IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 October 2023, accepted 23 December 2023, date of publication 1 January 2024,
date of current version 10 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3348663

RESEARCH ARTICLE

Analysis of the Impact of Lens Blur on Safety-Critical Automotive Object Detection

DARA MOLLOY[1,2], PATRICK MÜLLER[3], BRIAN DEEGAN[1,2], DARRAGH MULLINS[1,2], JONATHAN HORGAN[4], ENDA WARD[4], EDWARD JONES[1,2], (Senior Member, IEEE), ALEXANDER BRAUN[3], AND MARTIN GLAVIN[1,2], (Member, IEEE)

[1]School of Engineering, University of Galway, Galway, H91 TK33 Ireland
[2]Ryan Institute, University of Galway, Galway, H91 TK33 Ireland
[3]Department of Electrical Engineering, Hochschule Düsseldorf, 40476 Düsseldorf, Germany
[4]Valeo, Tuam, Galway, H54 Y276 Ireland

Corresponding author: Dara Molloy (d.molloy13@universityofgalway.ie)

This work was supported in part by the Science Foundation Ireland under Grant 13/RC/2094, and in part by the European Regional Development Fund through the Southern and Eastern Regional Operational Programme to Lero—the Science Foundation Ireland Research Centre for Software (www.lero.ie).

ABSTRACT Camera-based object detection is widely used in safety-critical applications such as advanced driver assistance systems (ADAS) and autonomous vehicle research. Road infrastructure has been designed for human vision, so computer vision, with RGB cameras, is a vital source of semantic information from the environment. Sensors, such as LIDAR and RADAR, are also often utilized for these applications; however, cameras provide a higher spatial resolution and color information. The spatial frequency response (SFR), or sharpness of a camera, utilized in object detection systems must be sufficient to allow a detection algorithm to localize objects in the environment over its lifetime reliably. This study explores the relationship between object detection performance and SFR. Six state-of-the-art object detection models are evaluated with varying levels of lens defocus. A novel raw image dataset is created and utilized, containing pedestrians and cars over a range of distances up to 100-m from the sensor. Object detection performance for each defocused dataset is analyzed over a range of distances to determine the minimum SFR necessary in each case. Results show that the relationship between object detection performance and lens blur is much more complex than previous studies have found due to lens field curvature, chromatic aberration, and astigmatisms. We have found that smaller objects are disproportionately impacted by lens blur, and different object detection models have differing levels of robustness to lens blur.

INDEX TERMS Object detection, lens blur, ADAS, autonomous vehicles, intelligent transportation system.

I. INTRODUCTION

Cameras that capture the visible spectrum of electromagnetic radiation are most similar to human vision. The road environment has been created and tailored for human vision with visible features such as signs, traffic lights, and lane markings; thus, object detection aims to enable computers to perceive objects in the road environment similarly to humans. In recent years, significant advances in deep learning architectures, datasets, and hardware accelerators have solidified deep learning models as the predominant architecture in object detection [1], [2], [3]. Object detection models, utilizing deep learning, learn object representations from large annotated datasets and achieve state-of-the-art performance on object detection benchmarks [1]. However, without explicit testing, it is difficult to determine how much the characteristics of an object's appearance can change while still being detectable by the model due to the black-box nature of deep-learning object representations. The appearance of objects can vary significantly due to occlusions, lighting, precipitation, and lens soiling. However, even within a camera system, the appearance of objects can vary due to defocus caused by temperature, sensor/lens misalignment, sensor noise, compression artifacts, and different image

The associate editor coordinating the review of this manuscript and approving it for publication was Sukhdev Roy.

signal processing (ISP) parameters. A safety-critical camera-based object detection system must detect objects reliably even when impacted by these factors.

Camera-based object detection is also currently being considered for smart infrastructure supporting safety-critical applications [4]. Intelligent transportation system (ITS) infrastructure nodes are sensor units along the roadside that communicate with connected and autonomous vehicles (CAVs) to improve transportation through various applications, as discussed by Clancy et al. [5]. Cooperative collision avoidance [6] is one such application in which an infrastructure node perceives the environment and provides third-party information to autonomous vehicles to increase their safety and reliability. Advanced driver assistance systems (ADAS), autonomous vehicles, and Intelligent Transportation System (ITS) infrastructure nodes are expected to increase road safety as the technologies advance [7]. Contemporary ADAS features [8] such as lane keeping, emergency braking, and blind spot detection use deep learning models for environmental perception [9]. While the cameras surrounding the vehicle can provide a much greater field of view and spatial resolution than a single human driver can achieve, the robustness of deep learning models to degraded images is much lower than in humans [10].

