Tampere University

Patrik Kakko

QUANTIFYING THE IMPACT OF OCCLUSION RATIO AND TYPE ON VISION MODEL DETECTION PERFORMANCE

Evaluating Robustness in Controlled Synthetic Occlusion Scenarios

Bachelor’s Thesis
Faculty of Information Technology and Communication Sciences
Examiner: Esa Rahtu
Instructor: Tapio Sorvajärvi
December 2025

ABSTRACT

Patrik Kakko: Quantifying Occlusion Ratio and Type in Vision Model Detection Performance: Evaluating Robustness in Controlled Synthetic Occlusion Scenarios
Bachelor's Thesis
Tampere University
Bachelor's Programme in Computing and Electrical Engineering, Information Technology
December 2025

The increasing deployment of deep learning-based computer vision models in diverse applications demands functionality beyond ideal conditions. These models often face various forms of noise, one of the most prominent being different occlusions. This thesis examines occlusions by employing three isolated occluders (solid, perforated, volumetric) in a controlled synthetic environment, assessing their impact on a YOLO-based model trained in a similar environment by varying both occlusion ratio and camera viewpoint angles. The occlusion-related metrics were derived using two distinct methods, with occlusion type treated as a separate variable. Existing background work is also harnessed to support the research, which compares and contrasts existing findings with the experimental results, highlighting novelties that arise.

Analysis of the extracted metrics reveals a strong correlation between occlusion ratio and vision model performance, which can be quantified and modeled continuously rather than at discrete levels. Occlusion ratio is found to cause the performance of the model to follow a logistic-like collapse, with different occluder types and viewpoints producing distinct degradation profiles. While the findings both build on the reviewed studies and introduce novel insights, the narrow scope of the experiment limits the applicability of the results largely to an isolated, controlled environment. The experiment shows that occlusion is not a single-variable phenomenon, and should not be treated as such, though its practical applications outside the experimental environment remain limited. A broader study of a less limited nature is required in order to apply the insights beyond this environment, allowing for the evaluation of robustness of vision models in general under more realistic conditions.

Keywords: Computer vision, Object detection, Occlusion modeling, Synthetic data, Model robustness, Camera viewpoint

PREFACE

This thesis was conducted as the result of both academic interest in computer vision and collaboration with the Finnish Defense Forces. The topic was highly relevant from an applied perspective, offering valuable practical experience while also highlighting current limitations in real-world deployment. The work integrates closely with my degree studies by combining signal processing and machine learning methodologies and has contributed both to my academic development and professional experience.

All experimental data used in this thesis was generated synthetically. While a proprietary 3D model was used as the training target, no classified or internal datasets were employed directly. Only derived model metrics were reported, and no sensitive or restricted information is included in this work.

I would like to thank the Finnish Defence Forces and the members of the hosting division for their support and for making this project possible, as well as for providing the academic freedom necessary to conduct this study. I would further like to express my gratitude to Tapio Sorvajärvi for their exceptional guidance and supervision throughout the research process, Petteri Mikkola for their invaluable assistance with the organizational arrangements necessary for conducting this study, and Ippa Seppälä for their rigorous scholarly feedback and responsibility for the implementation and operation of the vision model.

Tampere, 4th December 2025

Patrik Kakko

USE OF AI IN THESIS

I have utilised AI tools in my thesis:

☐ No
☒ Yes

The AI tools utilised in my thesis and their purposes are described below:
ChatGPT (GPT-5.0, GPT-5.1)

- Reviewing and advising on thesis structure and organization
- Polishing and refining language and syntax
- Confirming clarity and readability
- Reformulating complex phrasing

Sections where AI tools were used:

- Throughout all narrative sections of the thesis, excluding the title page and reference list.

I acknowledge that I am fully responsible for the entire content of my thesis, including all AI-assisted content, and accept accountability for any violations of applicable ethical standards.

CONTENTS

1. Introduction 1
  1.1. Background 1
  1.2. Research question and objectives 3
  1.3. Hypothesis 4
  1.4. Literature Review 4
  1.5. Advancements beyond existing research 10
  1.6. Significance of the study 13
2. Materials and Methods 14
  2.1. Research design 14
  2.2. Synthetic Environment Modeling and Model Training Datasets 17
  2.3. Experimental Setup and Main Dataset Generation 19
  2.4. Analysis techniques 23
3. Results 26
  3.1. Data presentation 26
  3.2. Observations and findings 32
4. Discussion 33
  4.1. Interpretation of results 33
  4.2. Model-specific insights and methodological implications 36
  4.3. Limitations of the study 39
  4.4. Contributions of the study 41
  4.5. Comparison with existing research 42
  4.6. Future research directions 45
5. Conclusion 47
References 49

LIST OF FIGURES

Figure 2.1. Visualization of the scripted spherical cap with camera points. 19
Figure 2.2. Camera viewpoints in the synthetic environment. 20
Figure 2.3. Occluder setup at 0.5 occlusion ratio at a camera viewpoint of 45°. The depicted 3D model is an open-source asset used for illustrative purposes only and does not correspond to the target object used for training or inference. 21
Figure 3.1. Standardized IoU plotted against occlusion ratio in the case of a solid occluder at a camera viewpoint of 20°. 27
Figure 3.2. Standardized IoU plotted against occlusion ratio in the case of a solid occluder at a camera viewpoint of 45°. 27
Figure 3.3. Standardized IoU plotted against occlusion ratio in the case of a solid occluder at a camera viewpoint of 70°. 28
Figure 3.4. IoU plotted against occlusion ratio in the case of a perforated occluder at a camera viewpoint of 20°. 29
Figure 3.5. IoU plotted against occlusion ratio in the case of a perforated occluder at a camera viewpoint of 45°. 29
Figure 3.6. IoU plotted against occlusion ratio in the case of a perforated occluder at a camera viewpoint of 70°. 30
Figure 3.7. IoU plotted against normalized occlusion ratio in the case of a volumetric occluder at a camera viewpoint of 20°. 31
Figure 3.8. IoU plotted against normalized occlusion ratio in the case of a volumetric occluder at a camera viewpoint of 45°. 31
Figure 3.9. IoU plotted against normalized occlusion ratio in the case of a volumetric occluder at a camera viewpoint of 70°. 32
Figure 4.1. Low IoU detection in the case of the volumetric occluder at a viewpoint of 20°. 35

LIST OF SYMBOLS AND ABBREVIATIONS

Abbreviations

CA-CompNet	Context-Aware Compositional Convolutional Network
CNN	Convolutional Neural Networks
COCO	Common Objects in Context
DETR	Detection Transformer
EXR	OpenEXR image format
IoU	Intersection over Union
mAP	Mean Average Precision
PNG	Portable Network Graphics image format
RGB	Red-Green-Blue color model
YOLO	You Only Look Once

Symbols and Metrics

Density	Shader parameter controlling volumetric occlusion
$\text{Density}_{\text{max}}$	Maximum density
I	Transmitted light intensity
I_0	Incident light intensity
k	Logistic slope coefficient
L (Beer-Lambert)	Path length (m)
L (Logistic)	Upper asymptote of the logistic function
Obj	Object region
Occ	Occluder region
O_{norm}	Normalized occlusion ratio
R	Occlusion ratio
T	Transmittance
x	Normalized occlusion ratio variable
x_0	Logistic inflection point
σ_a	Absorption coefficient

1. Introduction

1.1. Background

Computer vision is a rapidly developing field driven by advancing technology and its expanding range of applications, propelled in the past decade by developments in deep learning architectures and large-scale datasets. [1] One of the core tasks within computer vision is object detection, which has become a key capability in many applied domains [2]. In particular, deep learning-based vision models, such as Convolutional Neural Networks (CNNs) and transformers, have achieved remarkable accuracy under ideal conditions, while remaining limited in robustness [1]. In real-world environments, such conditions are rare, however, and therefore evaluating the robustness of these vision models is essential to ensure their functionality in practical deployments.

Object detection and classification are both central concepts that are often encountered when discussing computer vision. Although these two terms are often conflated, they are not to be used interchangeably as they denote distinct processes. In the context of this study, object detection is used to refer to both the process of recognizing and localizing an object within a scene via bounding boxes, whereas classification is only concerned with assigning the object with a class label. A bounding box denotes a rectangular region enclosing a detected object, defined by its pixel coordinates. These definitions align with standard usage in contemporary computer vision literature. [1, 2]

CNNs are deep neural networks specialized for grid-structured data, making them inherently well-suited for visual pattern recognition. On a fundamental level, CNNs consist of stacked convolutional layers that apply small, shared weight-kernels over local receptive fields. This process yields translation-equivariant feature maps that enable the network to learn hierarchical visual patterns, forming the foundation of the learned representation. These weights are randomly initialized and optimized via backpropagation using labeled training data. [3, 4]

Real environments often introduce various types of uncertainties, most prominently within visual data received by the model. Occlusion is one of the primary sources of visual uncertainty, occurring when the target object becomes partially or fully obscured [5, 6]. In this context, occlusion refers to visual obstruction within image space, where part or all of the target object is hidden from the camera’s view. Occlusion can result from a wide range of factors, including environmental elements that block the view [5], or in some domains, intentional concealment of the object [6]. The ratio of occlusion can greatly affect model performance [7], and while minimizing its occurrence is desirable, its complete avoidance is rarely feasible [6]. Consequently, understanding and modeling the effects of occlusion are critical for improving the robustness of vision models.

The effect of occlusion remains a major factor limiting the performance of vision models. Even partial occlusion leads to a loss of visual information, which can substantially reduce accuracy by disrupting the feature extraction and spatial reasoning stages of the model. [7] Contemporary state-of-the-art models such as You Only Look Once (YOLO) and Detection Transformer (DETR) still frequently misclassify or fail to detect objects entirely under moderate levels of occlusion [5, 8]. Since such models are increasingly deployed in safety-critical and structured environments, ranging from industrial robotics to autonomous driving and defense applications, their reliability is of utmost importance. [9-12]

The nature of occlusion itself also introduces additional complexity. Occluders vary widely in geometry, material, and appearance, which makes it difficult to categorize or model them in a consistent manner [6, 13]. Different occluder types may produce occlusions of similar extent but of very different perceptual severity, complicating model interpretation [2, 6]. Consequently, a systematic and controlled study of occlusion effects is essential for understanding how these variations influence model performance and for developing strategies to mitigate their impact. [7]

Due to the relatively recent development of this research area, existing work addressing the phenomenon of occlusion remains limited in scope. [6] While occluders have certainly been acknowledged in examined studies, their role has typically been treated only as a performance-degrading factor rather than as an independent variable with distinct severities and effects on model behavior. [5, 6, 14] Detection accuracy has often been the primary focus, meaning that although the influence of occlusion has been recognized, its underlying mechanisms and causes have not been examined in depth. [5, 14, 15] This limitation is also reflected in how occlusion has been measured, as comprehensive quantitative and continuous modeling remains scarce. [5, 6, 14, 15]

In the context of existing work, synthetic data does not necessarily refer to fully computer-generated content, but rather to digitally composed or augmented datasets. In practical terms, synthetic datasets are often produced either by compositing real-world image elements into new configurations or by applying procedurally generated textures and patterns to existing data to simulate occlusion. In contrast, within the scope of this study, computer-generated data refers to fully synthetic imagery rendered in a virtual environment. Furthermore, the examined studies have predominantly relied on real-world datasets or synthetic data, but only rarely on computer-generated datasets. While such data sources provide realism, they impose constraints on control and reproducibility. [14, 15] Conversely, the use of purely computer-generated imagery enables a more systematic and controlled framework for analyzing the effects of occlusion under clearly defined conditions [6, 16]. This study adopts such a framework to model and quantify how occlusion ratio and type influence detection accuracy within controlled virtual environments.

In addition to occlusion-specific effects, existing research has also shown that camera viewpoint alone can adversely influence detection performance, even when the scene content remains unchanged [17]. In real-world datasets, camera poses are typically implicit and constrained by the data collection process, limiting the extent to which viewpoint can be examined as an independent variable. In contrast, fully computer-generated environments allow camera position and orientation to be varied systematically, enabling a more controlled analysis of viewpoint-dependent robustness [16, 17]. In the context of the present study, camera viewpoint refers to the spatial position and rotation of the camera relative to the target object, defined by its location and orientation parameters.

1.2. Research question and objectives

Building on the presented background, the primary research question of this thesis is formally stated as: “What is the effect of occlusion ratio and type on the detection performance of computer vision models trained on synthetic visual environments?” This question outlines the core focus of the study, defining occlusion as the primary factor of interest in the context of computer vision models. Further, it clarifies the key variables to be assessed and the environment in which the study is conducted.

A secondary research question arises from the distinct nature of the data employed in this study, which allows the camera viewpoint to be controlled precisely. Accordingly, it is stated as: “What is the effect of camera viewpoint variation on object detection performance?” This question complements the primary one by addressing an additional factor that may influence model robustness, thereby further distinguishing the scope and methodological contributions of the present work.

The primary objective of this study is to precisely assess the impact of occluders on model performance by establishing a quantifiable measurement standard for occlusion. To support this, a secondary objective is to develop a systematic method for labeling and distinguishing between different occluder types. Achieving these objectives allows the relationship between occlusion and detection accuracy to be modeled accurately and analyzed in detail. This, in turn, enables the research question to be answered comprehensively and the effects of occluders to be understood in greater depth. To address the secondary research question, each object-occluder configuration is rendered and analyzed from multiple camera viewpoints rather than being limited to a single angle. To avoid background variation or changes in occluder orientation from influencing model performance, these factors are kept constant across all configurations. The only variable that changes between setups is the relative angle of the object to the camera itself.

1.3. Hypothesis

Outlined by the research question and observations discussed earlier, the hypothesis of this study concerns the relationship between occlusion and vision model performance. Due to the underlying nature of deep learning-based object detection, this relationship is expected to be non-linear, with model accuracy declining more sharply beyond a certain occlusion threshold. Furthermore, it is hypothesized that the intrinsic type of the occluder also influences model performance to a measurable extent.

1.4. Literature Review

In the literature review, past and current work in the field is examined. The main focus is on identifying existing research around the effects of occluders in vision models based on deep learning. Current research does cover and discuss occluders around computer vision models and their implications in general [5], with some more narrow applications. However, the impacts caused by the occluders on the model's accuracy, and more specifically the nature of the relationship, is less documented [15]. As such, the focus is on highlighting what existing research does imply about occlusions, however acknowledging that the effects on the model's performance were not always the primary focus of these studies.

