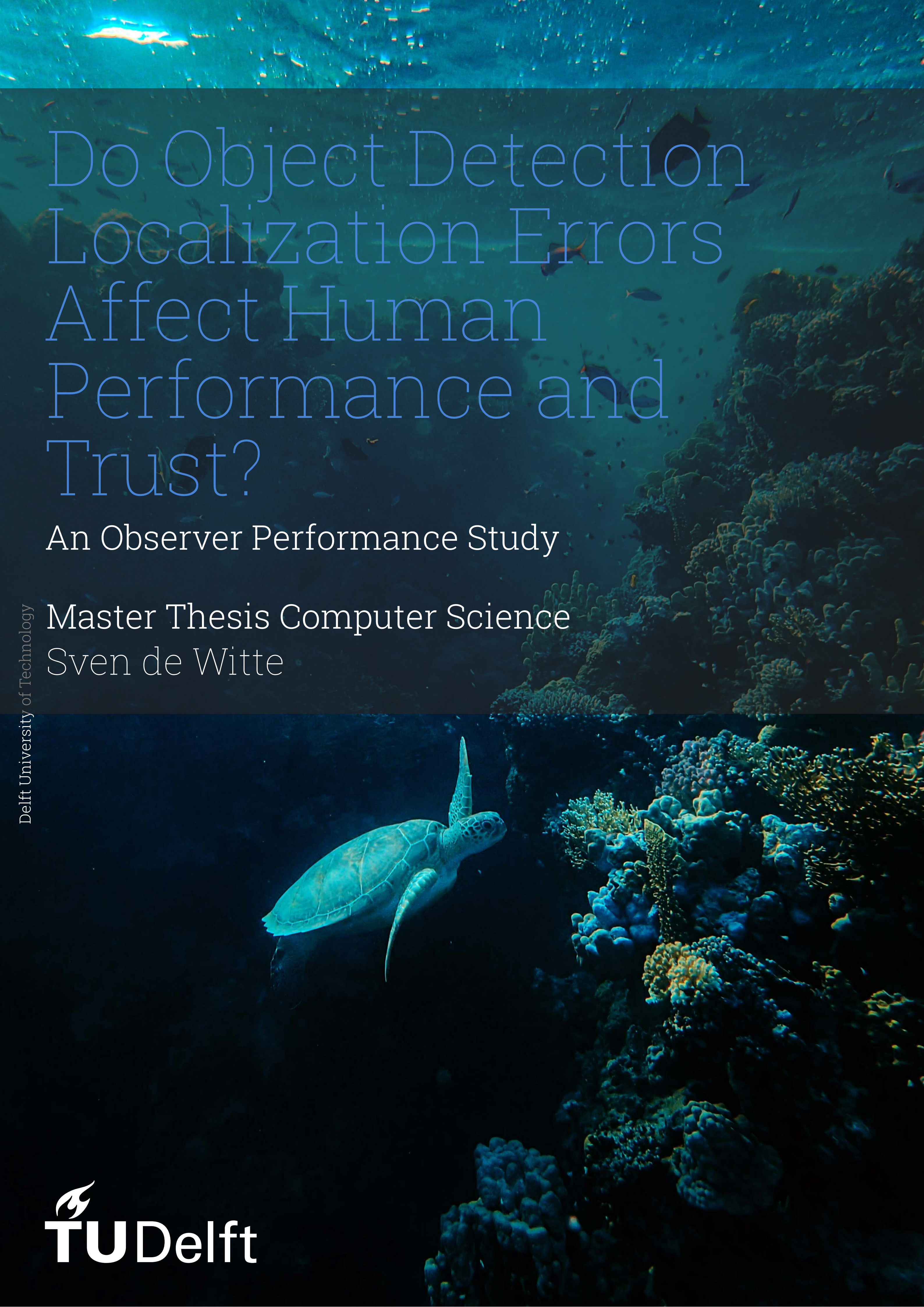

Do Object Detection Localization Errors Affect Human Performance and Trust?

An Observer Performance Study

Master Thesis Computer Science

Sven de Witte

Delft University of Technology

Do Object Detection Localization Errors Affect Human Performance and Trust?

An Observer Performance Study

by

Sven de Witte

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Thursday December 21, 2023 at 10:00 AM.

Student number:	4728254	
Thesis committee:	Prof. dr. ir. J. van Gemert,	TU Delft, supervisor
	Dr. E. O. Strafforello,	TU Delft, daily supervisor
	Dr. J. Martinez Castaneda,	TU Delft

Cover: Photo by Francesco Ungaro on Unsplash

Preface

This document marks the conclusion of my master’s thesis, titled ”Do Object Detection Localization Errors Affect Human Performance and Trust?” undertaken at Delft University of Technology. Navigating the rapidly evolving field of AI and computer vision, although challenging at times, has been primarily inspiring and a significant motivator for continuous improvement.

I undertook this project because I believe it is crucial to examine the interaction between humans and the system. The system can only operate at its full potential when humans can perform at their best. This can only be achieved by understanding the interaction between humans and the system. I extend my gratitude to the Computer Vision Lab for providing me with the opportunity and resources to conduct this research. A special acknowledgment goes to Jan van Gemert, my supervisor, for offering invaluable critical feedback and posing insightful questions. I would also like to express my appreciation to Ombretta Strafforello, my daily supervisor, for weekly meetings and guidance, helping me maintain a clear perspective on tasks and their execution. Lastly, I want to thank Jorge Martinez Castaneda for his presence and efforts as an additional committee member.

In conclusion, heartfelt thanks to my family, friends, and girlfriend for their unwavering support and motivation over the last few months.

Sven de Witte
Delft, December 2023

Contents

1 Introduction

Object detectors are algorithms designed and trained to recognize a particular object or set of objects within an image or video. In the research field of computer vision, object detection stands as one of the foundational technologies that empower machines to perceive and understand the visual world. Object detectors trained to identify specific objects or abnormalities can assist humans in their work, such as helping doctors understand medical MRIs or aiding individuals in quality control to identify irregularities within a product.