To create a camera-based object detection system that is robust enough to support safety-critical applications for the entire lifetime of a vehicle, which may be on the order of decades, all potential sources of degradation must be characterized. Sharpness is a key performance metric when manufacturing cameras because a sharper image yields higher subjective image quality. The sharpness of a camera system is known as its spatial frequency response (SFR), and it defines how well the system can resolve details in the environment. The SFR of a camera system has many determining factors such as cost, manufacturing variability, and, for fixed focus lenses, sometimes even drift of the sensor/lens alignment throughout the camera's lifespan due to temperature cycling, vibrations, or impacts. When mass manufacturing cameras, sharpness is measured at end-of-line (EOL) to ensure the cameras meet a minimum sharpness threshold. Camera sharpness is typically measured according to ISO12233:2017 using an edge SFR chart as seen in Fig. 4. The modulation transfer function (MTF) defines the relative contrast a camera obtains at a given spatial frequency and is the type of SFR that is used in camera manufacturing. Processing the edge SFR chart according to ISO12233:2017 provides MTF values for a range of spatial frequencies and in camera manufacturing these values must exceed minimum thresholds. Setting an appropriate MTF threshold is vital for manufacturing because setting the threshold far too low would create camera systems that yield severely blurry images that do not capture sufficient environmental information to carry out safety-critical object detection, such as ADAS. However, defining the minimum acceptable MTF threshold to be, for example, the theoretical maximum MTF of the sensor and lens would only allow a small percentage of cameras to meet this threshold, leading to uneconomic manufacturing. Currently, the minimum sharpness, or MTF, threshold is challenging to determine as the relationship between MTF and environmental perception has not previously been characterized, leading manufacturers to employ a high MTF value to ensure a margin of safety. The same MTF threshold is generally utilized in product validation testing, where the camera systems undergo accelerated life testing to ensure the system's SFR does not degrade to an unacceptable degree over the product's lifespan.

In this study, the impact of several lower SFR camera systems on object detection performance is evaluated and correlated to MTF values. Characterizing the relationship between object detection performance and MTF provides insight into creating acceptable MTF ranges on manufacturing lines and in product validation testing. The widely available and commonly used large object detection datasets [11], [12] could not be utilized as they include unknown variables that impact SFR, such as subjectively tuned ISP, compression, and multiple camera/lens combinations. MTF can be estimated from natural images [13]; however, it's not representative of measuring the MTF at end-of-line and in product validation testing.

To obtain accurate MTF measurements on cameras in an object detection dataset, a new dataset was collected and annotated from the perspective of an ITS infrastructure node. The cameras used to collect this dataset had their MTF measured according to the ISO12233:2017 standard in a controlled image quality lab. The dataset consists of cars and pedestrians traveling up to 100m away, with different cars and pedestrians being captured to maximize dataset variability. The camera's raw Bayer data was captured to ensure no unknown image processing was introduced, and a minimal software ISP [14] was applied. The ISP did not apply edge enhancement algorithms as they alter camera MTF. The averaging effect of lens blur disproportionately affects smaller objects, as it is analogous to reducing the number of pixels on a camera with a perfectly sharp lens, so it was vital to include a range of object sizes in the captured dataset.

A physically realistic lens model was implemented, using the point spread functions (PSFs) of a Cooke triplet lens model, to create accurate lens blur as described and validated in [15], [16], [17], and [18]. The benefits of a physically realistic lens model over previous blur models have also been discussed in [16]. PSFs describe the response of a lens to a point light source. They are generally not rotationally symmetric and vary across the image field, with the edges generally less sharp than the middle. Lens model PSFs for three wavelengths, red, green, and blue, were obtained from Zemax's ''OpticStudio'' [19]. The lens model was then defocused by adjusting the parameters of the Zernike polynomial to obtain blurred PSFs that represent both positive and negative defocus. The final degraded dataset contains three steps of defocus in the positive and negative directions, yielding seven dataset versions on which the object detection

models can be evaluated. The defocus is not symmetric around the nominal position in the positive and negative directions, with both directions yielding images with unique optical properties.