For this purpose, three relevant and varied studies around computer vision have been chosen. They were chosen specifically because each revolves around the more intricate topic of occluders and their impact on vision models [5, 14, 15], while approaching different problems from different perspectives. Furthermore, they are of methodological relevance, focusing on the effects of occluders and related challenges specifically. That is not to say that the entire field would be covered by or consist solely of said studies, but rather, that the selected studies provide the essential background required to support this research. As the field specifically around the relationship between occluders and model performance has only limited coverage, this section has been organized in a per-study structure, where the findings of each study are presented and assessed individually in relation to the research question. This approach was chosen over a thematic structure to retain coherence, as research in this field is still emerging and is not yet extensive enough to support a comprehensive thematic synthesis.

The first referenced study examined was published by Guo et al. in 2025 [5], studying the significance of occluders for object detection accuracy. The main focus of this research is the relationship between the different elements present, those being the occluder itself, the occluded object, and the spatial configuration of them. The study recognizes that even well-trained models that would usually identify the object correctly struggle when occluders are present. More importantly, they also recognize that occlusions can be of different severity, and model occlusion explicitly.

For object detection, the authors used YOLO and CenterNet. In their examples, they used common, distinct objects such as a person, occluded by a motorcycle or a car, or alternatively a horse occluded by a fence. While these occluders are varying in nature, the end result was still similar: the models often failed to detect objects correctly. Considering the occluders themselves, the study does not differentiate further among occluder types. The study concludes that occlusions adversely affect the evaluated detection algorithms, but their more discreet relationships are not analyzed further. Although the authors distinguish between the different levels of occlusions and report discrete results, their evaluation remains binary: either the object is detected, or it is left undetected. Thus, while the existence of a relationship between the variables is acknowledged, its underlying form remains unexplored. [5]

To assess different levels of occlusion, the authors introduce the concept of an occlusion ratio. In practice, it is defined as the proportion of the object that is blocked by the occluder, which is represented by the equation:

$$R = \frac{Area(Occ \cap Obj)}{Area(Obj)} \tag{1}$$

where *Occ* represents the area of the occluder and *Obj* represent the area of the object. [5]

This provides a quantitative measure for the ratio of the occlusion, ranging from 0 (no occlusion) to 1 (full occlusion). Furthermore, the authors recognize that at low levels of occlusion, their models retain a higher recall and precision, whereas at higher levels both declines. The research further extends beyond model performance and also assesses the extent of the presence of occluders in natural datasets as a whole. This supports the notion that the presence of occluders is prominent rather than an edge case, translating to real life applications. Most importantly, the study presents a formalized measure for the ratio of the level of occlusion. [5]

The second referenced study reviewed was published by Zhang et al. in 2024 [14], which, similar to the first study, recognizes degrading performance of standard models when occluders are introduced. Additionally, it presents a workflow context in industrial environments. To introduce occlusion, the authors employed synthetic data, with systematically varied occlusion ratios and levels. The study also discusses several mitigation strategies, from which useful implications can be drawn regarding how occluders affect computer vision. Furthermore, the study addresses the role of context cues, concluding that they should not be relied upon, as the environment is not always indicative of the object present. In this context, context cues refer to environmental elements within the surrounding scene that may indirectly assist in labeling the object, even when the object itself is occluded.

Two types of data were utilized in the study: real-world data and synthetic data. A notable aspect of the synthetic data is how it was constructed. Instead of being artificially generated, natural images were superimposed onto an object of interest to create occluders. The data was procedurally generated with predetermined occlusion levels ranging from 0% to 90% in 10% intervals, resulting in incremental occlusion. The authors divided this data into five subsets, with subsets 1 and 2 containing no occlusion and serving as training data. Subsets 3, 4 and 5 introduced varying levels of occlusion as well as changes in lighting conditions. While lighting variation does not directly affect the occlusion level itself, it does influence the model performance by altering the background and thus the usability of context cues. Furthermore, the authors measured the level of occlusion for both the object and the background, the latter being defined as context occlusion. While this allows for assessing detection performance independently of background content, in many real-world cases the background can act as a context cue. Consequently, the authors concluded that such cues should not be relied upon independently and that the object and background should be treated separately rather than as a unified context. [14]

The way the real-world data was used in the study by the authors also suggests practical implications for balancing automatic and manual annotation. The real-world data consisted of an assembly line at a production plant, which would naturally include various occlusions by nature. Interestingly, the data was grouped into different lighting levels to assess light occlusion manually without use of automation. However, the data was further annotated semi-automatically using a trained model to create pseudo-labels, which were then later checked manually. This highlights that automation is possible, and should be striven for with large datasets, to reduce manual workload. However, due to the subjective nature of occluders, some form of manual work is often still required. [14]

Another central concept introduced in the second study is the mitigation strategies used to handle occluders. The authors introduce two distinct methods to improve accuracy in the presence of occluders, Context-Aware Compositional Convolutional Network, referred to as CA-CompNet, and Bounding Box voting. In the first method, the object and background are essentially separated from each other and assessed separately. Or in the case of the target object being overly occluded, the background may be used by itself in an attempt to identify the object. For example, a heavily occluded car might still be guessed from the presence of a road, but this proved to be unreliable at high occlusion levels. The conclusion was that with a mild occlusion, the context cues could be used to aid in detection, but when heavy occlusion was present, the context cues proved unreliable. It did, however, aid in the localization of the object. The main contribution of CA-CompNet was to separate object features from the context cues, so that the model wouldn’t base its labeling solely on the background, essentially to eliminate false positives. The Bounding Box Voting approach, in turn, was used to take advantage of the fact that a model can often output multiple bounding boxes, often fragmented or misaligned from the actual object. Instead of simply picking the top confidence bounding box, in this method multiple bounding boxes are aggregated instead, to construct a consensus-based bounding box, weighted by confidence. When tested on the Common Objects in Context (COCO) dataset, the authors reported a 4.6% increase in mean Average Precision (mAP). Furthermore, the method also reduced the presence of false negatives when no single box exceeded the threshold value. This became especially apparent at medium levels of occlusion, where models more often output multiple bounding boxes with increasing inaccuracy. [14]

The third and final referenced study to be discussed was written by Kassaw et al. in 2024 [15]. This study is particularly relevant, as it examines the robustness of vision models under effects of occlusion specifically. In the study, the authors use both real and synthetic labeled data to test the robustness of vision models. The occluders are labeled based on their severity, introducing a structured classification. The authors tested the datasets on various models extensively, not limiting their research to just a single model, in order to determine the relative performance across models. Interestingly, even the recognition accuracy of humans was not only tested, but also quantified, providing a practical baseline against vision models.

Often, synthetic data is easier to procure and adjust for the purpose of testing a vision model, but its validity over real data isn't axiomatic. The researchers directly addressed this question by validating the use of synthetic data through comparative performance analysis. It was found that while using synthetic data does increase model performance slightly, the model rankings are preserved, meaning that if one model outperformed another on real data, it would hold true for synthetic data as well. This is of importance, as it does not only validate the use of synthetic data, but also gives empirical justification for its use. However, the study does not model the occluder type more specifically, nor does it address their spatial arrangement. [15]

To diversify the results, the authors tested numerous models on the datasets, which would also mitigate the effects of any outliers. This ensured that the results were not tied to the peculiarities of any single model but instead provided a relative assessment of vision model performance under occlusion in general. It was found that across all models, degradation followed a similar consistency, as the ratio of occlusion was increased, indicating a stable relative performance. It is important to note, however, that this study is concerned with classification, rather than detection. The underlying mechanism does remain the same, however. This suggests that the phenomenon isn't limited to any single vision task, but rather, that it is generalized beyond detection. However, while the study does recognize the existence of a relationship between occlusion and model performance, the more intricate relationship between the two was not explored in a greater depth and lacked a quantitative approach to measuring their relationship. [15]

Beyond computer models, the authors also introduced humans to the study. This brings forth some unique aspects that can be used when assessing vision model performance. The authors did show that humans still outperform the models at all levels of occlusion, but that their accuracy too does degrade with increasing severity. An interesting aspect is that human performance can be used as a theoretical ceiling when contrasted to the performance of vision models, which provides a reference point of assessing the current gap between model robustness and human performance. Furthermore, the observation that human accuracy is affected by the level of occlusion as well, does suggest that the relationship is not merely artificial, but rather a universal pattern. [15]

Despite the different approaches and end goals, the three referenced studies collectively support the idea that introducing occlusion degrades model performance. The degradation in model accuracy is not random by nature but correlates deterministically by the amount of occlusion present. Correlation is established, but its exact nature, whether linear, exponential, or threshold-based, remains unassessed. It is also recognized across the studies that even well-trained and advanced models are indeed not immune to occlusions, and the impact will start to show even at relatively lower levels. [5, 14, 15] Importantly, the degradation was found to be universal among models, suggesting that it is not simply a quirk of a single model or algorithm design, but rather a fundamental limitation in data representation. The use of synthetic data was also validated for controlled experiments, as the degradation pattern of it follows that of real data. [15] Some of the more profound differences in the studies, however, come from how they approached occlusion, and what the study tried to achieve. In the first study, occlusion was modeled explicitly in order to study its spatial configuration [5], whereas the second study focused largely on mitigation strategies and workflow optimization [14]. The third study placed less emphasis on deriving occluders themselves, instead focusing on how their presence impacts the performance of vision models [15]. Furthermore, an important distinction is that the first and second study focused on object detection, whereas the third study focused on classification, but their findings in performance degradation reveal the same trend [5, 14, 15]. The way the studies define occlusion itself also differs, but the nature of the definitions is largely similar. The first study introduced a formula for occlusion ratio [5], whereas the second study used discrete ratios [14], and the third study opted for labeled ratio levels [15]. The second study also considers the concept of context cues, and whether background information can aid in labeling occluded objects [14]. The first and third study do not assess this aspect, focusing on the features of the objects occluded themselves [5], [15].

However, several important aspects were not addressed in the referenced studies, which will be examined here. Firstly, none of the studies quantified the relationship between occlusion ratio and model accuracy beyond a trivial level. While all three studies acknowledged the existence of correlation, none modeled or characterized its functional form. In addition, the effect of combining the ratio and type of occlusion was not considered. The studies largely treat occlusion as a single-variable problem, which may be an oversimplification if occluder type influences accuracy. In addition, the studies do not compare occluder types with respect to inherent geometry. Although simplifying occluders to a common geometry is a reasonable assumption in some settings, our study is going to treat the type as a separate variable, and as such they will also be cross compared. This would include assessing not only their shape and material, but also whether they are solid or transparent, and to what extent. Another distinction to be made is that while occlusion was defined by discrete steps or levels, a continuous evaluation was not undertaken. [5, 14, 15] While the formula in the first study permits continuous measurement, its evaluation remained discrete [5]. Finally, the effect of occluders on the accuracy of detecting a known object instance was only partially explored. The selected studies emphasized classification or categorical recognition tasks rather than on consistent detection accuracy across identical object instances [5, 14, 15]. This distinction is particularly relevant in the present study, which focuses on single-instance detection across systematically varied occlusion conditions. Regarding the secondary research objective of this study, the role of the camera and its spatial positioning with respect to the target object was not assessed in the referenced studies. This is, to an extent, expected, as viewpoint plays a comparatively smaller role in pre-existing real-world datasets, where camera perspectives are implicitly fixed. In contrast, in fully synthetic environments, where datasets are custom-generated and parameters are explicitly controlled, its influence can be examined more systematically. Nevertheless, the role of viewpoint variation has been acknowledged in existing research, where changes in camera poses alone have been shown to adversely affect model performance. [17]

1.5. Advancements beyond existing research

As was identified in the literature review, while existing research on the topic does exist, there were also several aspects that were left unaddressed. The following section elaborates on these gaps, outlining how this study contributes and advances beyond existing research. In particular, this section outlines how the relationship between occlusion and accuracy is quantified, how occlusion type is considered as an additional factor, how a continuous evaluation approach is applied, and how a controlled framework is established by focusing on single-instance object detection. Finally, due to the nature of the datasets used in the referenced studies, the role of camera viewpoint and its orientation in object detection have not been considered. [5, 14, 15]

The first and main contribution of this study, as outlined by the research question, is to quantify the relationship between occlusion ratio and model accuracy beyond a trivial level. In practice, this means that instead of just assessing whether the model correctly detects the object from the scene in a binary sense at a given occlusion level, the accuracy is also assessed in relation to the occlusion level and type. The accuracy of the model is plotted against the calculated occlusion level and type at different severities, enabling the relationship to be modeled and analyzed quantitatively. As such, the occlusion level and type are used as independent variables, which are systematically varied, whereas detection accuracy serves as the dependent variable, being measured. Controlled variables, on the other hand, include the aspects of the scene that are kept constant. These include the object instance, the lighting setup, the camera distance, and the model architecture. Concisely, the scene configuration functions as a controlled environment, ensuring that only the independent variables are altered.

The second important aspect that was largely unaddressed in examined works is the concept of treating both the type of occlusion and its ratio as separate variables, instead of simplifying all occluder types into a single variable. The core idea is that, rather than considering the occluder as a single variable where only the level at which it obscures the object instance is varied, the intrinsic type of the occluder is also recognized and varied systematically. In practical terms, two variables are extracted from each sample. First, the level of occlusion is quantified by dividing the occluded area of the object by its total area, following the method established in the selected studies [5]. Second, the type of occlusion, including its material, shape, transparency, and geometry, is assessed. However, it is important to note that for this study, while the occlusion ratio is measured quantitatively, the occluder type is not, as it is inherently qualitative in nature. Because the outlined properties resist simple scalar quantification, the occluder types are treated through categorical separation. For instance, in the case of a perforated occluder, a narrower study might distinguish between individual variants by comparing the total area with the area that it visually obscures. This could be used to define an ad hoc measure of occluder transparency, but within the scope of this study, it is sufficient to classify occluders categorically. That is, to distinguish between solid, semi-transparent, and irregular occluders. Of course, when assessing occluders within the same category, it can be observed that even when occluders encompass the same area, one may obscure the object to a greater extent than the other. This enables both a separate analysis of each occluder type and a meaningful cross-comparison between them.

The third considerable aspect is to distinguish between discrete and continuous occlusion measurements. As discussed earlier, although a way to measure occlusion ratio continuously was introduced, none of the studies addressed the findings beyond a discrete level [5, 14, 15]. In this study, besides measuring the occlusion ratio continuously, it will also be applied continuously, instead of using binned intervals, for instance. As the main intention is to model the continuous relationship between occlusion and accuracy, introducing arbitrary discretization such as bins or thresholds would weaken the analytical precision. Furthermore, discrete evaluation risks capturing only broad trends and missing transitional behaviors, such as non-linear performance drops. Instead, continuous evaluation allows these intricate trends to be captured. This in turn enables the fitting of analytical or regression models, which ultimately constitutes the key to answering our research question. The more precisely the relationship is modeled, the more accurately it can be quantified rather than merely observed.