One way to visualize the outcomes of such object detectors is by using bounding boxes. As the name suggests, bounding boxes are rectangular boxes drawn around the object of interest. When examining an image, the bounding box is drawn in a manner that aligns its edges with the left, right, top, and bottommost points of the object.

The second chapter of this report is an article that describes the findings of my master's thesis, which explores the relationship between the accuracy of a set of bounding boxes within an image and the performance and trust of a human working with that image. The research for this master's thesis took the form of an observer performance study, aiming to uncover the connection between changes in a system and the performance of humans working with these systems. These studies are crucial for analyzing and improving processes where humans need to make decisions based on visual information. Domains where this form of research is applied include medical imaging, security screening, and Computer-Aided Detection Systems.

The last chapter of this report covers some background information that aid in understanding some of the terms and concepts described in my thesis. It will explain in more detail what a observer performance study is, some core concepts of computer vision and some of the tests used within the thesis.

The insights from this study not only enhance our understanding of human performance but also provide guidance for engineers and researchers developing algorithms for human-computer tasks.

2
Scientific Article

Do Object Detection Localization Errors Affect Human Performance and Trust?

Sven de Witte, Ombretta Strafforello and Jan van Gemert
Computer Visions lab, Delft University of Technology, Delft, the Netherlands
S.deWitte@student.tudelft.nl, O.Strafforello@tudelft.nl, J.C.vanGemert@tudelft.nl

Keywords: Observer Performance, Object Detection, Bounding Box, Trust

Abstract: Bounding boxes are often used to communicate automatic object detection results to humans, aiding humans in a multitude of tasks. We investigate the relationship between bounding box localization errors and human task performance. We use observer performance studies on a visual multi-object counting task to measure both human trust and performance with different levels of bounding box accuracy. The results show that localization errors have no significant impact on human accuracy or trust in the system. Recall and precision errors impact both human performance and trust, suggesting that optimizing algorithms based on the F1 score is more beneficial in human-computer tasks. Lastly, the paper offers an improvement on bounding boxes in multi-object counting tasks with center dots, showing improved performance and better resilience to localization inaccuracy.

1 INTRODUCTION

Automatic object detectors are used to localize and classify objects appearing in images and videos. These algorithms have application in a number of fields, including autonomous driving, surveillance, medical imaging, augmented reality, robotics and visual inspection. In this work, we are interested in the applications that involve humans as end users of object detectors. Important examples are anomaly detection in surveillance footage and examination of medical images. A common approach of showing the outcome of a object detection system to a human is with bounding boxes. Bounding boxes are rectangular boxes drawn around each object of interest. Object detectors are trained to predict bounding boxes that closely match ”ground truth” bounding boxes drawn by humans.

The quality of object detections is assessed using standard evaluation methods that do not consider the detectors' intended application. Often, the evaluation is achieved by means of the mean average precision (mAP), a metric that combines object classification and localization accuracy. In particular, an object is considered accurately localized if there is sufficient overlap between the bounding box predicted by the algorithm and the ground truth. The overlap is calculated using intersection-over-union (IoU). Object detections with IoU greater than 0.5 or 0.75 and correct classification are considered acceptable (Lin et al., 2014; Everingham et al., 2010).

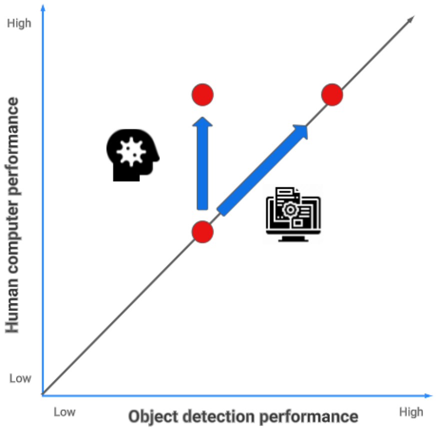

Figure 1: Illustration visualizing improvement of human performance in human computer task. A design choice focused on the human in the system could improve performance without need of object detector improvement.

However, relying on IoU (and therefore, on mAP) can be misleading as IoU is highly affected by small annotation errors present in current datasets (Murrugarra-Llerena et al., 2022). In addition, previous work shows that the IoU does not consistently align with users preference (Strafforello et al., 2022). In this work, we explore how the performance of the end human users is affected by the object detectors localization errors. We do this using observer performance studies on a practical inspection task including multi-object counting task. In addition, we measure how the human trust towards the object detection system varies as we vary the object detections localization quality. Our study reveals that using center dots instead of boxes not only improves the efficiency of the human in the process, but also increases the overall resilience to inaccurate localization. Generating a performance increase without the need for increased object detector performance as illustrated in Figure 1.

This paper contributes by (1) Designing a study to test human accuracy and performance on a simple object counting task. (2) Showing the relation between object detection accuracy and human performance. (3) Showing the relation between object detection accuracy and human trust. (4) Lastly, showing how simple design choices can increase user performance.

2 RELATED WORK

Research on object detectors aims to enhance the localization and classification performance on a variety of images and datasets, accelerate the inference speed and reduce the computational requirements. Current models comprise two-stage models (Cai and Vasconcelos, 2018; Girshick, 2015; Girshick et al., 2015; Ren et al., 2015), single stage approaches (Lin et al., 2017; Liu et al., 2016; Redmon et al., 2016; Redmon and Farhadi, 2018), pointwise/anchorless methods (Duan et al., 2019; Law and Deng, 2018; Zhou et al., 2019), transformers-based detectors (Li et al., 2022; Beal et al., 2020; Carion et al., 2020; Dai et al., 2021; Zhu et al., 2020) and, more recently, diffusion models (Chen et al., 2023).