Many unique deep learning object detection architectures achieve state-of-the-art performance on object detection benchmarks. The main differences between the architectures utilized in this study are the number of parameters, single-stage versus two-stage networks, use of anchor boxes, differing loss functions, and a comparison between convolutional neural network and transformer neural network architectures. The six deep learning models chosen were FCOS, with a ResNet50 FPN backbone, YOLOv8m, Faster RCNN, with the ResNet50 FPN backbone, Cascade RCNN, with the ResNet50 FPN backbone, Faster RCNN, with the transformer-based Swin backbone, and Deformable DETR. Model performance will be evaluated over each object class for small and large objects to investigate how lens blurring impacts model performance on each variable. Evaluating the model's performance on the defocused datasets facilitates drawing the correlation between MTF and deep learning model performance.

This paper is organized as follows: related works are discussed in section II. Section III contains the methodology for analyzing the performance of six object detection algorithms on seven degraded datasets and measuring the MTF of the various degradations. The deep learning model results and the correlation between MTF and deep learning models are outlined in section IV, with section V drawing the conclusions.

II. RELATED WORKS

Deep learning object detection models are currently deployed in safety-critical systems because they greatly outperform traditional methods based on handcrafted feature extraction [20], [21]. The need for these systems to be reliable in every situation has motivated additional research into the robustness of deep learning models to outlier cases and degradations. Popular object detection benchmarks such as COCO [11] provide metrics to analyze outlier cases by segmenting performance based on object area. The AP_{small} metric, created for the COCO benchmark, specifically characterizes the performance of objects less than 32×32 pixels. Other object detection dataset benchmarks contain occlusion annotations [22], [23], [24], [25], [26], enabling evaluation of occluded object performance. The impact of other external factors such as precipitation [27], [28], [29], [30], lighting [27], [28], [29], and lens soiling [27] on deep learning model performance have also been investigated. There have also been investigations into degradations that are applied from within the camera system, such as varying image exposure [27], [28], [29], compression artifacts [28], [29], [31], [32], sensor noise [22], [28], [29], [32], [33] and varying ISP parameters [22], [29], [32], [34].

In 2016, Dodge and Karam [32] evaluated the effect of Gaussian blurring, alongside other degradations, on deep-learning classification performance and found that image blurring significantly impacts classification performance. Other studies have analyzed the impact of different types of blur on deep learning models [22], [27], [28], [29], [33], [35]. Mitigation strategies against blur and other degradations such as pre-processing with another deep learning model and adding degraded data to the training set [29], [35] have also been investigated. Recent research into evaluating object detection robustness has led to the creation of datasets containing images that have undergone multiple degradation types, including blurring [27], [28], [29], [36]. Training deep learning models with blurred images has been shown to increase robustness against blurring [35], [36]; however, there is a limit to the amount of additional robustness that can be gained as blurring an image reduces the amount of information contained within the image. In 2019, Geirhos et al. [36] showed that deep learning classification currently relies heavily on object texture, whereas humans also avail of object shape to inform object classification decisions. Object textures consist of high spatial frequencies that are removed when blurring occurs, explaining the considerable impact of blurring on deep learning model performance.

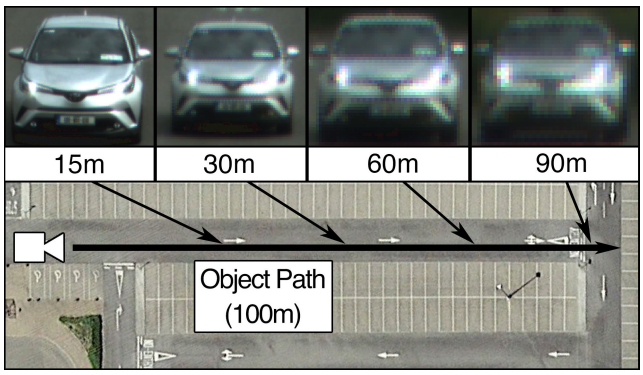


FIGURE 1. Aerial view of the data acquisition setup showing the controlled 100m path that the objects traversed and some example car data at four points along the path.