The fourth of discussion arises from how the object itself is detected. In the examined works, the models were often tested across multiple object categories, and the output was considered correct when the object was classified accurately [5, 14, 15]. This approach is not inherently incorrect, but it differs from the one used in this study. Instead of classification, the present setup focuses solely on detecting the specified object instance that the model was trained to recognize. In other words, the model is not required to distinguish between different object classes or variants, but only to determine whether the target object is present and where it appears within the scene. Relying on detection rather than classification allows for a more quantitative assessment of how occlusion affects model performance.

A final point of advancement arises from recognizing the camera viewpoint as a distinct variable rather than an implicit property of the data. While this factor may be less prominent in the two-dimensional datasets used in existing studies, it remains conceptually relevant. In this study, occlusion is examined in a three-dimensional setting. By utilizing a virtual synthetic environment, both the camera and the objects can be adjusted freely in relation to one another, allowing the same object to be detected from multiple angles and the effect of viewpoint variation to be measured systematically. The training process is also considered as a model that has not been exposed to all relevant object angles during training would naturally be expected to perform less accurately when encountering unfamiliar viewpoints.

Recognizing the aspects discussed clarifies how this study diverges from existing research and builds upon existing knowledge through new methodological approaches. By adopting the methodological distinctions outlined earlier, this study enables a more controlled and quantitative analysis of occlusion effects. This not only helps in answering the research question, but also in understanding why and how the chosen methods are applied in this study.

1.6. Significance of the study

The academic significance of this study lies in its exploration of the previously unquantified relationship between occluders and vision-model accuracy. A deeper examination of this relationship provides novel insights that can advance future research on model robustness under occlusion. While these methodological advancements were outlined in the previous section, this section instead focuses on their broader significance and impact.

Starting with the quantification of the relationship between the outlined variables, the significance of this lies beyond merely establishing a measurable foundation for understanding model degradation. By enabling robustness to be modeled as a continuous and quantifiable function, the relationship between the variables can be analyzed with far greater precision. In turn, this enables a deeper understanding of how these variables interact, reinforcing the research question of this study.

Beyond the relationship between the variables, the way in which the occluder is defined and treated as a variable is of particular significance. Distinguishing and separating the different dimensions of occlusion prevents the oversimplification of its nature and encourages more nuanced research around its effects. Furthermore, adopting a continuous form of measurement over discrete intervals not only increases analytical precision, but also supports the detection of subtle non-linear effects, improving interpretability and overall depth of analysis.

Finally, training the vision model on a single object instance and focusing on detection rather than classification provides an isolated perspective on the effects of occlusion itself, minimizing the influence of unrelated factors such as category noise. The significance of this lies in offering a more precise and relevant benchmark for evaluating model behavior, eliminating confounding variables and concentrating the analysis on the core phenomenon of interest.

2. Materials and Methods

2.1. Research design

To conduct the practical experiment of the study and measure the metrics, a fully controlled virtual synthetic environment is initialized. The experiment follows a factorial design, in which multiple independent variables are systematically varied to examine both their individual and combined effects on model performance. A distinct scene is built and configured for each occluder configuration, which allows for precise control over its elements without interference. Environmental conditions are kept constant across scenes, which ensures comparability across scenes, and supports consistent quantitative analysis of the results. The experiment incorporates various independent, dependent, and controlled variables, each of which is defined and described in the following sections.

Independent variables represent the controlled parameters in the virtual environment, which are systematically adjusted per scene in order achieve a desired configuration to observe their influence on model performance. The variables examined in this study are occlusion ratio, occlusion type, shader density, and camera viewpoint.

Occlusion ratio is a continuous quantitative variable, which represents the extent to which the target object is occluded within the image. It is determined in two complementary approaches, depending on the occluder type. For the solid and perforated occluders, a geometric method is applied to compute the ratio between the occluded area and the total visible surface area of the object. The resulting value ranges from 0 to 1, where 0 denotes no occlusion and 1 denoted full occlusion. The measurement is derived from object index masks, which are described in detail in Section 2.4. For the volumetric occluder, attenuation is used to describe visibility reduction through a participating medium. Due to the medium's optical properties, visibility in this case decreases exponentially, and never reaches zero. Again, the specific implementation and derivation of the attenuation model used is detailed in the analysis section. Occlusion ratio, as a variable, does however provide a quantitative basis for assessing the ratio of the occlusion present in each configuration, and enables a systematic evaluation of its influence on the performance of the model.

Occlusion type is a categorial qualitative variable, which describes the intrinsic characteristics of the occluder present in the scene, beyond the extent of visual coverage. The categories of the occluders correspond to the distinct types employed in the experiment. To ensure diversity in structural and visual complexity, three distinct occluder types were selected. These types are solid, perforated, and volumetric. The solid occluder is dynamic, meaning that it is moved in the scene, whereas the perforated and volumetric are static, with only their intrinsic parameters adjusted parametrically. The occluders themselves are adjusted in each scene to achieve a full range of occlusion from 0 to 1 for all types, with each scene containing only a single occluder at a time to maintain experimental control.

Camera viewpoint, defined by the camera's extrinsic parameters, is a categorical discrete variable comprising three angular configurations. These elevation angles are 20°, 45°, and 70°, defined relative to the object's horizontal plane. In this setup, 20° corresponds to a low-elevation lateral viewpoint, whereas 70° corresponds to a high-elevation view approaching the zenith axis. These angles were chosen to represent low, medium, and high perspective viewing perspectives. In practice, displacing the camera itself in the scene to achieve viewpoint variation would also require for the occluder and background to be adjusted relatively as well to minimize their influence on the model. Instead, the target object itself is rotated around its perpendicular axis to achieve the same effect. This ensures that the occluder, background and lighting remain fixed throughout different angular configurations, with only the orientation of the target object itself being rotated relative to the camera. Importantly, viewpoint angles are defined by their relation to the horizontal plane of the object rather than the plane of the scene, ensuring that they are treated as controlled experimental variables. The camera setup is illustrated visually in Section 2.3.

Dependent variables are the variables which are measured and that change as a function of the independent variables. These variables represent the measurable outcomes and metrics from which quantitative results are derived and analyzed. The dependent variable in this experiment is detection accuracy.

Detection accuracy is treated as a quantitative continuous variable that measures the correspondence between the model's predicted detections and the ground truth annotations. In this study, detection accuracy is measured using two key metrics, which are Intersection over Union (IoU) and detection confidence. IoU measures the spatial overlap between the bounding box predicted by the model and the ground truth bounding box and is defined as the ratio of their intersection area to their union area. Detection confidence is the model's internal measure of prediction certainty, defined by its estimated probability that the predicted bounding box corresponds to a true positive instance. Together, these two metrics allow for a quantitative assessment of how the independent variables influence model performance.

Controlled variables are parameters of the scene that are defined once and kept constant through different configurations. The controlled variables in the experiment include the background, lighting, object instance, intrinsic camera parameters, and render settings. By keeping these factors constant, potential variations in detection accuracy caused by them are minimized, ensuring that only the independent variables influence the results.

Background is treated as the surrounding environment of the object instance, and it is kept unchanged across configurations. This ensures that the influence of contextual cues on dependent variables is minimized.

Lighting is likewise kept consistent throughout different configurations by employing a uniform lighting setup with fixed intensity and direction. This further isolates the effects of occlusion from other factors.

The object instance, representing the target object itself, is single 3D model kept constant across all the scenes. In addition to maintaining a standardized geometry for detection, the material properties are kept uniform. This ensures that any variation in model performance would arise solely from occlusion and camera viewpoint.

Intrinsic camera parameters refer to the internal optical characteristics of the camera. Although the camera is in a synthetic environment, it is still defined by the same intrinsic parameters as a real camera. In this setup, the key parameters are focal length, field of view, and sensor size. Keeping these parameters as constant ensures consistent geometric projection across scenes.

Render settings are what define the characteristics of the data generated from the environment. All the settings are kept unchanged, however, perhaps the most influential parameters for model performance are the resolution and sample size of the renders. While minor variations other variables may influence visual quality, altering these two would compromise consistency and comparability across scenes.

Based on the outlined variables, the primary relationships examined are those between detection accuracy and the independent variables. These consist of occlusion ratio, occlusion type, and camera viewpoint. Importantly, the metrics are obtained only after model inference, following the rendering of all images. No data is derived from the virtual environment itself, only the rendered outputs are used for analysis.

2.2. Synthetic Environment Modeling and Model Training Datasets

The experiment relies extensively on synthetically generated materials and data, which were all produced for this experiment specifically. Blender, a free 3D computer graphics software, was used to create the synthetic environment itself, including the camera setup, as well as the occluders. Blender is suited for the purpose of this study, as it is open source, ensuring that the replicability of this study isn't limited by exclusive software. Furthermore, it is scriptable, allowing for seamless implementations of scriptable logic, and it provides deterministic renders. The only external data sources employed in the experiment were the base 3D model for the target object, and the computer vision model used for inference. For each configuration and frame, the generated dataset consists of an RGB beauty render used for model inference, an object index mask for ground-truth bounding box extraction, and the detection confidence scores produced by the model. IoU values are calculated from the bounding boxes predicted by the model.

Starting with the materials themselves, first is the visual environment. This refers to the scene in which the configurations take place, and it is intentionally kept rather simple, to ensure only the variables we adjust affect the model performance. Essentially, it only consists of background that separates the target object from the rest of the space, a simple light source for lighting, the camera, and the occluders. The background is made to resemble a sort of oversimplified gravel backdrop. It is not meant to be realistic, but rather, contrasting enough to separate the target object from the rest of the scene. The light source is a simple point lamp, meaning it radiates light uniformly in all directions. The camera is used to take the renders and adjust the render output. This is done through setting a resolution for the renders, which in this case is a 1:1 image of 640x640 resolution. As for Blender-specific settings, the render engine is set as Cycles, and the sample size at 32 for beauty shots, and 1 for object index masks to conserve resources. The camera is also adjusted when switching between beauty and object index mask renders.

The occluders themselves are the final materials to be discussed. The first occluder is an opaque one by type. This can be categorized as a solid occluder, having no transparency or perforation. Fully opaque structure by nature, it ensures total obstruction of the target object in the occluded region, with the occlusion ratio being adjusted by displacing the occluder relative to the target object. The second occluder is more structured by type. This occluder can be categorized as a perforated occluder, as it employs partial masking by discrete visible and occluded areas. The extent of perforations is controlled iteratively by adjusting the level of perforations. The occlusion ratio is adjusted by this alone, and the occluder is not moved relative to the target object. The third and final occluder is uniquely transparent by type, where instead of controlling its geometric properties, its material is controlled instead. This can be categorized as a volumetric occluder, as it has no innate geometry. Instead, it is based on diffusion to scatter light, occluding the target object. This occluder is not moved relative to the target object either, but instead, the occlusion ratio is controlled by the transparency of the material.

The external data sources in this experiment were limited, as the intention was to use create custom materials wherever suitable. However, as was mentioned, the model for the target object, as well as the computer vision model used, were external. The target object used is an armored combat vehicle, proprietary model owned by the Finnish Defence Forces, and its use is therefore restricted. While it was not custom created for this experiment like the other materials, it was manipulated to suit the experiment. This included creating a custom material for it, as well as adjusting it for the scene itself. As for the vision model, YOLOv11 was used, with a confidence threshold of 0.25. The model was trained to detect this target object specifically, instead of relying on a pretrained model. Training, validation, and test datasets were custom created for the vision model separately, and the model was trained with them. The model was trained with five datasets: three for training, one for validation, and one for testing. Each dataset consisted of three separate scenes of 440 frames covering the target object from 30° to 90° vertically, and a full 360° coverage horizontally. Each dataset consisted of 1320 frames, meaning the model was trained with 6600 custom images in total. The material was created using a script that moves the camera around the target object, tracking it in a spherical cap shape. This shape encapsulates the target object from above, covering the region from 30° elevation down to 90° elevation, illustrated visually in Figure 2.1. The model was therefore not exposed to viewpoints below the target object, which follows directly from the vertical coverage of the spherical cap. Furthermore, because points near the rim of the spherical cap span longer arc lengths relative to the object, the script was adjusted to render each frame at the same relative displacement along parallels of the spherical cap, to ensure variety in the frames. As the rings shrink in circumference toward the top of the spherical cap, the sampling procedure results in more frames being rendered at the 30° elevation region of the spherical cap. Consequently, fewer frames are rendered at the 90° elevation along the positive Z-axis. As such, the vision model is exposed to target object more at the 60° elevation angle, which can lead to higher relative robustness from this region.

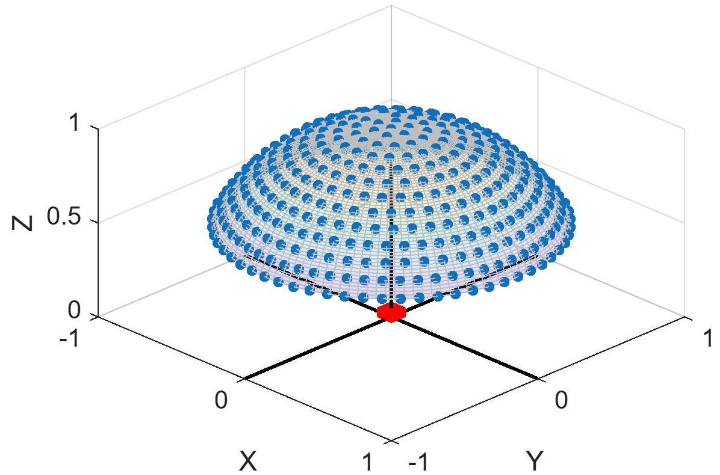

__Figure 2.1.__ Visualization of the scripted spherical cap with camera points.

In Figure 2.1, the spherical cap used in the creation of the material has been modeled. Here, the blue dots along the spherical cap indicate the positions the camera goes through iteratively, and the red cube at origin represents the target object. The scale is relative, but both visually and mathematically the spherical cap corresponds to that used in the material creation process.

2.3. Experimental Setup and Main Dataset Generation

To enable the experiment to be repeated faithfully, the setup of the synthetic environment and the generation and capturing of the experimental data is outlined in detail. This includes outlining the setting up of the scenes, their configuration, and the rendering pipeline. Emphasis was placed on automation, utilizing scripting where applicable, to ensure reproducibility and consistency.

First, a background terrain was crafted for the scene, to separate the target object from the rest of the space. The background itself is, by intention, simple in design and material, as its main functionality is to distinguish the target object, not add to it. The background is a simple half-sphere, with an elongated base to fit the target object. The material used is reminiscent of a gravel-like texture and was created utilizing noise.