Although the latest state-of-the-art models have achieved competitive performance on standard benchmarks, progress in object detection is often carried out overlooking what the intended application of object detectors is. In fact, (Strafforello et al., 2022) showed that when object detections are meant to be shown to humans, standard evaluation metrics are unreliable. In this work, our objective is to assess the impact of object detections localization accuracy when they are intended for applications involving human users. In this regard, we conduct an observer performance study to compare object detectors localization accuracy with users performance.

Using observer performance study to analyze and improve the workflow between humans and computer vision systems is not new. (Thompson et al., 2013) deployed observer performance studies to assess assistive imaging techniques in radiology. Similarly (Sunwoo et al., 2017) looked at the impact computer aided detection for radiologist using 3D MR imaging.

When looking at the mixed field of psychology and computer science, *trust* is an important factor to consider. Extensive research has been conducted on effectively measuring human trust in computer systems, yet it remains a challenging task. Multiple studies examining trust measures and methodologies are available (Kohn et al., 2021; Brzowski and Nathan-Roberts, 2019; Mcknight et al., 2011). However, it has been demonstrated that many studies employ trust measures that are either inadequately validated or specifically designed for a particular use case (Kohn et al., 2021). In this study, we use the multi-dimensional scale proposed by (Gulati et al., 2019) to assess user trust in human-computer interactions.

3 METHOD

Our goal is to investigate the relation between object detection localization accuracy and human performance and trust. To do this, we conduct an observer performance study. We design a task where participants of the study have to count the number of aquatic creatures in images.

In the experiments, participants are shown images containing bounding boxes and center dots with varying localization accuracy. The answers and response times per image are recorded. At the end of the task, the participants are asked to fill in a survey related to the trust in the system that generated the bounding boxes. Participants are informed that the same system could potentially be utilized to monitor the growth and well-being of creatures in the aquarium or for an automated feeding system, where errors might result in incorrect meal sizes. The survey consists of 12 question about the perceived risk, benevolence, competence and reciprocity. Respondents answer through a Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree).

The errors introduced are using 2 different measures. The IoU that gives a bounding box a score based on the overlap of the placed box compared to the ground truth. The Shifted boxes and Shifted Dots are shifted into a random direction from the ground truth to create an IoU of 0.5. The other measure we use is the F1 score. The F1 score is calculated using the precision and recall. Precision measures how accurate a object detector is looking at the amount of true positive boxes placed compared to the total amount of boxes placed. This score is varied by placing additional boxes. Recall measures the ability of the object detector to find all the objects of interest calculated by dividing the number of true positive boxes by the total number of boxes that the ground truth has. A lower recall can be achieved by removing some of the ground truth boxes. The object detections proposed in our study are generated by manipulating ground truth bounding boxes. We do not make use of real object detectors to ensure full control of the object detections localization error. Our study participants are recruited through Amazon Mechanical Turk (mtu,).

In total we perform 7 experiments, each containing a set of 30 images divided into 3 smaller tasks of 10 images. This choice was made to reduce the workload per participant, increasing their cooperation and reducing the chance of noise in our results. Every task is performed by 10 to 12 participants, resulting in approximately 36 participants per experiment following the recommendations of (Carvalho et al., 2016).

Within this study, we measure the participants error by calculating the absolute difference between the real number of aquatic creatures in an image and the provided answer. We also calculate the agreement between the participants using the Krippendorff's Alpha (Krippendorff, 2010) and use it to assesses the participants' reliability. To test the significance of our findings we use the T-test with Bonferroni correction. This study involves 7 different groups of data resulting in significance threshold $\alpha = \frac{0.05}{7} = 0.0071$.

4 Experiments

4.1 Dataset

All images used in this study come from the Aquarium combined data set from Roboflow(Dwyer and Nelson, 2022; Dwyer and Nelson, 2014). The dataset contains images with multiple aquatic creatures. The images were labeled by the Roboflow team with help of SageMaker Ground Truth. From this dataset the 30 most suitable images for the experiments were hand picked. To be suitable for the experiment an image has to have multiple creatures in the image, the creatures must be in the water (images containing puffins outside the water could lead to confusion), the image must be clear and the creatures in the image have to be big enough to be able to recognize as a aquatic creature. The bounding box information of some of the images were updated after finding additional unmarked fish that were in the image or reflections of fish that were marked by the original team. All bounding boxes and dots are drawn is a bright red color. As red contrasted the best with the dark and blue tones that were most present in the images. Examples of the different kind of images are visualized in the grid in Figure 2.

4.2 Pilot Testing

Prior to conducting the main survey on Amazon Mechanical Turk, we conducted pilot surveys in the MTurk sandbox to evaluate the clarity of images and accuracy of bounding box surveys. The purpose of this pilot study was to gather feedback on the survey and assess the task's difficulty level. If the task was too easy, participants would likely achieve close to 100% accuracy with short response times, potentially diminishing the performance differences between various experiments. On the other hand, if the task was too difficult, overall performance across all experiments would likely be poor. The accuracy results from the two pilot surveys indicated that the task's difficulty level was appropriate for testing the hypotheses.

Additionally, the pilot studies served as a means to estimate the average time required to complete the survey. This information was essential for ensuring that participants received appropriate compensation for their time investment in the task. Moreover, the pilot surveys helped identify any errors in the images and any ambiguities in the task or images that could lead to divergent subjective interpretations of the correct answers. Such discrepancies would undermine or interfere with the hypotheses being tested.

4.3 Clean Image Baseline

The first experiment aimed to determine the task difficulty and establish a baseline for overall performance. Clean images, consisting of the 30 images used in this study without additional information such as bounding boxes, were employed. These 30 images were divided into three counting tasks, each containing 10 images.