It is clear that blurring significantly reduces deep learning model performance; however, the correlation between optical quality and deep learning model performance is still unknown. The studies above quantify blur amount based on the kernel size of a uniform disk-like blur kernel; however, this can not be directly related to optical quality. The blur kernels used in the studies above are also not physically realistic as they yield uniform defocus around the image field and are color channel independent.

Previous studies simulated blur on widespread object detection or classification datasets, so the impact of post-processing steps, such as edge enhancement and image compression, have not been isolated. The blur level has also not been correlated to an objective SFR metric to allow researchers and manufacturers to recreate these results with real defocused cameras. In 2022, Müller and Braun [15]

utilized their physically realistic lens model, the same methodology as in this study, on the Berkeley Deep Drive [12] dataset to validate their approach for superposition approximation by comparing SFR and object detection performance with a single detection model for the pedestrian class with no offset on the defocus coefficient. Although no defocus was applied, adding the Cooke triplet lens model reduced the object detection AP by 6% due to the reduction in image sharpness. Another area in need of investigation is the impact that blurring has on object detection performance with smaller or further away objects because both smaller objects and object textures consist of higher spatial frequencies, which are disproportionately impacted by lens blurring.

III. METHODOLOGY

A. DATASET

Previous studies investigating the impact of blur on object detection performance have used publicly available datasets. Publicly available datasets provide access to a significant amount of data; however, there are many unknowns surrounding their capture, such as details of the image signal processing (ISP) algorithms and compression that have been applied, which may have impacted these studies' results. A typical ISP pipeline consists of algorithms to increase subjective image quality, such as increasing sharpness and removing noise. These algorithms often modify the spatial frequencies within the image [37]. Lossy compression algorithms, such as JPEG, are generally applied to large image datasets to reduce dataset storage requirements. Lossy compression exploits the human visual system by removing higher spatial frequencies from images because humans tend to perceive object shapes more so than object texture [36], so the difference in image quality can be imperceivable while achieving a significant reduction in file size. Image blurring disproportionately affects these higher spatial frequencies, so they must be present in the dataset used to investigate the impact of blur on object detection models.

In this study, a controlled dataset of uncompressed raw images was collected using a FLIR BlackFly-S 8.9MP camera. The images were acquired at a frame rate of 30Hz in an 8-bit raw Bayer format. The dataset was taken from an ITS infrastructure node perspective, with the cameras mounted at a height of 4 meters above the ground and pointing at an angle of 20° towards the ground. The dataset consists of pedestrians and cars traveling 100 meters away from the camera on the path shown in Fig. 1. The dataset was captured to identify the changes in object detection performance due to lens defocus by minimizing the impact of external variables such as lighting, weather, and object occlusions. The test setup in Fig. 1 shows the sensor position and object path in a car park. The dataset consists of 450 images with 660 object instances, located throughout the object path, that were labeled in CVAT [38]. The variables within the dataset are object distance and class. The dataset was split by class and object size to evaluate the variables individually. The object size thresholds were determined by finding the object sizes corresponding to near objects (between 15m and 30m from the camera) and far objects (between 50m and 100m from the camera). The resulting thresholds are 500px^2 to $1{,}500\text{px}^2$ for the small objects within the person dataset, $3{,}500\text{px}^2$ to $20{,}000\text{px}^2$ for the large person objects, $1{,}250\text{px}^2$ to $3{,}500\text{px}^2$ for the small car objects and $10{,}000\text{px}^2$ to $40{,}000\text{px}^2$ for the large car objects. The individual datasets are evaluated to analyze how the models perform on different classes and object sizes when impacted by lens blur.

B. BLUR

Previous works [22], [27], [28], [29], [32], [33], [35] have utilized a uniform disk-like blur kernel that is spatially uniform and wavelength independent. However, a real lens generally has a non-rotationally symmetric point spread function (PSF) that varies across the field of view, resulting in different levels of blur in the outer and middle image areas, as seen in Fig. 2. The PSF is also wavelength-dependent, with red, green, and blue light refracting differently and producing varying levels of blur depending on their location in the image [39]. Production tolerances such as focal length variations and sensor/lens misalignment can also affect image quality.