The light source used is an omnidirectional point light source with a bright, neutral color to evenly brighten the whole scene, but not tint anything. The light source emits isotropically providing uniform illumination across the scene. Most importantly, this ensures that all the sides of the target object visible to the camera are clear, and not overly dark as would be the case if using a directed light source.

The camera is placed laterally to the target object, so that its optical axis is perpendicular to the frontal axis of the target object. This allows the relative viewpoint to be controlled directly, as rotating the frontal axis of the target object effectively transforms the viewpoint. With this approach, to alter the viewpoint, only the target object has to be rotated, and everything else can be left as is.

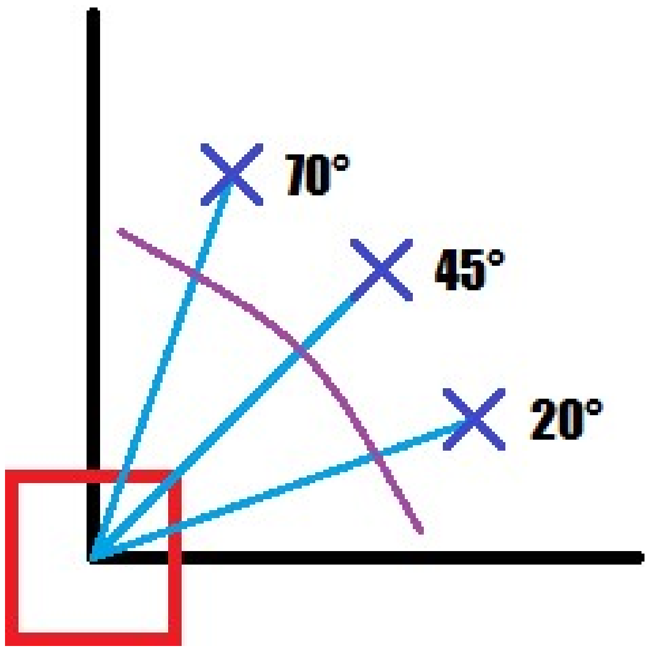

Figure 2.2. *Camera viewpoints in the synthetic environment.*

Figure 2.2 illustrates the concept of how the camera is setup in the environment. The red square represents the object, the purple curve represents the occluder, and the blue crosses represent the three camera viewpoints. As was outlined in Section 2.1, the camera itself is not moved, only the object is rotated. However, this figure outlines how the camera viewpoints relate to the object, and what was functionally achieved while keeping other variables controlled.

Each occluder exists in the same scene simultaneously, but only one is rendered per configuration. This allows for everything to exist in the same scene instead of separate files, allowing for a more seamless workflow. The occlusion functionality is keyframed for the solid and volumetric occluder, while the logic of the perforated occluder is handled with a script. In the case of the solid occluder, its movement is keyframed to start from one side of the object, and end on the other, fully passing over the object and occluding it in the process. In this case, the occluder is moved from the front to the rear of the target object, as can be seen in Figure 2.2. The perforated occluder is scripted to iteratively add horizontal and vertical divisions to an occluder frame outlining the object. This means that the occluder starts with only the thin outline of the occluder, essentially one big perforation. As additional grid lines are introduced, the number of perforations increases while their individual size decreases, resulting in greater overall occlusion of the object. This means that occlusion is introduced by discrete structural changes as the grid line are added, but that each iteration redefines the occlusion pattern spatially in a non-monotonic manner, as some regions occluded in the previous iteration are not occluded in the next iteration. This is due to the fact that the grid lines also move spatially in each iteration. As such, while the total amount of occlusion increases monotonically, its spatial distribution does not. For the volumetric occluder, the density of its material shader is keyframed to increase, eventually fully occluding the target object. Importantly, all the animations and iterations follow a linear interpolation, ensuring consistency.

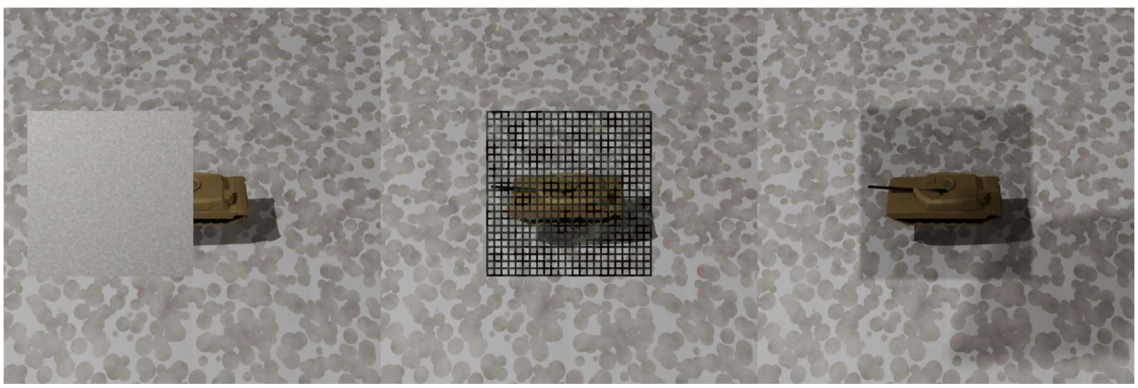

Figure 2.3. *Occluder setup at 0.5 occlusion ratio at a camera viewpoint of 45°. The depicted 3D model is an open-source asset used for illustrative purposes only and does not correspond to the target object used for training or inference.*

Figure 2.3 shows the synthetic experimental setup, illustrating the occluder configurations. From this, the different visual characteristics and properties of the occluders can be seen. It is important to note, however, that the model used in this figure is a placeholder intended purely for demonstration. For the case of the solid and geometric occluder, 50% of the projected geometry of the model is occluded. In the case of the volumetric occluder, 50% of the original intensity of the light is transmitted through the occluding volume.

Once the occluder logic is set up, each configuration is rendered from three different angles. The scene spans 300 frames per configuration, and all parameters are held constant throughout the scene except for the relative angle between the target object and the camera. With three occluder types and three camera viewpoints, this results in a total of nine configurations. For the solid and perforated occluders, composition logic and render settings are adjusted between the beauty renders and the object index mask renders. For the volumetric occluder, the base object index mask requires its own composition logic and render settings as well. However, between the beauty and EXR renders, it is sufficient to modify only the output properties.

The resulting renders constitute the main experimental datasets, which are then processed by the vision model to obtain the desired performance metrics, including IoU and detection confidence. For the solid and perforated occluders, each frame produces one beauty render and two object index masks, which are used to compute the occlusion ratio geometrically. For the volumetric occluder, each frame produces one beauty render, a non-occluded reference EXR image, and an occluded EXR image, allowing the occlusion ratio to be computed volumetrically. A clean object index mask is rendered for all configurations to provide ground-truth annotations. Occlusion ratios for both geometric and volumetric approaches are derived directly from the rendered imagery prior to model inference. Each configuration consists of 300 frames, yielding a total of 2700 base samples. For each sample, a beauty render and a clean object index mask are produced, with additional occluded masks or volumetric EXR passes rendered as required. Beauty renders are stored as 8-bit PNG files, object index masks as 1-bit PNGs, and volumetric passes as 32-bit EXR images. All files are indexed by frame number and configuration and named accordingly to enable automated dataset parsing during analysis.

Although Blender largely automates rendering animations innately, the output was setup to organize the produced renders. In this case, each dataset of each configuration was outputted into their own directory. This included organizing the datasets firstly based on the occluder type, then by image type, and finally per angle.

Following the rendering process, object index masks are processed into bounding boxes, to acquire the ground truth. Before the datasets are inferred by the vision model, the occlusion ratios are computed as well, both for the geometric and volumetric approaches. The datasets are then passed through the vision model for inference, after which IoU and confidence are extracted for each configuration. This data is exported in a structured manner to enable further analysis. As the size of the data remains very manageable, both the produced renders and the inferred data are checked for proper occluder alignment and bounding box correspondence.

Metric acquisition occurs at the end of the pipeline, and any structured data is acquired only after the rendering process. The occlusion ratio is the only metric measured that does not require any model inference to be acquired, as it is derived directly from the rendered datasets.

2.4. Analysis techniques

The main analyses employed in the experiment revolve around how the occlusion ratio is measured, how the bounding boxes are extracted, model inference and metric extraction, as well as data structuring and visualization.

The geometric approach computes occlusion ratio from the clean object index mask with no occlusion, whereas the volumetric approach compares the transmitted and incident intensity of the light. The occlusion ratio between the different types of occluders is thus parametrized differently, but the ratio is calibrated to a scale from 0 to 1 to ensure comparability and consistency across configurations.

Object index mask, in this case, refers to creating a mask based on the indexes on the object. In this case, the target objected is given a different index from everything else present in the scene. For example, the target object could have an object index of 1, and the rest of the models present in the scene could have an object index of 0. We then set the object index to represent the alpha value of the object in the render, and with this setup, it causes everything indexed at 0 to be black, and everything indexed at 1 to be white, creating an object index mask. The first object index mask is a clean one, where the occluder is not present. This works as a ground truth to compare to. The second one has the occluder present, to model its effects. Comparing the visibility of the object in these object index masks allows us to compute the occlusion ratio geometrically.

In the case of the volumetric occluder, occlusion ratio is defined a bit differently. While adjusting the density of the shader controls the occlusion, it is measured by computing the transmittance from the EXR renders. This is achieved by comparing the base and occluded EXR outputs, providing us with the intensity ratio. Density is the factor by which the volumetric material absorbs lights, and in this setup, it follows the Beer-Lambert law [18]:

$$I = I_0 e^{-\sigma_a L} \tag{2}$$

Here, I is the transmitted intensity, I_0 is the incident intensity, σ_a is the absorption coefficient, and L is the length. From this, we multiply dividing sides with I_0:

$$\frac{I}{I_0} = e^{-\sigma_a L} \tag{3}$$

to derive our measured variable, the transmittance

$$T = \frac{I}{I_0} \tag{4}$$

As the occluder employing this shader is exactly one meter in width on the camera's and the target object's axis, the density scales the absorption coefficient directly, meaning that

$$\sigma_a = \text{Density} \tag{5}$$

As the absorption color is set as pure black, meaning the occluder absorbs all wavelengths of light equally, we can express transmittance as

$$T = e^{-\text{Density}*L} \tag{6}$$

which, since we use a constant length of 1 along the projected axis, the expression simplifies to

$$T = e^{-\text{Densit}} \tag{7}$$

which allows us to calculate transmittance directly from the density, which we manipulate.

As density is manipulated from 0 to 5, we can calculate the range of the occlusion ratio in this case:

$$T = e^{0} = 1 \tag{8}$$

as the lower bound, where all light is transmitted through and there is no occlusion, to

$$T = e^{-5} \approx 0.00674 \tag{9}$$

meaning only around 0.674% of the original light is transmitted through the occluding volume. As transmittance follows an exponential curve, if the occlusion ratio is directly derived from transmittance by taking its complement $1 - T$, the occlusion ratio would in this case follow an exponential curve as well. Due to this, when plotting the volumetric occluder, a normalized occlusion ratio is instead used, defined by taking the fraction of maximum density. From the formula for transmittance, we can derive the definition of density as

$$\text{Density} = -\ln T \tag{10}$$

which we can further use to define a normalized occlusion ratio

$$O_{\text{norm}} = \frac{\text{Density}}{\text{Density}_{\text{max}}} = -\frac{-\ln T}{-\ln T_{\text{min}}} \tag{11}$$

as density ranges from 0 to 5, and the minimum transmittance is achieved with the maximum density used.

The bounding boxes are extracted from the object index masks in a rather straightforward manner. The highest and lowest values for both the x and y coordinates of any pixel with the index of 1 are extracted, outlining the extent of the projection of the target object. For the clean mask, this provides the ground truth, and for the occluded mask, this provides the visibility after occlusion in the geometric approach, enabling the rate to be computed. By extracting the bounding box based on the actual pixels projected to the camera, a pixel-perfect ground truth is enabled for IoU comparison.

In contrast to the perforated and volumetric occluder, the solid occluder introduces occlusion directionally, progressively obstructing the target object from one side rather than uniformly across its geometry. This causes the visible area to decrease monotonically but non-uniformly, and while the occlusion ratio can still be calculated without an issue, the bounding box predicted by the model is misleading. Since the vision model hasn't been taught that the target object can exist beyond what is visible to it, it interprets partially occluded regions as truncated object boundaries, even when the ground truth bounding box spans beyond the occluder. This causes the IoU to deteriorate, but not necessarily because the vision model is misplacing it due to inaccuracy, but rather due to this technicality. In order to provide a valid metric baseline, the IoU is normalized for the case of the solid occluder, by multiplying the inverse of the occlusion ratio with the IoU. With this, penalization from the invisible occluded region of the ground-truth is removed.

Each image of each configuration is inferred by the vision model, which outputs a predicted bounding box for it, along with an associated confidence score. This allows for the IoU to be extracted, which is computed by taking the area intersection of the ground truth and the predicted bounding box, and dividing it by the area of their union. The other value provided by the model is detection confidence, which in this context indicated the probability of the detection being a true positive. Only the highest probability prediction is extracted, and the vision model tends to lose the target object before any confidence threshold is approached.

Finally, the data is extracted in a structured manner, and labeled neatly into a single file, per configuration. The file includes the frame index, visibility ratio, confidence, and IoU. This enables free examination of the values, and further mathematical analysis, such as plotting the values.

Separate plots are done per occluder configuration. For the solid occluder, the standardized IoU is plotted against occlusion ratio, whereas normal IoU is used with the perforated and volumetric occluder. Furthermore, due to the unsymmetrical way the solid occluder introduces occlusion, the plot is split to frames 1 to 150 and 151 to 300, to highlight the difference to vision model performance depending on which side is obstructed first by the occluder. The rest of the occluder configurations use the standard IoU and plot normally from frames 1 to 300.

3. Results

3.1. Data presentation

The following section presents the quantitative results obtained from the experiment. The results are categorized by occluder type and camera viewpoint, yielding nine configurations in total. Data visualization employs a logistic function, selected for its ability to model threshold-like and saturation behavior characteristic of detection performance as occlusion ratio transitions from low to high values.

The logistic function is defined as:

$$\mathrm{IoU}(x) = \frac{L}{1 + \mathrm{e}^{-k(x-x_0)}} \tag{12}$$

where L is the upper asymptote, in this case represented by the maximum attainable value of IoU. The variable k is the slope coefficient, which also governs the steepness of the transition. The variable x is the occlusion ratio, and x_0 is the inflection point at which IoU equals half of its asymptotic value.

The logistic model is not intended to interpolate or represent individual detection outcomes but rather to describe the mean robustness trend by providing a smooth envelope over the data for visual interpretation.