Analyzing the results of the individual counting tasks gives an average error of 3.64, 3.68, and 3.88. These results indicate that the task difficulty remains relatively consistent across all the counting task. In this study, the Krippendorff's Alpha value was found to be 0.76, indicating a substantial level of agreement among participants regarding the correct answers. A low score would indicate disagreement or unreliability in the data, which would mean that the images or task are ambiguous or data is labeled wrong. A score of 0.76 suggest that the data and task are clear and we can use the outcome.

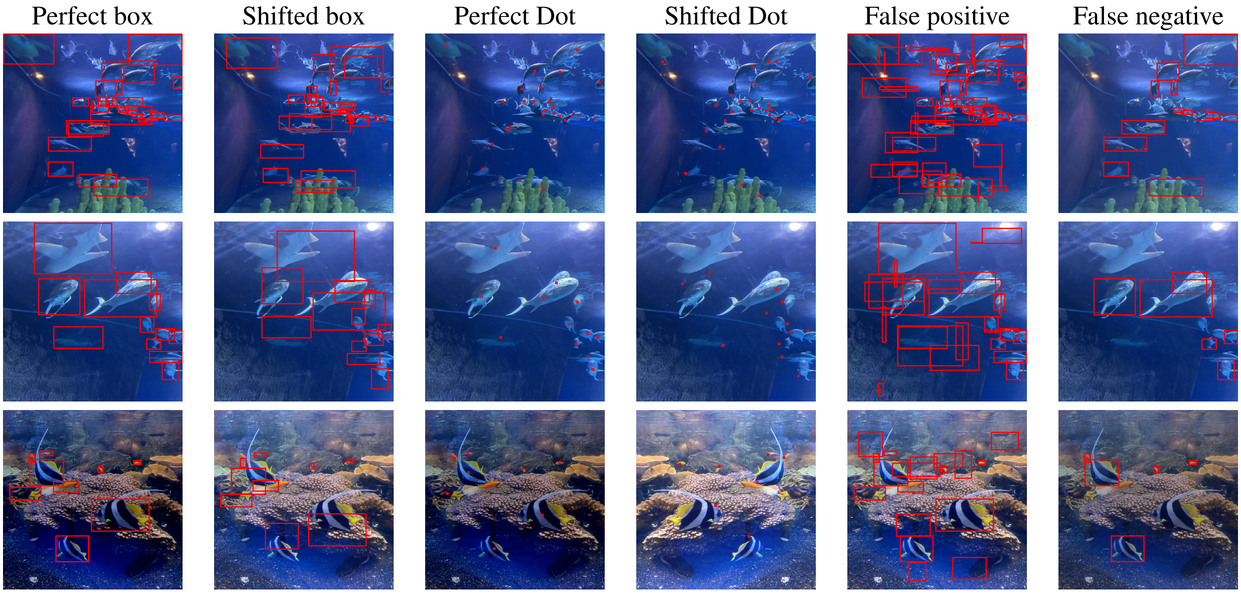

Figure 2: Example of images used within the study. Going from left to right accurate bounding boxes (*Perfect box*), bounding boxes with 0.5 IoU (*Shifted box*), accurate Dots (*Perfect Dot*), Inaccurate dots (*Shifted Dot*), additional boxes(*False positive*) and missing boxes(*False negative*).

The overall average error for the clean image task was 3.7, with a standard deviation of 3.6 (see Table 1). The standard deviation was calculated based on the absolute errors for all questions, without distinguishing between errors above or below the correct answer. These findings serve as a valuable baseline for the subsequent experiments, providing insights into the task difficulty and establishing a reference Dot for performance comparison.

4.4 Perfect Bounding Boxes

The objective of the second experiment is to assess whether adding a correct bounding boxes to the images aids the participants' performance. We use the same experimental setup as for first experiment.

As presented in Table 1, the average error and standard deviation of errors in this experiment were significantly lower than those observed in the baseline experiment. As anticipated, participants exhibited improved performance in accurately counting the fish when provided with the correct bounding box information. Furthermore, participants spent less time on the images in this experiment. However, we found no statistical significant difference in response times. This implies that participants spent roughly the same amount of time on images both with and without bounding boxes, despite their enhanced performance when the bounding boxes were available.

These findings suggest that the inclusion of correct bounding boxes in the images led to a significant reduction in errors and improved participant performance in counting the fish. Moreover, the response times remained comparable, indicating that participants efficiently processed the images with or without the presence of bounding boxes, albeit with superior results when aided by the bounding box information.

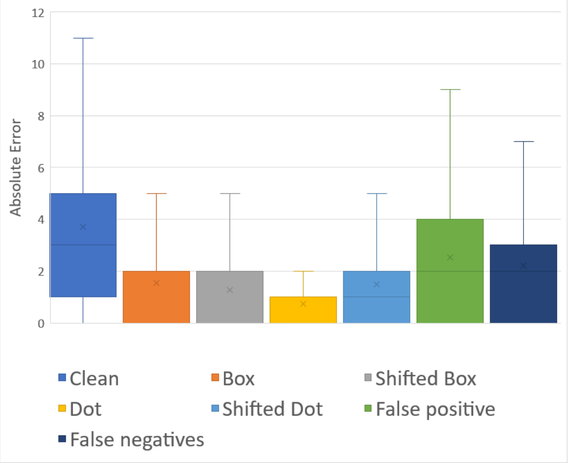

Figure 3: Boxplot comparing the absolute error of the different experiments.