In this study, the Zernike Fringe coefficients of a Cooke triplet model for three wavelengths were utilized from Zemax's software ''OpticStudio''. The Cooke triplet was optimized for a wide field of view, resulting in strong aberrations due to the limited number of lens elements. The wavefront errors from our Zernike polynomials were used to simulate the propagation of light through the lens. The Zernike coefficients (A_c) were linearly interpolated to produce PSFs at different locations on a 4096×2160 pixel imager. For simplicity, the coefficients are sampled on the imager's diagonal and then rotated for the required azimuth, assuming a rotationally symmetric lens. The resulting PSFs showed different shapes and levels of aberrations, with the tendency for the PSF to become smaller and rounder towards the center. The PSFs were normalized to constant energy using the l_1 norm. The algorithm used for this process is an approximation of the so-called superposition algorithm (SP): The iso-planar patches approach [40] assumes that the PSF does not vary substantially over a particular region, which allows for approximating the SP with convolution in overlapping patches. The patches are usually bilinearly interpolated. The approximation allows for greatly reduced storage and runtime requirements for the defocused dataset creation. If all PSFs across the imager have approximately equal sizes and shapes, the location dependence in the PSF is eliminated, reducing the equation to a simple convolution. This study utilizes the iso-planar patches algorithm with square patches of size $150px^2$. The PSF for each patch is taken from the patch's center. A total of 464×3 overlapping PSFs of size 28×28 were available for precise simulation.