For the solid occluder, standardized IoU is plotted against occlusion ratio, for all three camera viewpoints. As outlined in Section 2.4, the plots are also split. The plot with the label A consists of frames 1 to 150, whereas the plot with the label B consists of frames 151 to 300.

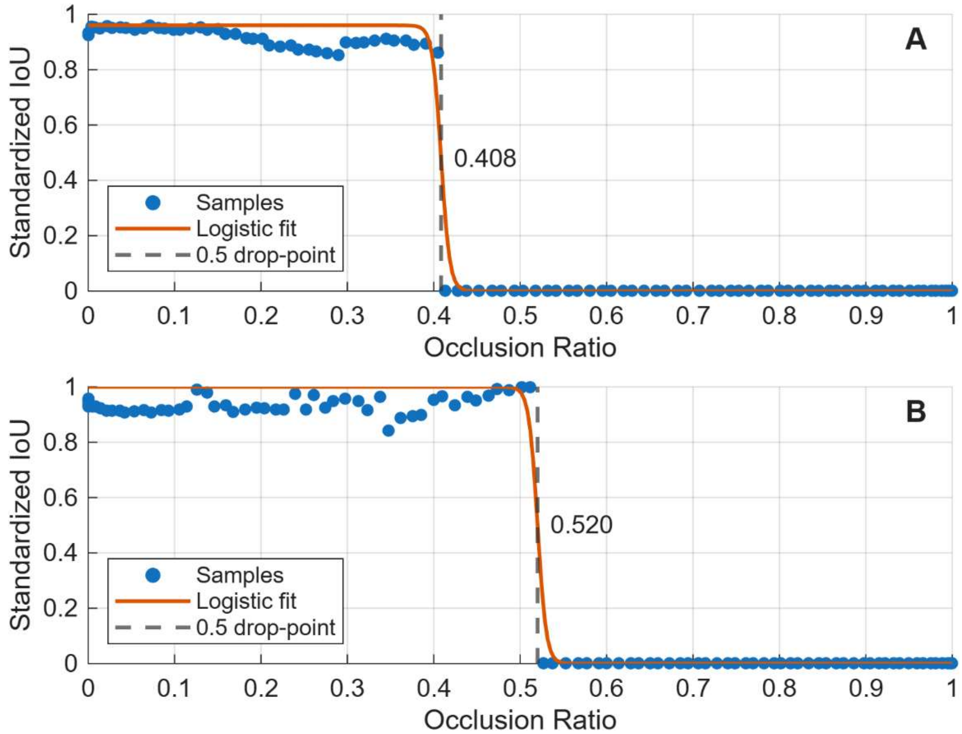

Figure 3.1. *Standardized IoU plotted against occlusion ratio in the case of a solid occluder at a camera viewpoint of 20°.*

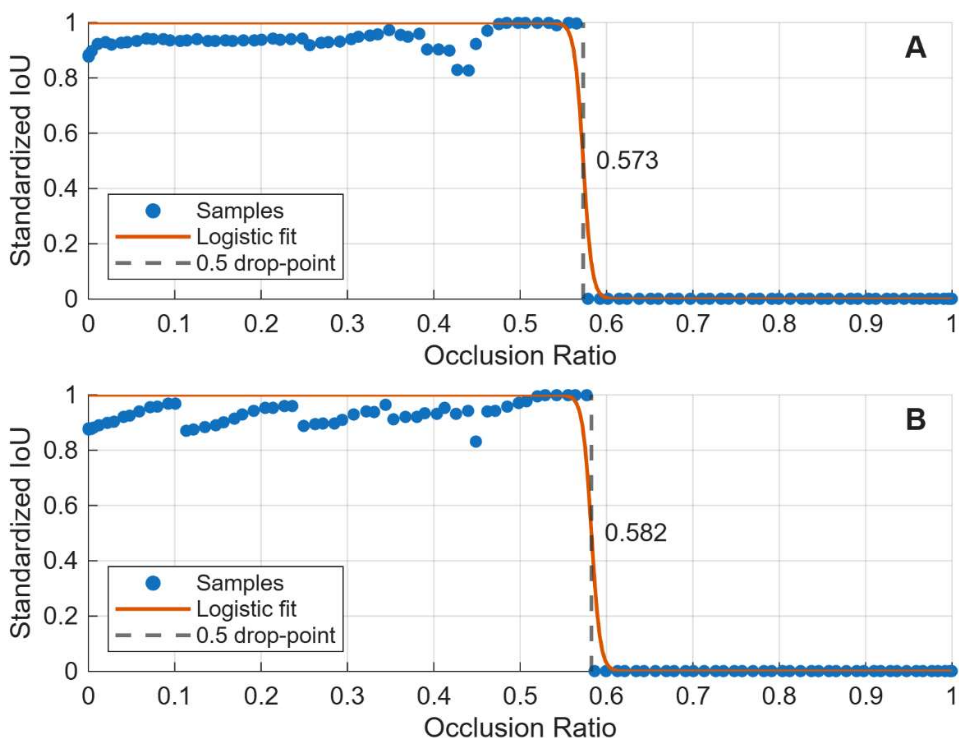

Figure 3.2. *Standardized IoU plotted against occlusion ratio in the case of a solid occluder at a camera viewpoint of 45°.*

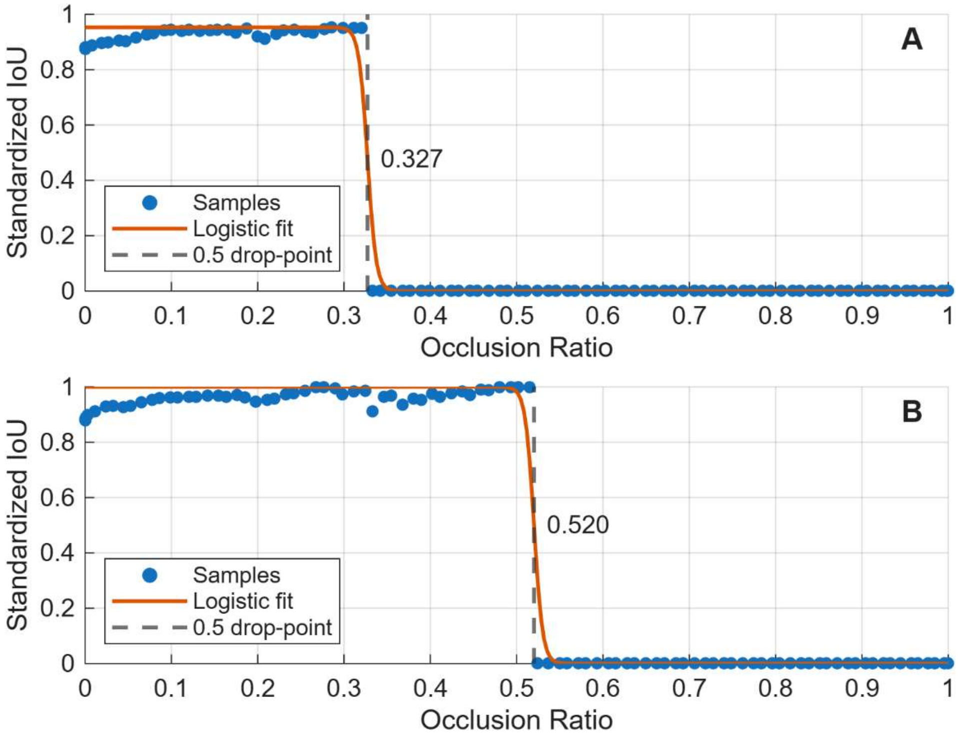

Figure 3.3. *Standardized IoU plotted against occlusion ratio in the case of a solid occluder at a camera viewpoint of 70°.*

For the perforated occluder, IoU is plotted against occlusion ratio, for all three camera viewpoints.

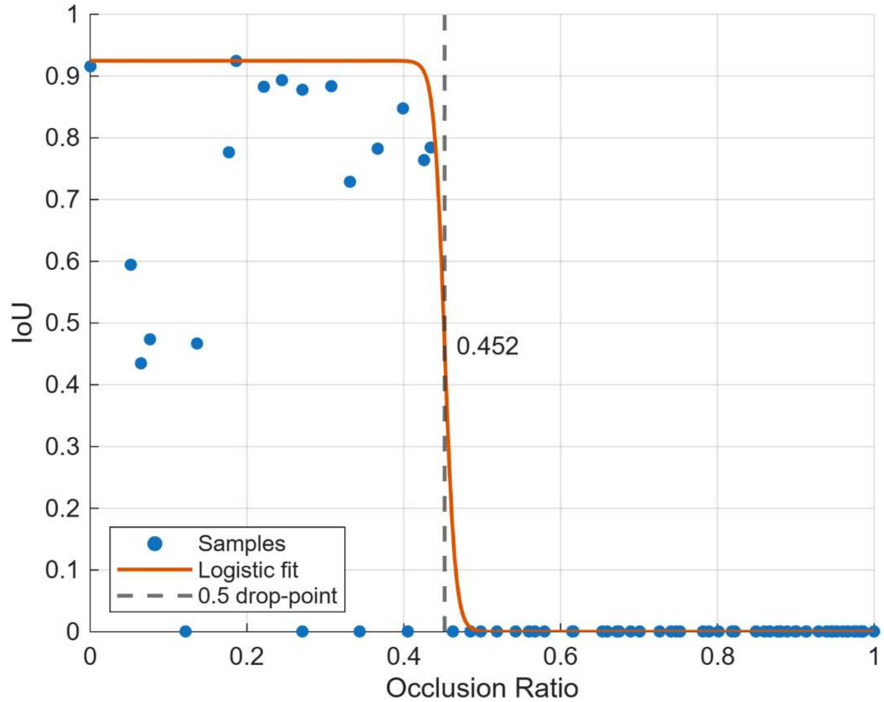

Figure 3.4. *IoU plotted against occlusion ratio in the case of a perforated occluder at a camera viewpoint of 20°.*

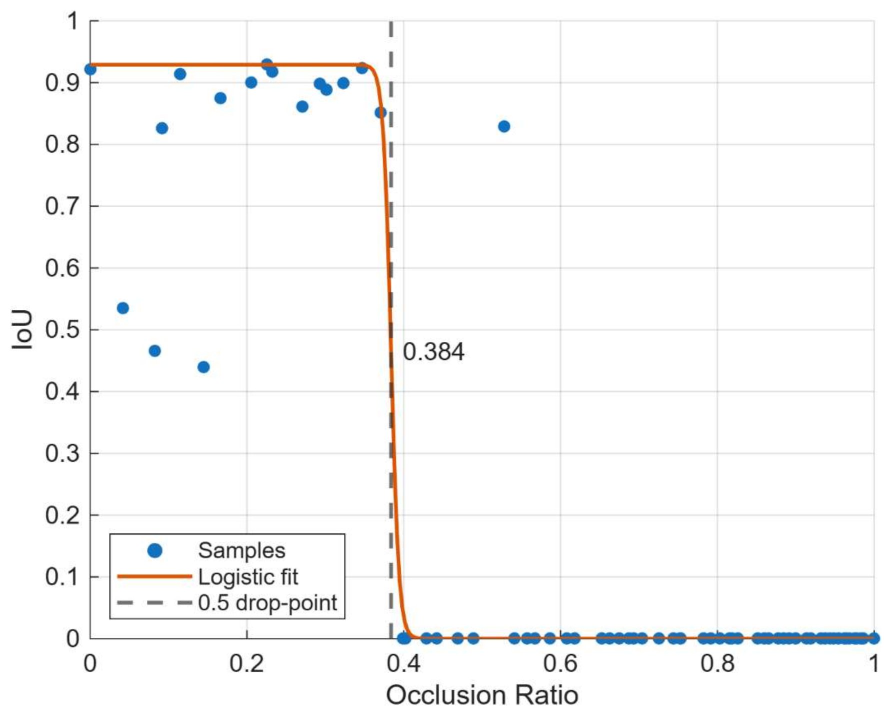

Figure 3.5. *IoU plotted against occlusion ratio in the case of a perforated occluder at a camera viewpoint of 45°.*

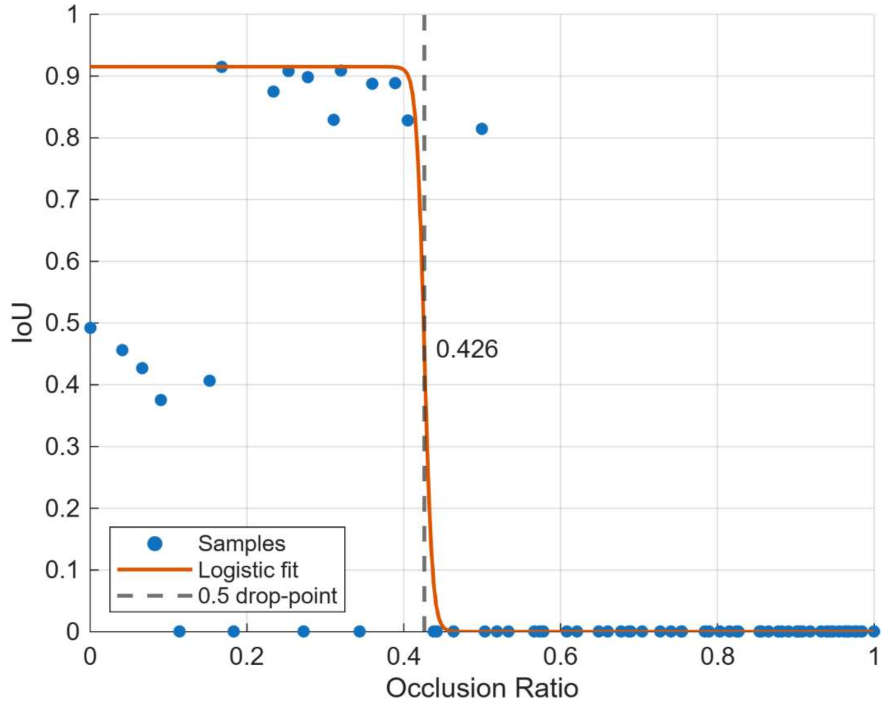

Figure 3.6. *IoU plotted against occlusion ratio in the case of a perforated occluder at a camera viewpoint of 70°.*

For the volumetric occluder, IoU is plotted against the normalized occlusion ratio derived from volumetric transmittance, for all three camera viewpoints.