4.4.1 Shifted Bounding Boxes

The purpose of the third experiment is to evaluate the performance of humans when presented with images

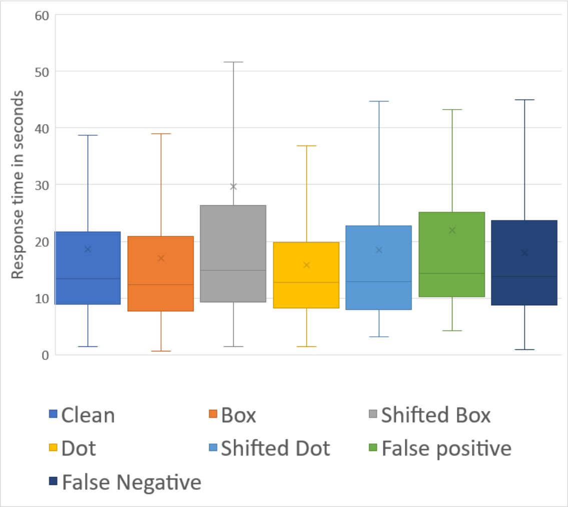

Figure 4: Boxplot comparing response times within the different experimets.

containing bounding boxes with lower object detection localization accuracy. In this experiment, the bounding boxes in the images are randomly shifted in direction to create an IoU value of 0.5. The scale of the bounding boxes remains consistent, as variations in bounding box scale could introduce additional factors that might influence the experiment's outcomes. Similarly to the previous experiments, the images were divided into three counting tasks, with a total of 36 participants.

As shown in Table 1, the average error observed in this experiment does not appear significantly different compared to the experiment with correct bounding boxes. Surprisingly, the participants' performance is even better when confronted with the shifted bounding boxes. Conducting a T-test on the data confirms that there is no significant difference in accuracy between the two types of images. However, a notable difference is observed in the response times for the images with participants taking an average of 12 seconds more of the images containing the shifted boxes. This finding suggests that overall performance on the images is better when participants are provided with correct bounding boxes. The discrepancy in response times can be attributed to participants needing to count more carefully on the images with shifted bounding boxes compared to those with correct bounding boxes.

4.5 F1 Errors

Localization errors are not the only type of errors that object detectors struggle with. A object detector can fail to find the objects of interest or might detect a object where there is none. These are called type II and type I errors or false-positives and false-negatives. The F1 score is used to give the object detector a score based on these errors .

Two experiments were run to analyze the impact of type I and type II errors on the human performance and trust. The hypotheses for these two experiments were that the overall performance and trust of the humans would be lower than in the case of the correct or shifted boxes. This because the amount of boxes shown does not equal the correct amount of sea creatures within the image.

Table 1: Participant mean error and response time when presented with Clean images (*Clean*). accurate bounding boxes (*Perfect box*), bounding boxes with 0.5 IoU (*Shifted box*), accurate Dots (*Perfect Dot*), Inaccurate dots (*Shifted Dot*), additional boxes(*False positive*) and missing boxes(*False negative*).

	Mean error	Mean response time
Clean	3.7 ± 3.6	18.6 ± 17.4
Perfect Box	1.5 ± 2.7	$17.0s \pm 15.8$
Shifted Box	1.3 ± 2.0	$29.7s \pm 7.3s$
Perfect Dot	0.7 ± 1.8	$15.8s \pm 11.7s$
Shifted Dot	1.5 ± 2.5	$18.5s \pm 16.8s$
False positive	2.5 ± 2.7	$21.9s \pm 21.8s$
False negative	2.2 ± 2.2	$17.9s \pm 12.2s$

4.5.1 False-Positives

In this experiment, the images were altered to have additional boxes on random locations in the image as demonstrated in the 5th column of the image grid in Figure 2. The bounding box size was in a range around the average box size within the image. The amount of boxes in the image was doubled to create a precision score of 0.5.

As the data in Table 1 shows, the average error is 2.524. Using a t-test with an $\alpha = 0.0071$ it shows that this difference is significant compared to that of correct bounding boxes. The same goes for the average time spent per image with an average of 21.891 seconds. This result is important as it shows that false-positive errors impact the performance of the human both in accuracy and speed.

4.5.2 False-Negatives

In this experiment, we remove 50% of the bounding boxes in each image, leading to a decrease of the recall score to 0.5 as demonstrated in the 6th column of figure 2. This way not all sea creatures are as easy to find as in the experiments with perfect recall, and there are clearly visible mistakes from the object detector.

The data in Table 1 shows an average error that is significantly higher compared to the correct bounding boxes. With an average error of 2.216. This shows that the participants had an overall worse performance when not all of the boxes are present in the image. However, we find no significant difference in the response time. This means that, on average, people spent about the same amount of time on false negative images as they did on the images with the correct bounding boxes.

4.6 Perfect Center Dots

The introduction of center dots is proposed as a potential solution to address challenges associated with bounding boxes such as occlusion, overlapping and clutter. In this experiment, images are modified to include red dots positioned at the center of where the original bounding boxes would be, effectively marking all aquatic creatures with a red dot. The hypothesis states that images with center dots will yield superior human performance and accuracy compared to images with correct bounding boxes. This hypothesis stems from the notion that bounding boxes in a multiple object counting task may introduce additional difficulties for certain images.

As depicted in some of the images in Figure 2 the proximity of multiple bounding boxes can result in the formation of new boxes through overlap. These new boxes could potentially confuse or mislead participants attempting to count the number of aquatic creatures as quickly as possible, leading them to either perceive empty boxes or count additional boxes. To address this issue, the presented approach utilizes center dots, represented by down scaled versions (9 by 9 pixels) of the original bounding boxes. The expectation is that employing center dots will enhance human performance and accuracy while potentially reducing the overall Intersection over Union (IoU) metric.

Table 1 illustrates that the overall task performance using center dots surpasses that of correct bounding boxes. With a mean error of 0.74, the center dot approach achieves significantly lower error rates compared to the 1.53 error associated with correct bounding boxes in this simple task. Utilizing a T-test with a significance threshold α of 0.0071 demonstrates that the difference in performance is statistically significant and not merely a result of random chance. Conversely, no significant difference is observed in response times between the two approaches. This finding is intriguing, as it indicates that participants spent an equivalent amount of time on both types of questions, while accuracy with center dots was higher.