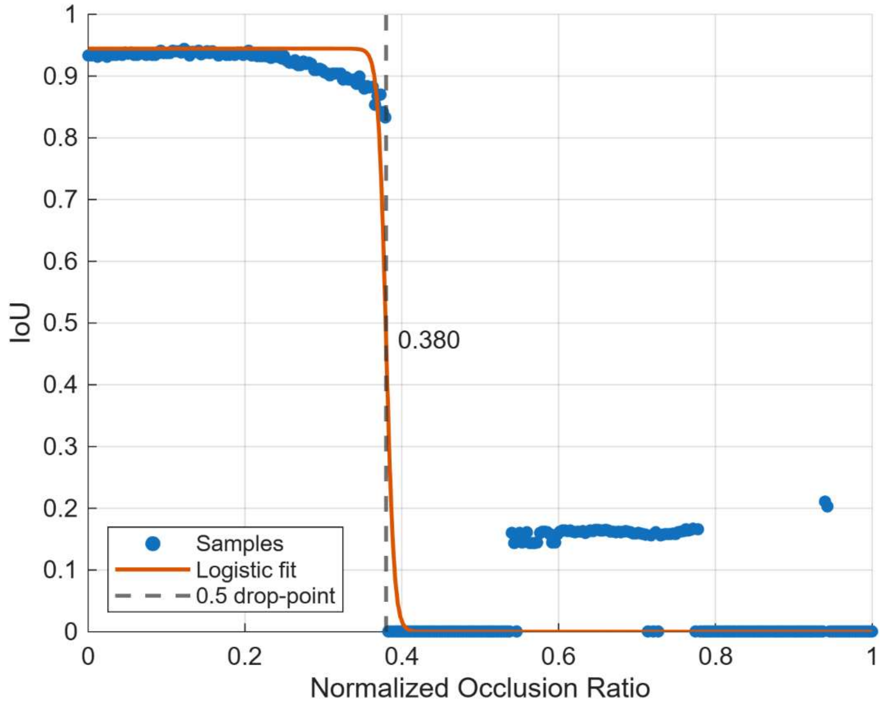

Figure 3.7. *IoU plotted against normalized occlusion ratio in the case of a volumetric occluder at a camera viewpoint of 20°.*

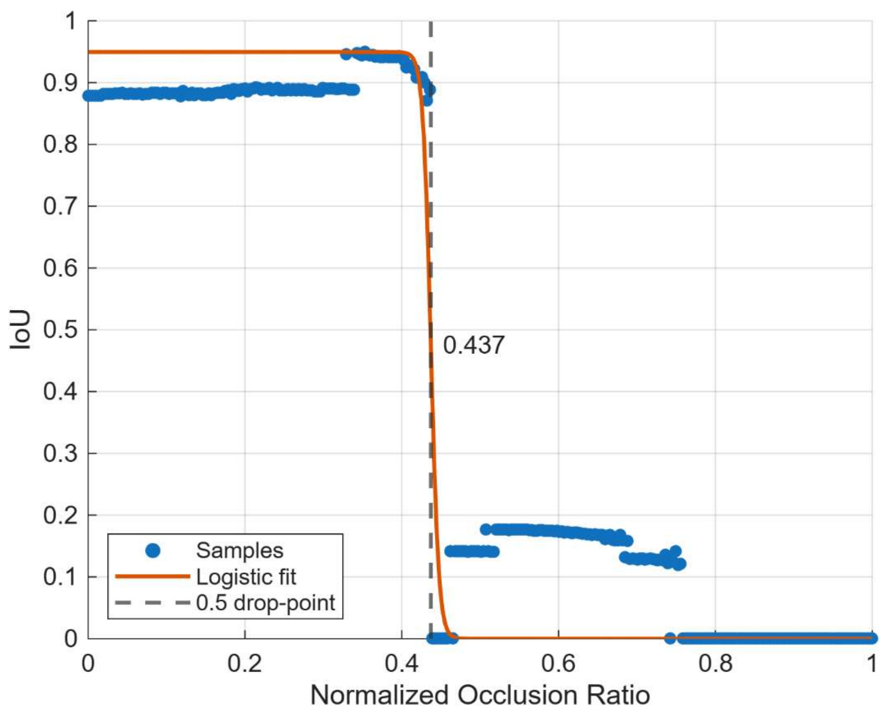

Figure 3.8. *IoU plotted against normalized occlusion ratio in the case of a volumetric occluder at a camera viewpoint of 45°.*

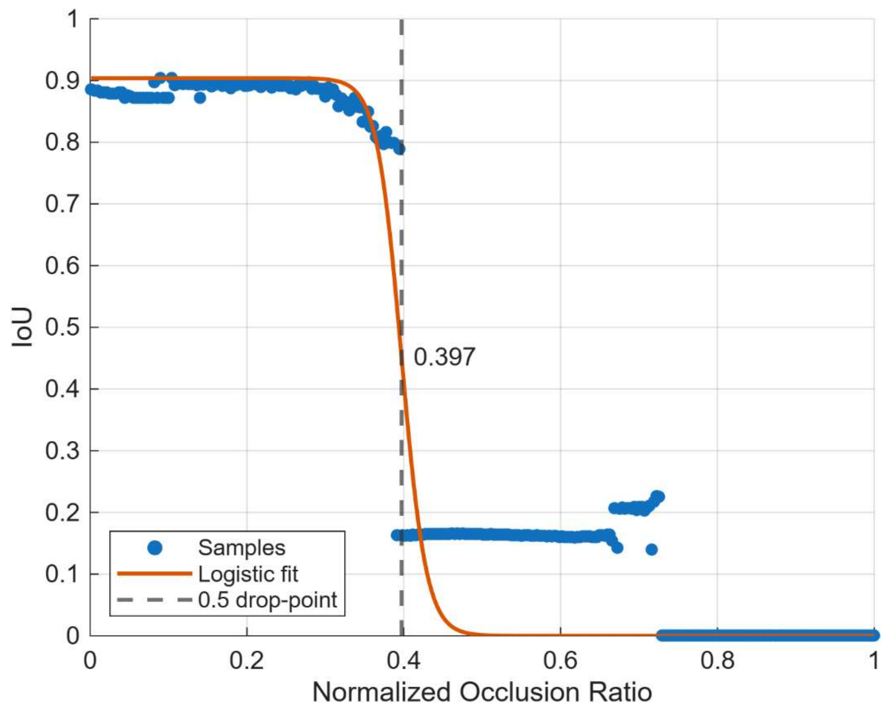

Figure 3.9. *IoU plotted against normalized occlusion ratio in the case of a volumetric occluder at a camera viewpoint of 70°.*

3.2. Observations and findings

This section will focus on the immediate observations and prominent findings from the data, before going into deeper interpretation. This includes any high-level patterns, differences between occluder types and camera viewpoints, as well as any anomalies.

Across all occluder types and camera angles, a sharp drop can be observed at a certain point. Although this point does vary by configuration, it is clearly present. Although this point varies across occluders, it can be seen that when comparing between camera viewpoints of a certain occluder, these points are often grouped. For the solid occluder, this point occurs within the occlusion ratio of 0.327-0.582, the largest deviation present. As for the perforated occluder, the point lands in the range 0.384-0.452. For the volumetric case, this range is 0.380-0.437. For the most part, the plots exhibit a stable high-IoU plateau before the drop-off point, and a low-IoU plateau after it. This is to be expected, as when confidence drops below the threshold, detections disappear entirely.

The solid occluder shows the most uniform sample distribution across all occluder types, with the samples mostly fitting with the curve, and a sharp drop-off point present. Further, the average drop-off threshold appears higher than in the perforated or volumetric case. The perforated occluder shows the highest variance among all the types, with the least obvious drop-off point. While it plateaus strongly towards higher occlusion ratios, it noticeably does show several low-IoU samples at very low occlusion ratios, something that is not present in other occluders. The volumetric occluder exhibits very stable behavior at a low occlusion ratio, but distinct clustering patterns become present at higher levels. While the drop-off point remains clear, the high-IoU plateau leading up to it does produce a gradual downward curvature, unique to this occluder.

Varying the camera viewpoint within the occluder configurations does create a difference some variation. While the level of IoU stays mostly the same among different camera viewpoints, the point at which the drop-off occurs does change. For the solid and volumetric occluder, the decline begins the latest at an angle of 45°. In the case of the perforated occluder, the difference in drop-off point remains minimal, however the drop-off beginning the earliest at the 45° angle instead.

Out of all the configurations, the solid occluder exhibits the least anomalies, with the only notable behavior occurring in the B plot of Figure 3.2, where a small alternating oscillatory pattern is visible before the drop-off point. The perforated occluder exhibits a lot of data points lying outside the fit, in a non-grouped manner, but only before the drop-off point. The volumetric occluder does retain the data points grouped to an extent, but some groups deviate at a certain level from the fit. Notably, the outlying data behavior per occluder does persist between camera viewpoints. Despite this, the logistic fit does align appropriately with the samples.

4. Discussion

4.1. Interpretation of results

The results of the experiment raise several points that warrant further discussion, which will be examined here. These main topics include why the patterns and outliers present in the data occur in a broader sense, and why the different occluders and camera viewpoints produce different results.

The data shows the vision model is capable of retaining a clear detection stability, until key features become occluded, which causes the outlined threshold behavior to emerge. Due to the salient, non-uniform features that the vision model relies on, introducing occlusion does not produce a linear degradation in IoU. Compared to a linear fit, the logistic shape of the performance reflects the fundamental feature-based detection of the model, rather than pixel-area dependence. The difference in the drop-off point between occluders is caused by the fundamental characteristics and traits of the occluders, further emphasizing that they should not be treated as a single variable. This is due to how each of the occluders interact differently with model feature hierarchies. The solid occluder removes spatial chunks, the perforated occluder disrupts local feature continuity, and the volumetric occluder removes global contrast and texture. In convolutional feature spaces, these effects are fundamentally non-equivalent. As such, assuming occlusion behaves the same regardless of the type oversimplifies model perception to a high extent and can lead to a fundamentally misleading basis for reasoning.

In the case of the solid occluder, the oscillatory pattern observed likely originates from YOLO-specific augmentations, which was inferred to be indicative of random erasing specifically. The shape of this pattern arises from the periodic misalignment between the loss of important features and model expectation. This also affects the drop-off point, which is largely determined by whether the front or rear is obstructed first, because the occluder hides salient features in a directional and monotonic way. The threshold is clear in this case, as the occlusion is spatially coherent.

In the case of the perforated occluder, the early instability is indicative of local occlusion obscuring high-value visual cues, critical for the performance of the model. The magnitude and irregularity of this variance are partly caused by the shifting windows of visibility, creating iteration-dependent fluctuations. This arises from the non-monotonic nature of the occluder, and its use of high-contrast edges of grid cells which deteriorate feature extraction. Consequently, perceptual difficulty is underestimated by the global occlusion ratio.

The cluster of low IoU detections after the drop-off point when the volumetric occluder was employed was deduced to be a direct result of the confidence threshold being set too low, identified as false positives. They are not to be interpreted as meaningful detections, but rather just as a result of the detector's fallback tendency. The smooth pre-collapse decline reflects uniform attenuation reducing contrast across the entire spatial field and indicating intermediate visibility states. The drop-off point marks the threshold at which the signal-to-noise ratio becomes too low for the model to perform.

The camera viewpoints used affect the model in a broadly similar manner amongst different occluder configurations, but the extent of the degradation varies. For most cases, the 45° angle is the most robust, due to it also aligning with the majority of the training distribution. As 20° and 70° were underrepresented in the training process, the performance expectedly suffered, unrelated to occlusion physics. In the case of directional occlusion induced by the solid occluder, the distinct difference in the way it interacted with each viewpoint was the most apparent.

Beyond the overall threshold behavior, the occluder-specific patterns further highlight how different occlusion mechanisms interact with the detector. The solid occluder exhibits the most spatially coherent and monotonic degradation, producing a distinct directional sensitivity whereby feature loss depends on which regions of the object are obscured first. The perforated occluder, in contrast, destabilizes detections even at very low global occlusion ratios, as fragmented, high-contrast structure near the target interferes with early feature extraction before any substantive geometric overlap occurs. The volumetric occluder displays a gradual pre-collapse decline in IoU, consistent with uniform attenuation reducing contrast continuously across the spatial field, followed by post-threshold clusters of low-IoU detections attributable to near-confidence false positives rather than meaningful detections. This can be confirmed by inspecting the predictions, and Figure 4.1 illustrates a representative example of such a detection.

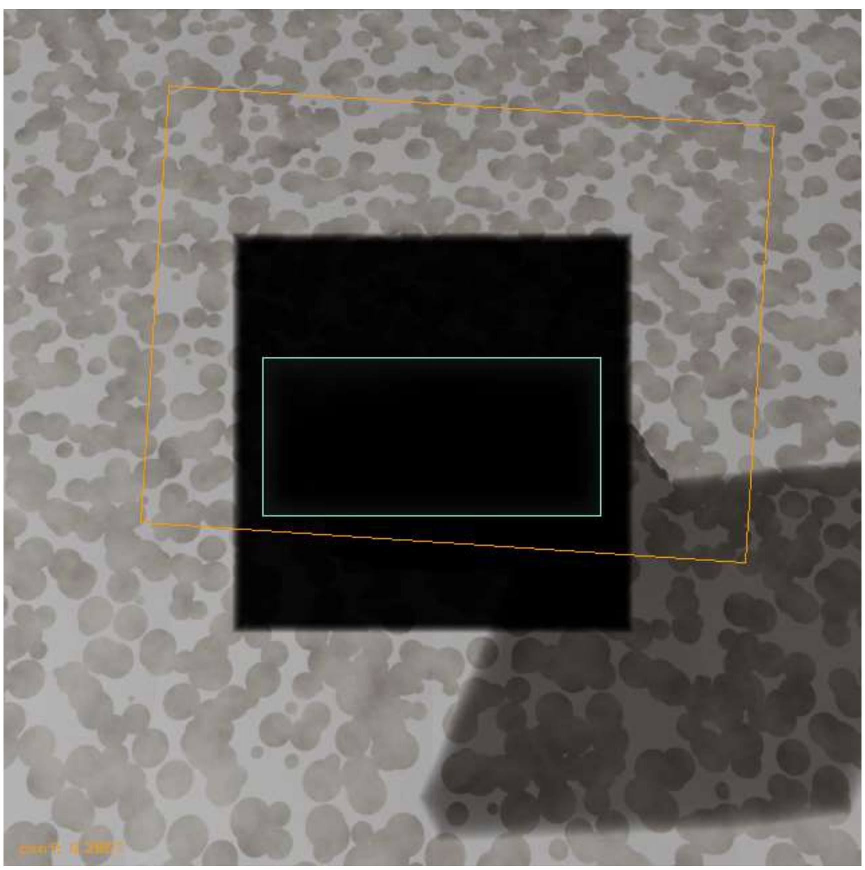

Figure 4.1. *Low IoU detection in the case of the volumetric occluder at a viewpoint of 20°.*

In Figure 4.1, the turquoise rectangle represents the ground-truth bounding box, and the orange rectangle represents the predicted bounding box. Also, the confidence (0.2857) is printed in the bottom left corner. As can be seen, the bounding box predicted by the vision model is not very precise or accurate and does not seem to be based on the actual projected geometry of the target object. As the occlusion ratio present in the figure is high at 0.933, a detection should not be expected in the first place. The vision model confidence threshold being at 0.25 implies that this detection is unlikely to represent a reliable true positive in the first place, and the clusters present in the plot could be avoided by increasing the confidence threshold above it. On the other hand, the vision model may interpret the volumetric occluder as partially resembling the target object. During training, the model learns to associate object detection with characteristic high-contrast edges and corners separating the target from its background. Similar edge-like features present within the volumetric occluder may therefore weakly activate detection responses, resulting in low-confidence detections under volumetric occlusion.

These findings further demonstrate that occlusion does not constitute a single uniform variable, but instead represents distinct interaction mechanisms between the visual scene and the learned feature hierarchies of the model. Directional geometric obstruction, fragmented high-contrast interference, and volumetric attenuation each deteriorate detection performance through different pathways, producing non-equivalent degradation profiles. Consequently, global occlusion ratio alone cannot serve as a reliable indicator for detection performance alone, when occlusion is spatially fragmented or perceptually heterogeneous. Treating occlusion ratio independently of occlusion type therefore oversimplifies model behavior and risks drawing misleading conclusions about robustness.

4.2. Model-specific insights and methodological implications

Next, the vision model itself is considered in greater detail in context of the findings. For model insight, this discussion will cover training distribution bias and architecture-driven sensitivity. To assess methodological implications, this discussion will cover measurement methods, analysis techniques, and the use of the experimental setup.

The most apparent aspect of the training distribution in this experiment was the biased representation of the 45° viewpoint, leading it to be the most robust. This gives insight into the model functionality during the training process, as this elevation was sampled the most, shaping the performance across configurations. Consequently, the model robustness is not limited to just occlusion physics but is tightly coupled with the distribution of viewpoints in the training material, showing how deep learning models tend to internalize biases from the training material. In the case of the solid occluder, the asymmetric drop-off between case A and case B implies that the model's learned salience map places greater weight on frontal features than rear features. This suggests that the vision model learns behavioral salience maps of the object, and whenever high-salience regions are exposed to occlusion, an earlier collapse is to be expected. These insights are useful in many applications and designs, for example detection industrial or automotive detection contexts. The instability caused by the perforated occluder shows that the vision model struggles especially with fragmented local signals, high-contrast artificial edges, and broken spatial continuity, which CNN-based detectors, in particular, can be sensitive to. This further emphasizes the importance of distinguishing between separate types of occlusions and their characteristics, as the vision model reacts in a distinct way to directional occlusion, spatial discontinuity, and attenuation.

The use of continuous measurement over discrete bins was found to be of great value, as it allowed for important metrics to be measured despite the intricate nature of occluders. The logistic-like collapse could not have been plotted with discrete bins, which provides the basis for large part of the analysis and discussion. Further, continuous measurement allowed for the observation of a curved decline, pre-collapse slopes, early instability, and oscillatory patterns, which discrete bins would have hidden. This emphasizes that to ensure methodological validity, studies measuring occlusion benefit from avoiding binning as stand-alone measurements for metrics. Of further methodological relevance is the recognition of occlusion type and occlusion ratio as distinct variables. It is critical to distinguish between them, as they are not interchangeable. Treating occlusion as a visual percentage blocked alone is insufficient for vision models based on CNNs, as this architecture reacts strongly to spatial coherence, texture destruction, contrast attenuation, and discontinuity. As such, occlusion type should be classified as a separate variable unrelated to the ratio of occlusion present and instead model them independently.

Besides providing insight related to vision model architecture in general, augmentation-induced artifacts specific to YOLO were observed as well. The oscillatory pattern induced by the Random Erasing augmentation reveals that model behavior cannot be assumed to reflect solely real-world logic but is also prone to at times unintended side effects of augmentations. As this augmentation introduces structured priors into the network, any resulting artifacts tend to appear as periodic patterns rather than isolated anomalies. This can be misinterpreted to be not tied to the vision model itself, which can lead to misleading deductions of the model performance. Consequently, vision model augmentations should be analyzed considering the observed occlusion patterns specific to the domain. In addition, the false positives in the case of the volumetric occluder provide further insight into the performance of YOLO specifically. The wide IoU spread among false positives indicates increased variance in the detector's classification and localization heads near the confidence threshold. Furthermore, it should be noted that a collapse in IoU will not always eliminate possible detection. However, the false positives did manifest themselves strongly at all viewpoint angles, implying that this phenomenon is not necessarily tied to camera viewpoint bias. On the other hand, the oscillatory pattern and difference in IoU drop-off points in the case of the solid occluder was strongly tied to the viewpoint. As such, it would not be unreasonable to deduce that not all artifacts induced by architecture or augmentations of a certain model manifest at the same levels of occlusion, or that they would be tied to the level of performance to the same extent.

The strong viewpoint dependence observed further emphasizes that robustness must be evaluated relative to training distributions rather than assumed to be viewpoint-invariant. Assessing vision model performance using a single viewpoint or occlusion scenario may therefore lead to substantial overestimation of true robustness. Additionally, reliance on aggregate metrics alone can conceal underlying failure modes, as detections may persist despite severe localization degradation. Continuous measurement proved critical in exposing pre-collapse slopes and logistic threshold behavior, which would remain obscured under discrete binning or class-based evaluation approaches.

While the use of a synthetic environment was found to create unique constraints, it on the other hand is capable of modeling behavior invisible in real-world data. Synthetic environment provides an environment for clean signals, interpretable patterns, and reproducible pathology. This supports the use of synthetic environments as valid tools for research and are conducive for diagnosing behavior tied to the vision model specifically. However, it is also important to address the limitations and the aspects that arise from them. These results further illustrate how training distribution biases shape viewpoint-dependent robustness. Further, the findings indicate that robustness is occluder-type specific, implying that exposing a model to multiple occluder mechanisms during training could influence generalization. In addition, it is important to recognize that the benchmarked robustness only applies to the type of occluder employed, and that this robustness does not transfer to another type. As such, testing the model for only a single type of occlusion can cause a dramatic overestimation of robustness. From this, it can be reasoned that scenario-specific variation may reveal vulnerabilities that random augmentation alone does not capture, improving vision model robustness. Consequently, assessing vision model performance based on a single viewpoint or occlusion scenario alone may overestimate robustness.

4.3. Limitations of the study

In part due to the highly controlled approach undertaken in our study, it consequently also restricts its scope. While this allows us to focus on examining our core interest within the scope of the study, it does limit it to just that in the context of the broader field.

Due to the synthetic environment employed, the experiment does not assess the impact of noise in this setup. Although real-world noise varies a lot, it could be implemented in the synthetic environment as well to an extent, but was deliberately left out, as it fell outside the scope of this research. Furthermore, all the geometry and materials used were geometrically idealized, lacking natural imperfections, and appeared just the way they were intended. In addition, the background was clean, uncluttered, and used a controlled material. This is, again, rarely the case in real-world data, which may reduce the ecological validity of the results. The camera model itself was also artifact-free, noise-free, and due to it being in a synthetic environment, it always produced deterministic renders lacking stochastic variation. In real-world data, conversely, factors such as occlusion dynamics, motion cues, and temporal stability are present and of importance.

While the occluders used allowed for a modular configuration, their perfect nature and constraints in the configurations also limit the generality of their findings regarding their impact on vision model performance. Starting with the solid occluder, it introduced occlusion only directionally, and furthermore, was limited to a single shape. Arguably, variations in shape, even in the context of directional occlusion, would conceivably create variations in the vision model performance. Conversely, the perforated occluder was rather unrealistic in terms of how it created occlusion. Natural, irregular perforations could be more expected and realistic to face, in contrast to an artificial grid-like pattern. While it allowed for great control, as was the intention, it does introduce non-monotonic occlusion patterns, which is not typically representative of real-world occluders. Finally, the volumetric occluder, while being uniform and possessing homogeneous density, is once again not representative of real-world conditions. While this allowed for precise control and a way to measure the isolated effects, true smoke swirls and fluctuates, and gradually disperses into the environment. Another notable limitation is the use of only three occluder categories. Furthermore, only a single occluder was introduced at a time per scene. While this, again, ensures that consistency is maintained, it would be unrealistic to assume that only a single type of occlusion would occur at a time, especially regarding real-world data. While this was largely dictated by the design constraints, it does limit the extent to which these findings can be applied, and therefore generalization to broader real-world occlusion scenarios should be made with caution. Instead, they are better assessed as providing the characteristic traits of the occluders used.

The target object used comprised only of a single, non-mutating mesh, with no variation in scale, shape, material, or other conditions. Notably, only a single target object was present per scene, with no use of multiple target objects or decoy objects. Again, this was deliberate to regulate and isolate the effects of the occluder specifically. What this means, however, is that the results are most applicable when describing robustness to occlusion under perfect single-instance familiarity, rather than real-world recognition of the effects of occlusion in general.

The use of vision models in the experiment was also restricted to a single model architecture, namely YOLOv11. While deliberate due to the designed framework of the study, this again introduces certain limits. Firstly, some observed effects, such as the oscillations, may be specific only to the used architecture or augmentations, rather than the effect of occluders in general. It is therefore possible that some findings reflect characteristics specific to YOLOv11 rather than general properties of vision models.

The training data itself was also originally created for a purpose not related to this thesis, and as such it introduces further limitations. Namely, its biased exposure to mid-elevation angles, rather than being exposed to the model at a 90° elevation angle, in a balanced manner. As the vision model was inherently less familiar with the 20° and 70° angles employed, compared to the 45° angle, it can meaningfully affect the performance of the vision model at these angles, separate of the effects of occlusion.

Although the occlusion ratio metric was handled rigorously and deliberately, they do cause some limitations as well. The geometric occlusion, for instance, assumes perfect silhouette segmentation, based on the object index masks. The volumetric occlusion, in turn, is approximated via derived transmittance, as a geometric approach was not conceivable. As they were derived from the data of the synthetic environment itself, rather than being visually extracted from the renders, real-world data cannot be expected to be perfect to the same extent. Not only does real-world occlusion rarely behave in a binary or purely exponential way, comparability among occluders themselves is not perfectly equivalent.

The sample size itself was also not exhaustive, as the three occluders and three viewpoints only allowed nine unique configurations to be employed. Each setup consisted only of a single occluder, a single target object, and a single light source, with very limited variations per configuration. This lack of environmental diversity is another limitation, especially given that the examined studies recognized environmental variation as a factor that can affect performance.

4.4. Contributions of the study

The findings of this study contribute to existing research by introducing novel methodological, empirical, and theoretical insights. This includes more fundamental findings, as well as those only acquirable through the quantifiable way this study approached the research. These contributions include both foundational insights that support a more fundamental understanding of occlusion behavior, as well as more fine-grained findings made possible only through the fully quantifiable and controlled experimental design of the study.

The main novelty of the contributions of this study arises from its methodological contributions. This study contributes a systematic dual framework for quantifying geometric and volumetric occlusion ratios. For geometric occlusion, this is achieved through object index masks, and for volumetric occlusion, through computing transmittance. Further, continuous measurement was applied instead of discrete bins, enabling the occlusion to be measured quantifiably. This enabled observations of intricate patterns and anomalies regarding model performance under occlusion that would otherwise remain obscured. The use of a fully controlled synthetic environment with fixed parameters enabled clear isolation of occlusion effects. Due to the modular and parametric framework built, each occluder type with varying rates and viewpoint angles could be tested in the same environment in a reproducible and scalable manner, while allowing for each to be isolated at a time. Most examined studies could not isolate the variables in the same manner, leading to occlusion-related variables commonly being treated in a simplified or undifferentiated manner. Conversely, this study formalizes and operationalizes them as distinct variables within a controlled framework.

Empirically, this study demonstrated that performance degradation induced by occlusion follows a logistic pattern, with non-linear drop-off threshold behavior, supporting and quantifying existing theoretical foundations. The study also showed evidence that in the environment used, each occluder type produces a distinct degradation profile, shown through the comparison of solid, perforated, and volumetric occluders. Further, the effects of different viewpoint angles were also quantified, showing that in isolation, viewpoint influences performance separately from occlusion. The study also identified previously unreported patterns, such as those caused by augmentations or occlusion asymmetry.

Conceptually, this study shows that occlusion is not a single variable, and that the type and underlying mechanism are just as important as the rate itself. Further, feature-salience was also shown to be a factor vision model robustness is dependent on, providing evidence that degradation in performance is linked to learned salience maps. Training bias was also shown to be a strong determinant of robustness, as an uneven distribution in viewpoint angles during training strongly shaped collapse points, aligning with and enriching theories on data-induced bias in deep learning.

Practical applications that can be drawn from the study are broad and widely applicable. The insights provide guidance on dataset design, emphasizing the importance of including multiple occluder mechanisms and ensuring viewpoint diversity during training to avoid systematic robustness biases. The use of controlled synthetic data to diagnose model weaknesses was also shown to be a valid method, particularly for uncovering failure modes that may remain masked in natural datasets. The threshold-based collapse points provide clearer insight into model performance characteristics than intrinsic model parameters alone, and underscore the importance of appropriate detector confidence tuning in deployment scenarios, especially under volumetric attenuation where false-positive clustering may distort performance assessment. The evaluation further showed how different occluders require distinct robustness testing procedures, supporting the differentiation of occlusion mechanisms rather than treating occlusion as a single uniform variable.

4.5. Comparison with existing research

Although existing research relevant to our experiment has been discussed and outlined in Section 1.4, it has not been directly contrasted to our work. Accordingly, this section will synthesize the results from our experiment in the context of established findings. The broad alignment is assessed first, after which occluder types, camera viewpoint effects, and contradictions and novelties will be compared and contrasted, before summarizing the findings and how this work adds to the existing research.

Starting with the general patterns, our study confirms the notion that occlusion degrades vision model performance in a consistent, predictable manner. Existing work outlined that performance collapses at high levels of occlusion, and our study quantified this by pinpointing the drop-off ranges for each occluder and camera viewpoint. Consistent with referenced work, our findings show that performance remains robust at low levels of occlusion, with the exception of the perforated occluder. The degradation was also stated to be systematic instead of random, which is supported in our work by not only identifying correlation, but also demonstrating its logistic-like relationship across different configurations. Kassaw et al. [15] similarly reported that degradation behavior is preserved in synthetic datasets, a finding supported by our experimental results.