The results indicate that incorporating center dots leads to superior performance and accuracy in the task compared to the use of correct bounding boxes. This outcome holds even though participants spent similar amounts of time on both types of questions

4.6.1 Shifted Center Dots

How do participants perform when the location of the center dots is inaccurate? This experiment is similar to the *Shifted Bounding Box* (Section 4.4.1), but in this case, the dots are deliberately shifted to achieve lower accuracy. The images in this experiment include center dots that are randomly displaced in a direction based on where the center of the box would be with a 0.5 IoU when drawing the bounding boxes. We hypothesize that the advantage of over bounding boxes found in the *Perfect Center Dots* experiment will still hold, even with reduced accuracy. This is because there is a greater likelihood that the center dot will still connect to a portion of the object. Additionally, center dots may aid in human perception of a less accurate system. Bounding boxes make it easier for humans to perceive the boundaries of the box, as demonstrated in the study by List and Bins (List et al., 2005).

The results indicates that the participants performed with similar accuracy and time compared to the correct bounding boxes. Using a T-test with a significance level of α 0.0071 on the times of the shifted bounding boxes and shifted dots reveals a significant difference. This suggests that the overall performance of the shifted dots, although inferior to that of the center dots, is better than that of the shifted bounding boxes. The aforementioned perceptual factors could also play a role, where the reduced accuracy of the dots might be less noticeable, resulting in people being less cautious compared to their interaction with the shifted bounding boxes.

4.7 Trust

In addition to the participants performance, we measure the participants trust towards the object detector. Table 2 shows the collected trust scores, ranging on a scale from 1 to 100. Surprisingly, there is no significant difference in trust when showing participants correct bounding boxes versus boxes with localization errors. This might suggest that users solved the counting task by counting the number of bounding box present in the images, while not focusing too much on the exact box location.

Table 2: Result summary of trust scores per experiment out of 100. The data shows no signifact drop in trust with perturbations to the box and dots location, while adding or removing false and true positive bounding boxes does. This suggest to optimize for precision and recall in human computer tasks.

	Perfect box	Shifted Box	Perfect Dot	Shifted Dot	False positive	False negative
Trust score	63.7 ± 17.5	62.5 ± 16.0	57.7 ± 10.0	63.7 ± 16.4	28.8 ± 12.9	23.9 ± 8.7

On the other hand, the trust scores of the experiments with the false negatives and false positives are significantly lower than the *perfect box* baseline. This suggests that showing an incorrect amount of bounding boxes significantly impact the participants perception of the object detection system.

The trust scores collected in presence of false negatives and false positives align with the diminished performance of participants observed in these two experiments, suggesting that users' self-perception of their own performance plays a pivotal role in determining their trust in the system.

5 DISCUSSION

In this work, we show that the human performance in a visual multi-object counting task is improved when introducing an assistive object detection system. Secondly, we measure how human performance varies as we control for object detector localization errors. Interestingly, we find that perturbing the location of the object detections does not degrade the human counting accuracy nor their trust in the system, but only increases the completion time of the task. In addition, human performance is improved even when the object detections are not perfect, but contain precision and recall errors. On the other hand, we find that the human trust in the object detector system significantly decreases in presence of false positive or false negative detections. Therefore, we conclude that when object detections are meant to be presented to humans, it is more important to optimize the F1-score over the IoU.

Finally, we show that the visualization strategy used to present the object detections has an impact on the human performance. In fact, in our study, using center dots instead of bounding boxes increases the human performance in the object counting task and the resilience to localization errors. This highlights the importance of presenting data in a way that is optimal for the task.

REFERENCES

Amazon mechanical turk. https://www.mturk.com/. Accessed: 2023-02-20. 3

Beal, J., Kim, E., Tzeng, E., Park, D. H., Zhai, A., and Kislyuk, D. (2020). Toward Transformer-Based Object Detection. 2

Brzowski, M. B. and Nathan-Roberts, D. (2019). Trust measurement in human–automation interaction: A systematic review. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63:1595–1599. 2

Cai, Z. and Vasconcelos, N. (2018). Cascade r-cnn: Delving into high quality object detection. In *CVPR*. IEEE/CVF. 2

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. *Lecture Notes in Computer Science*, page 213–229. 2

Carvalho, A., Dimitrov, S., and Larson, K. (2016). How many crowdsourced workers should a requester hire? *Annals of Mathematics and Artificial Intelligence*, 78:45–72. 3

Chen, S., Sun, P., Song, Y., and Luo, P. (2023). Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843. 2

Dai, Z., Cai, B., Lin, Y., and Chen, J. (2021). Up-detr: Unsupervised pre-training for object detection with transformers. In *CVPR*, pages 1601–1610. IEEE/CVF. 2

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., and Tian, Q. (2019). Centernet: Keypoint triplets for object detection. In *ICCV*. IEEE/CVF. 2

Dwyer, B. and Nelson, J. (2014). Aquarium combined roboflow. https://public.roboflow.com/object-detection/aquarium. 3

Dwyer, B. and Nelson, J. (2022). Roboflow (version 1.0) [software]. available from https://roboflow.com. computer vision. 3

Everingham, M., Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338. 1

Girshick, R. (2015). Fast r-cnn. In *ICCV*. IEEE. 2

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. *PAMI*. 2

Gulati, S., Sousa, S., and Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, 38:1–12. 2

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y. C., and Shaw, T. H. (2021). Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Front Psychol*, 12:604977. 2

Krippendorff, K. E. (2010). *(Vols.* 1-0). 3

Law, H. and Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. In *ECCV*. Springer. 2

Li, Y., Mao, H., Girshick, R., and He, K. (2022). Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer. 2

Lin, T., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *ICCV*. IEEE/CVF. 2

Lin, T.-Y., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312. 1

List, T., Bins, J., Vazquez, J., and Fisher, R. (2005). Performance evaluating the evaluator. pages 129–136. 6

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., and Berg, A. (2016). Ssd: Single shot multibox detector. volume 9905, pages 21–37. 2

Mcknight, D., Carter, M., Thatcher, J., and Clay, P. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, 2:12–32. 2

Murrugarra-Llerena, J., Kirsten, L. N., and Jung, C. R. (2022). Can we trust bounding box annotations for object detection? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4813–4822. 1

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *CVPR*. IEEE/CVF. 2

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*. 2

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 2

Strafforello, O., Rajasekart, V., Kayhan, O. S., Inel, O., and van Gemert, J. (2022). Humans disagree with the iou for measuring object detector localization error. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1261–1265. IEEE. 1, 2

Sunwoo, L., Kim, Y. J., Choi, S. H., Kim, K.-G., Kang, J. H., Kang, Y., Bae, Y. J., Yoo, R.-E., Kim, J., Lee, K. J., Lee, S. H., Choi, B. S., Jung, C., Sohn, C.-H., and Kim, J. H. (2017). Computer-aided detection of brain metastasis on 3d mr imaging: Observer performance study. *PLOS ONE*, 12(6):1–18. 2

Thompson, J. D., Manning, D. J., and Hogg, P. (2013). The value of observer performance studies in dose optimization: a focus on free-response receiver operating characteristic methods. *J Nucl Med Technol*, 41(2):57–64. 2

Zhou, X., Zhuo, J., and Krahenbuhl, P. (2019). Bottom-up object detection by grouping extreme and center points. In *CVPR*. IEEE/CVF. 2

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2020). Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv preprint arXiv:2010.04159*. 2

3 Background

3.1. Observer performance study

This research for this thesis was done in the form of a observer performance study. Observer performance studies are research methodologies used to assess and quantify the ability of human observers to perform a specific task, such as detecting abnormalities in medical images, identifying objects in visual scenes, or making judgments in decision-making scenarios. These studies are particularly common in fields such as medical imaging, computer vision, and psychology and can be used for instance as a comparative analysis of two systems [4] or to improve detection and classification[1][2].

The primary goal of observer performance studies is to evaluate and understand how well human observers can perform a given task under specific conditions. Here's a general outline of the key components of these studies:

1. **Task Definition:** The first step is to clearly define the task that the observers are expected to perform. This could be anything from detecting tumors in medical images to identifying specific objects in a visual scene.
2. **Study Design:** The study design outlines the experimental setup, including the selection of observers, the presentation of stimuli (e.g., images or scenarios), and any variations in conditions (e.g., different imaging techniques, levels of noise, or presence of distractors).
3. **Data Collection:** Observers are then presented with the stimuli, and their responses are recorded. This might involve rating the severity of a condition, making binary decisions (e.g., presence or absence of a lesion), or ranking items.
4. **Performance Metrics**: Various metrics are used to assess observer performance. Common metrics include sensitivity, specificity, accuracy, positive predictive value, and negative predictive value. These metrics provide quantitative measures of how well observers can perform the task.
5. **Statistical Analysis:** Statistical methods are applied to analyze the data and determine the significance of any observed differences or trends. This helps ensure that the results are not due to random chance.
6. **Results Interpretation:** The results are interpreted in the context of the study objectives. Researchers may draw conclusions about the effectiveness of a particular diagnostic method, the impact of different conditions on observer performance, or the reliability of human decision-making in a given context.
7. **Implications and Applications:** The findings from observer performance studies often have important implications for practical applications. For example, in medical imaging, these studies can influence the development and optimization of diagnostic techniques, potentially leading to improvements in patient care.

Observer performance studies provide valuable insights into the strengths and limitations of human observers in specific tasks. They help refine and validate techniques and technologies, ultimately contributing to advancements in fields where human judgment and decision-making play a crucial role.

3.2. Bounding Boxes

Bounding boxes are rectangular regions that are used to encapsulate and define the spatial location of objects or regions of interest within an image or a frame of a video. These boxes are commonly employed in computer vision tasks, especially in object detection, localization, and tracking. Bounding boxes can be defined either by a set of coordinates relating to the corners or the coordinate of a center point with a height and width.

3.3. Mean Avereage Precision

Mean Average Precision (mAP) is a metric used to evaluate the performance of object detection models. The mAP consist of 3 elements mainly the Intersection over Union(IoU) which rates the localization of the detection and the recall and precision which rate the relation between correct and wrong detections.

3.3.1. Intersection over Union

The intersection over Union or IoU is used to calculate how good a bounding box is placed on a detected object on a scale from 0 to 1. Here 0 means that there is no overlap between the predicted box and the ground truth or correct location and 1 means a perfect overlap. The IoU is calculated by taking the intersection between the predicted area and the ground truth and dividing this by the union of these two boxes.

$$\text{IoU} = \frac{\text{Intersection Area}}{\text{Union Area}}$$

A threshold can be set up to decide what degree of overlap is acceptable to be considered as a true detection. An IoU of 0.5 or 50% is often used, however this is highly dependent on how important location is for the use of the object detector.

3.3.2. Precision and Recall

Ones a threshold for the IoU is set the performance of the object detector can be measured. This is done by looking at the amount of correct predictions (true positives), incorrect predictions (false positves) and missing predictions (false negatives).

The precision measures how accurate the predictions are or the percentage of correct predictions, this is calculated by taking the ratio of true positive predictions to the total number of positive predictions.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

The recall measures how good the object detector is in finding all the objects and is calculated by taking the ratio of true positive predictions to the total number of actual positives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

When calculating these scores with different IoU tresholds a Precision-Recall curve can be made from which a single average precision (AP) can be taken or mean Average precision when looking at multiple classes.