Regarding different types of occlusions, the referenced studies largely treated occlusion as type-agnostic. While different severities were recognized and treated separately in some cases, there was no differentiation between occluder types themselves, such as solid, perforated, transparent, or directional. Most cases employ fixed or simple geometric occluders, such as a car blocking a person, or in synthetic cases, a pasted image. Notably, occlusion geometry or material was not varied systematically in examined studies, which our study, by contrast, does. As such, our study brings forth some novelties around occlusion types specifically, revealing distinct behavioral signatures between them.

The solid occluder was found to be the most stable and monotonic, cleanly fitting the logistic curve, alluding to a strong vision model robustness relative to the geometry of the occluder due to delayed IoU drop-off. Further, due to recognizing its directional effect, the point of collapse also depends on which side is occluded first. The perforated occluder was found to have the highest variance at early stages of occlusion, while also having the smallest variety in drop-off between viewpoints. Early instability was found to arise from partial, non-monotonic occlusion, characteristic of the perforated occluder used. The volumetric occluder displayed a smooth decline in IoU with fine-grained attenuation, and also exhibited a unique, early downward slope before the drop-off. The selected studies have not examined volumetric occlusion in the context of continuous attenuation, which was modeled in detail in our study. Besides the novelties tied to certain occluder types, the main finding that our study brings forth is that the occluder type fundamentally shapes model degradation patterns. As such, treating occlusion as a single variable can obscure important distinctions, limiting methodological progress. Furthermore, most of the existing research had a greater emphasis on classification and categorization of the target object, and vision model recall and precision. Our study, in contrast, focused on IoU degradation specifically. In essence, instead of focusing on only if the vision model is capable of detecting the object, we focus on how capable it is.

Camera viewpoint is a relatively unique aspect inherent to the use of a synthetic environment, as the predominantly real-world datasets employed in the examined studies typically contain uncontrollable or fixed viewpoints [14, 15]. Our study treated it as an independent variable instead, which allowed its effects to be modeled systematically. Although existing research has acknowledged that shifting the camera viewpoint can adversely affect detection performance [17], this effect has not been quantified or evaluated in a controlled manner in the examined studies. Our study found that generally, the vision model is the most robust at the 45° camera angle. This, however, can be largely attributed to how the vision model was trained with an emphasis on mid-elevation points, which is outlined in detail in Section 2.2. This would also make sense considering the nature of machine learning, as it would not be inherently aware of an object's rotation and translation, other than through how it was trained. Further, due to the way our study modeled occlusion, it confirmed that it is not indeed always symmetrical, as was most clearly shown with the solid occluder. In general, our study demonstrated how a shift in viewpoint interacts with different occlusions and modeled it. Solid showed a strong effect, perforated a weak effect, and volumetric a moderate effect on viewpoint sensitivity.

Whereas the referenced studies identified correlation and modeled it to a certain extent, our study provided a logistic fit and showed that the relationship is not linear, nor strictly exponential. Occlusion was also treated as uniform, at times separated by ratio, whereas our study differentiated between them in a non-arbitrary manner, and provided characteristics for them. Solid was largely monotonic, perforated showed high variance, whereas volumetric was dense with a smooth continuum. Furthermore, our study revealed that each occluder type produced a distinct logistic curve shape, further emphasizing the importance of distinguishing between different types of occluders. Where examined studies relied on discrete bins or ratio labels, our study measured occlusion in two separate ways, depending on the occluder. Guo et al. [5] did provide a way to measure occlusion geometrically, which our study applied, but also expanded on by providing a way to measure occlusion through volumetric transmittance. Another novelty is the observation of directional occlusion, shown in the case of the solid occluder. It further suggests the presence of previously unexamined feature-salience patterns in the vision model. The augmentation-induced model artefacts witnessed through the oscillatory patterns in the case of solid occlusion is another novelty, likely tied to YOLO's innate augmentations, which has not been previously identified. Finally, examined studies largely relied on heterogenous categories and scenes, whereas our study used deliberately fixed variables in an independent parametric space, using the exact same target object, lighting, distance, and material, while controlling camera viewpoints and parameters. The controlled environment allowed our study to isolate the effects of occlusion, providing a much cleaner signal for analyzing its effects.

4.6. Future research directions

While the study explored a broad range of factors related to occluder effects on vision model performance, certain simplifications and constraints were required to keep the research within scope. These aspects were acknowledged in Section 4.3. Accordingly, this section reflects on how future research could build on both the limited areas and the contributions introduced by the study.

Although occluder diversity was achieved, three occluders are far from exhaustive. While they provide a strong base to build upon, introducing more occluders and camera angles would benefit the research greatly. Increasing the number of occluders is the first step, but further emphasis on realism could also be considered. Irregular and dynamic occluders, especially those varying temporally, would represent real-world scenarios more accurately. In addition, using multiple occluders per scene would allow for layered occlusion and more complex visibility conditions.

Using three camera viewpoint angles worked for the scope of this thesis, but this too could be measured continuously. Further camera transformations, such as panning, tilting, changing the field of view, adjusting resolution, or shifting object location within the frame, could also be introduced. Recognized camera motion constraints, such as those arising from drones, could also be explored, especially when considering real-world applications. Naturally, adding sensor noise to avoid idealizing the environment, as well as adjusting lighting, background materials, and shadows, would help bridge the gap between synthetic and real scenes.

The target object used is another limiting factor, as its material and geometric projection remained strictly controlled throughout the entire experiment. Future research could manipulate these characteristics directly. Using more than a single target object is also important, as it would allow researchers to explore the fundamentals of feature salience more deeply. Further research could model salience maps directly, assess occlusion effects per feature region, and determine which parts of an object carry the most predictive power.

Adding whole new object classes would allow both cross-class and intra-class effects to be examined.

The training process itself is a natural target for refinement in future studies, especially around avoiding viewpoint biases by ensuring rich and varied training datasets. Augmentations were already found to influence performance, so varying them systematically could help model their effects and aid in identifying augmentation-specific artifacts. Introducing occluders directly into the training process and assessing how this affects model robustness could also provide meaningful insight.

Using only a single vision model architecture is another major limitation, as generalization of the results is constrained and architecture-specific artifacts may have influenced the findings. Future studies could compare multiple architectures of different types to assess cross-model robustness and isolate effects that are model-specific rather than inherent to occlusion.

The separation of occlusion metrics into geometric and volumetric cases was a novel insight, but further categories could also be introduced. These could eventually be unified into a general occlusion ontology. The fundamental metrics could be validated using perceptual visibility models, and the measurement techniques themselves refined further. Building a universal measurement standard would benefit the field greatly and streamline future research.

Moving beyond drop-off thresholds derived from IoU and occlusion ratio also represents a promising direction. Some patterns may be more visible under alternative metrics, which could reveal additional aspects of model performance. The logistic collapse observed in this study suggests that occlusion may resemble a type of phase transition in the model, naturally inviting deeper analysis of collapse dynamics. Further computational modeling, such as exponential, critical-threshold, or hybrid models, could reveal whether the slope of the collapse correlates with occluder characteristics.

As real-world systems rarely rely on single-frame inference, introducing temporal analysis during the experimental process would help assess robustness in dynamic scenes. Tracking, motion prediction, and multi-frame aggregation could reveal new insights, especially for directional occluders whose effects evolve over time.

A natural next step beyond temporal modelling is moving from synthetic to real environments. Generating paired occlusion sequences in both domains would allow for direct comparison of collapse patterns and failure modes. Such cross-domain experiments would help determine which occlusion effects are fundamental to vision models, and which arise specifically from synthetic conditions.

5. Conclusion

This study investigated the effects of occlusion ratio, occlusion type, and camera viewpoint on the detection performance of a YOLO-based vision model trained in a synthetic environment. The work addressed the primary research question, "What is the effect of occlusion ratio and type on the detection performance of computer vision models trained on synthetic visual environments?", and the secondary research question, "What is the effect of camera viewpoint variation on object detection performance?", through quantitative analysis of occlusion-related performance metrics. The hypothesis was that these relationships would be non-linear, with sharp performance degradation occurring beyond a threshold, and that different occluder types would produce measurably distinct effects.

The observed degradation patterns across occluder configurations strongly supported the proposed hypothesis. Detection performance remained relatively stable at low levels of occlusion before exhibiting a sharp logistic-like collapse at a threshold point. This confirms that the relationship between occlusion ratio and performance is distinctly non-linear rather than gradual.

Per occluder, the solid occluder was found to exhibit the most stable behavior, with samples remaining tightly grouped. It also exhibited the latest drop-off threshold across the occluders, while also highlighting the unique aspect of directional occlusion, which has a different effect on performance depending on which spatial area is obstructed first. The drop-off threshold ranged from around 0.33 to 0.58, showing the highest variance. The perforated occluder was shown to exhibit the highest variance and was the only configuration to exhibit a degradation in performance at low levels of global occlusion. This phenomenon did not show strong correlation with the viewpoint angle, however. The drop-off threshold ranged from around 0.38 to 0.45, with notable early instability. The volumetric occluder displayed the smoothest, most gradual decline of performance before the drop-off threshold, which nonetheless was present and clearly observable. The drop-off threshold ranged from around 0.38 to 0.44, the tightest among the configurations. These patterns align with the second part of the hypothesis, and help answer the primary research question, by showing that model performance is measurably affected depending on the occluder type.

Camera viewpoint had a systematic effect on model robustness, with the highest performance consistently observed at the 45° elevation across occluder types. The 20° and 70° viewpoints exhibited earlier performance collapse, particularly in the solid occluder condition. These findings demonstrate that viewpoint variation contributes independently to detection robustness, consistent with the training bias toward mid-elevation perspectives outlined in Section 2.2.

Notable anomalies and secondary patterns were also observed across occluder types. The solid occluder exhibited oscillatory behavior at low IoU levels, most pronounced at the 45° viewpoint. The perforated occluder showed strong scatter in data points at low IoU values as well, largely independent of viewpoint. This instability was evident even at a global occlusion ratio of 0, likely due to the visible grid outline present in the scene. The volumetric occluder exhibited a gradual performance decline approaching the drop-off threshold, with post-threshold clustering of low-IoU false-positive detections near the confidence threshold.

Overall, occlusion ratio was shown to strongly affect detection performance, with degradation patterns varying systematically by occluder type. Camera viewpoint further influenced robustness across all configurations. Taken together, these results demonstrate that occlusion should not be treated as a single-variable phenomenon but must instead be categorized and modeled according to both ratio and underlying occlusion mechanism.

REFERENCES

[1] J. Liu and Y. Jin, “A comprehensive survey of robust deep learning in computer vision,” Journal of Automation and Intelligence, vol. 2, pp. 175–197, 2023. doi: 10.1016/j.jai.2023.10.002.

[2] J. Wang, Y. Lin, J. Yang, Y. Wang, and Z. Wu, “Deep Learning for Object Detection: A Survey,” Computer Systems Science & Engineering, vol. 36, no. 2, pp. 225–240, 2021, doi: 10.32604/csse.2021.017016.

[3] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[4] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016, ch. 9 (“Convolutional Networks”).

[5] P. Guo, K. Zheng, T. Liu, Y. Jiang, and H. Zhu, “Deep object occlusion relationship detection based on associative embedding clustering,” Technologies, vol. 13, no. 4, p. 143, 2025, doi: 10.3390/technologies13040143.

[6] A. Saleh, M. Al-Maadeed, A. A. Zaidan, and B. B. Zaidan, “Occlusion Handling in Generic Object Detection: A Review,” Artificial Intelligence Review, vol. 54, no. 1, pp. 341–393, 2021, doi: 10.1007/s10462-020-09858-8.

[7] A. Gilroy, S. Küppers, and A. Leonardis, “The Impact of Partial Occlusion on Pedestrian Detectability,” arXiv preprint, arXiv:2205.04812, 2022, doi: 10.48550/arXiv.2205.04812.

[8] H. Chen, X. Li, and Y. Zhou, “Investigating the Robustness and Properties of Detection Models,” arXiv preprint arXiv:2310.08772, 2023, doi: 10.48550/arXiv.2310.08772

[9] M. A. Ariffin, M. M. Tahir, and S. A. Ahmad, “A Survey on Dependability and Safety in Industrial Robotic Systems,” IEEE Access, vol. 10, pp. 84213–84232, 2022, doi: 10.1109/ACCESS.2022.3189811.

[10] Y. Ma, X. Zhang, and W. Ouyang, “A Survey on 3D Object Detection Methods for Autonomous Driving,” IEEE Transactions on Intelligent Vehicles, vol. 7, no. 3, pp. 387–404, 2022, doi: 10.1109/TIV.2021.3094382.

[11] A. D. Nguyen, T. V. Pham, and M. Q. Tran, “Computer Vision-Based Target Recognition and Vision Enhancement in Battlefield Environments: A Review,” Defense Technology, vol. 19, no. 1, pp. 227–245, 2023, doi: 10.1016/j.dt.2022.06.009.

[12] Z. Wang, H. Zhao, and R. Rajaraman, "Reliability Challenges in Safety-Critical Computer Vision Systems: A Comprehensive Survey," IEEE Transactions on Intelligent Systems, vol. 39, no. 4, pp. 564–578, 2023, doi: 10.1109/TIS.2023.3249012.

[13] B. Pepik, M. Stark, P. Gehler, and B. Schiele, "Occlusion Patterns for Object Class Detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3286–3293, 2013, doi: 10.1109/CVPR.2013.422.

[14] M. Zhang, H. Hu, Z. Li, and J. Chen, "Occlusion-robust workflow recognition with context-aware compositional ConvNet," Soft Computing, vol. 28, no. 6, pp. 5125–5135, 2024, doi: 10.1007/s00500-023-09225-2.

[15] K. Kassaw, F. Luzi, L. M. Collins, and J. M. Malof, "Are Deep Learning Models Robust to Partial Object Occlusion in Visual Recognition Tasks?," arXiv preprint, arXiv:2409.10775, 2024, doi: 10.48550/arXiv.2409.10775

[16] J. W. Smutny, "The effects of using created synthetic images in computer vision training," arXiv preprint, arXiv:2506.03449, 2025, doi: 10.48550/arXiv.2506.03449

[17] Y. Dong, S. Ruan, H. Su, C. Kang, X. Wei, and J. Zhu, "ViewFool: Evaluating the robustness of visual recognition to adversarial viewpoints," arXiv preprint arXiv:2210.03895, 2022.

[18] H. R. Sánchez, "Seven Derivations of the Lambert–Beer Law," arXiv preprint arXiv:2009.04847, 2020.

[19] Ultralytics, "YOLO Data Augmentation," Ultralytics Documentation. Accessed: Feb. 12, 2025. [Online]. Available: https://docs.ultralytics.com/guides/yolo-data-augmentation (accessed November 19, 2025)