3.3.3. F1 score

The precision and recall can also be used to calculate the F1 score. The F1 score is used when trying to optimize both recall and precision. The harmonic mean is used to prevent the score from being dominated by either precision or recall when one of them is very small.

$$\text{F1} = 2 * \frac{\text{Precision * Recall}}{\text{Precision + Recall}}$$

3.4. Object detection

Object detection is a computer vision task that involves identifying and locating objects of interest within an image or video. The goal is to not only classify the objects present but also to determine their precise locations by outlining bounding boxes around them. Object detection algorithms often make use of machine learning or deep learning to produce their results. What kind of algorithm you use depends on the problem you are trying to solve, the data you have and the amount of processing power you have access to. Some popular architectures and algorithms for object detection include Faster R-CNN (Regional-Based Convolutional Neural Network)[7], Mask R-CNN[3], YOLO (You Only Look Once)[6], SSD (Single Shot Detector)[**ssd**] and Retinanet[5]. These models are often pre-trained on large data sets and fine-tuned for specific tasks or domains. Object detection has seen significant advancements in recent years, ongoing research continues to improve the accuracy, speed, and efficiency of these models. Object detection enables a lot of computer vision based applications such as robot vision, autonomous driving and smart video surveillance.

3.5. T-Test

T-testing is a statistical method used to compare the means of two groups and determine if there is a significant difference between them. There are different types of t-tests, the paper uses an independent two-sample t-test, which is commonly used when comparing the means of two independent groups.

Assumptions: The data in each group should be approximately normally distributed. The variances of the two groups should be approximately equal.

- **Null Hypothesis (H0):** There is no significant difference between the means of the two groups.
- **Alternative Hypothesis (H1):** There is a significant difference between the means of the two groups.

The formula for the Two sample T-Test becomes:

$$t = \frac{\bar{X}1 - \bar{X}2}{\sqrt{\frac{s1^2}{n1} + \frac{s2^2}{n2}}}$$

Where:

- $\bar{X}1$ and $\bar{X}2$ are the sample means of the two groups.
- s1 and s2 are the sample standard deviations of the two groups.
- n1 and n2 are the sample sizes of the two groups.

If the absolute value of the t-statistic is greater than the critical value from the t-distribution table or if the p-value is less than the chosen significance level (commonly $\alpha = 0.05$), the null hypothesis is rejected.

3.6. Bonferroni correction

The Bonferroni correction is a method used to adjust the significance level of a statistical test when multiple comparisons are performed simultaneously. When conducting multiple statistical tests, the probability of making a Type I error (rejecting a true null hypothesis) increases. The Bonferroni correction helps control the familywise error rate by reducing the chances of finding a statistically significant result by chance.

The idea behind the Bonferroni correction is relatively straightforward. If you are conducting *k* tests and want to maintain an overall significance level of *α*, you adjust the significance level for each individual test to $\frac{\alpha}{k}$

The Bonferroni correction is a conservative method, meaning it reduces the risk of Type I errors but comes at the cost of potentially increasing Type II errors (failing to reject a false null hypothesis).

3.7. Kripperndorfs alpha

Krippendorff's alpha[**krippendorff2010vols**] is a reliability coefficient used to measure the reliability or agreement among multiple raters or annotators when assessing categorical data. It takes into account the possibility of chance agreement and provides a normalized measure of agreement. The formula for Krippendorff's alpha is as follows:

$$\alpha = 1 - \frac{Do}{De}$$

Where:

- *Do* is the observed disagreement.
- *De* is the expected disagreement.

The calculation involves comparing the observed disagreement to the expected disagreement under the assumption of random chance. The closer the value of Krippendorff's alpha is to 1, the higher the agreement among raters beyond what would be expected by chance. The resulting value of α ranges from -1 to 1. A value near 1 indicates high agreement beyond what would be expected by chance, while a value near 0 or negative suggests agreement is no better than random chance. Negative values may occur when there is less agreement than expected by chance.

References

[1] Joel G. Fletcher et al. “Observer Performance in the Detection and Classification of Malignant Hepatic Nodules and Masses with CT Image-Space Denoising and Iterative Reconstruction”. In: *Radiology* 276.2 (2015). PMID: 26020436, pp. 465–478. DOI: `10.1148/radiol.2015141991`. eprint: `https://doi.org/10.1148/radiol.2015141991`. URL: `https://doi.org/10.1148/radiol.2015141991`.

[2] David Gur et al. “Dose Reduction in Digital Breast Tomosynthesis (DBT) Screening using Synthetically Reconstructed Projection Images: An Observer Performance Study”. In: *Academic Radiology* 19.2 (2012), pp. 166–171. ISSN: 1076-6332. DOI: `https://doi.org/10.1016/j.acra.2011.10.003`. URL: `https://www.sciencedirect.com/science/article/pii/S1076633211004594`.

[3] Kaiming He et al. “Mask R-CNN”. In: *CoRR* abs/1703.06870 (2017). arXiv: `1703.06870`. URL: `http://arxiv.org/abs/1703.06870`.

[4] Maryam Jessop et al. “Lesion Detection Performance: Comparative Analysis of Low-Dose CT Data of the Chest on Two Hybrid Imaging Systems”. In: *Journal of Nuclear Medicine Technology* 43.1 (2015), pp. 47–52. ISSN: 0091-4916. DOI: `10.2967/jnmt.114.147447`. eprint: `https://tech.snmjournals.org/content/43/1/47.full.pdf`. URL: `https://tech.snmjournals.org/content/43/1/47`.

[5] Tsung-Yi Lin et al. “Focal Loss for Dense Object Detection”. In: *CoRR* abs/1708.02002 (2017). arXiv: `1708.02002`. URL: `http://arxiv.org/abs/1708.02002`.

[6] Joseph Redmon et al. “You Only Look Once: Unified, Real-Time Object Detection”. In: *CoRR* abs/1506.02640 (2015). arXiv: `1506.02640`. URL: `http://arxiv.org/abs/1506.02640`.

[7] Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *CoRR* abs/1506.01497 (2015). arXiv: `1506.01497`. URL: `http://arxiv.org/abs/1506.01497`.