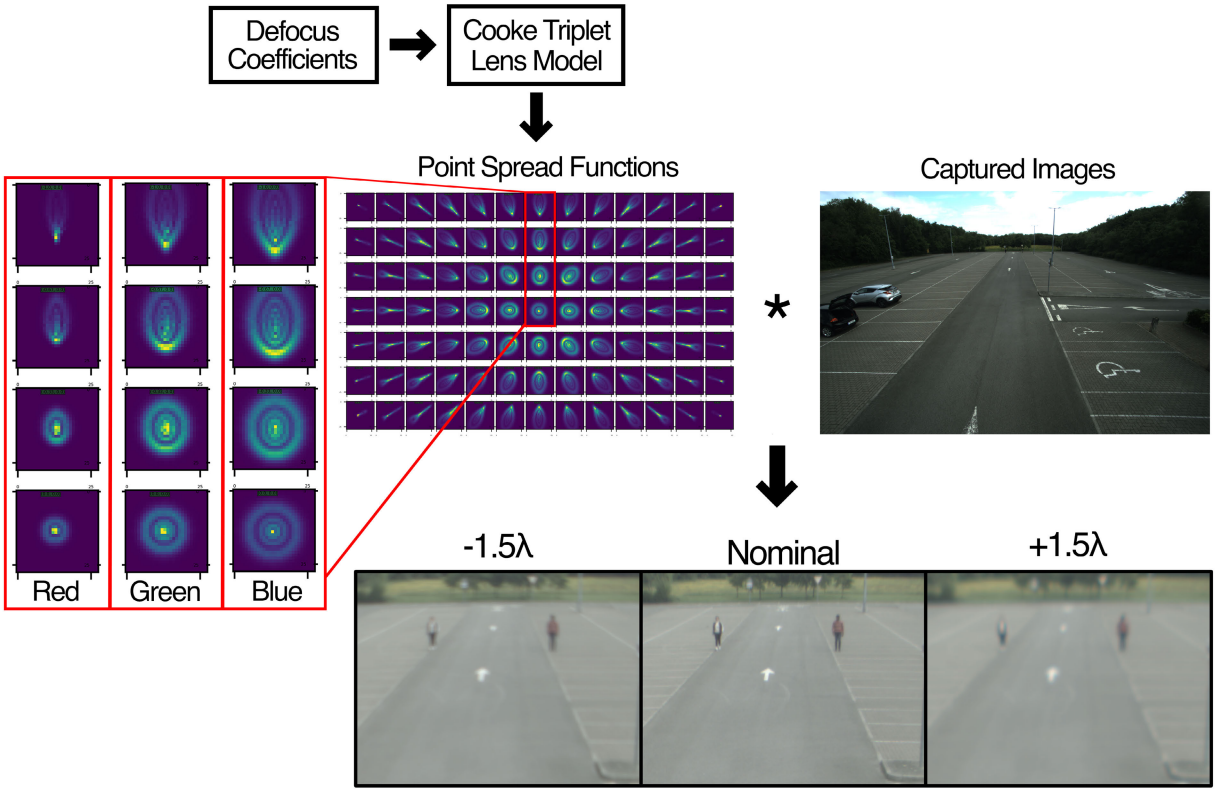


FIGURE 2. Defocused image generation pipeline used to generate a defocused dataset with varying levels of physically realistic lens blur. Example PSFs for the full imager field of view for a single color channel and blur level are shown in the middle of the diagram. Red, green, and blue PSFs are also shown on the left, highlighting that different wavelengths yield different PSFs. Three defocused images are shown on the bottom, with defocus coefficients of -1.5λ, 0λ(Nominal), and $+1.5\lambda$.

Fig. 2 shows an example of green channel PSFs where the middle PSFs are more uniform in size and shape than the outer PSFs. The red, green, and blue PSFs are also shown, and it is evident that the PSFs across different color channels have differing sizes and shapes. Larger PSFs cause more blur and elongated PSFs cause astigmatism to occur due to the PSF blurring the region non-uniformly in all directions. An example image from the dataset is on the right of Fig. 2, and examples of the physically realistic lens blur are shown at the bottom with varying Zernike defocus coefficients. Altering the Zernike defocus coefficient yields different PSFs, changing their size and shape, leading to unique optical blur and astigmatism for each wavelength.

The lens blur in this study is equivalent to inducing a sensor/lens misalignment by offsetting the Zernike defocus coefficient, A_4, with different values and using the wavefront expansion from Eq. 1 with $\Delta_4 \in \{-1.5, -1.0, -0.5, 0.0, +0.5, +1.0, +1.5\}$. Three levels of lens blur are visualized in Fig. 2.

$$W_\lambda(\rho, \varphi) = \sum_{c=1}^{C} (A_c + \Delta_c) \cdot Z_c(\rho, \varphi) \tag{1}$$

Wavefront aberration W_λ is evaluated at polar coordinates ρ, φ of a circular pupil, wavefront coefficient A_c, and the corresponding polynomial Z_c for Fringe index c. The wavefront is expanded into the first $C = 20$ polynomials, which represents a highly non-linear decomposition of the wavefront. The model additionally depends on wavelength λ. In this study, three wavelengths λ are sampled to model red, green, and blue color channels using $\{0.6563\mu m, 0.5876\mu m$ and $0.4861\mu m\}$ mapped to the red, green, and blue color channels of an image.

In our controlled dataset, the objects are, on average, located approximately halfway between the center and top of the imager. The application of the different defocus models produces different blurring: While the pedestrians appear relatively sharp with the nominal, or 0λ offset, model in Fig. 2, the two defocused images, with -1.5λ and $+1.5\lambda$ offsets, show that the pedestrians are severely blurred. Visually, the two defocused images are not identical, and the blur on the pedestrians appears more extensive for $\Delta_4 = +1.5\lambda$. The difference in blurring between the two defocused images, even though they share the same magnitude, can be understood from Fig. 3: The different RMS wavefront errors correlate to what is observed in Fig. 2.

At the pedestrian's location in the imager (dotted line in Fig. 3), the blue curve, visualizing $+1.5\lambda$ defocus offset, proceeds at around $9\mu m$, whereas -1.5λ (red) proceeds significantly lower.

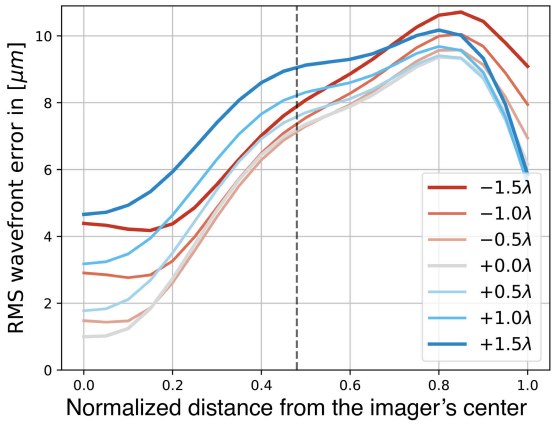


FIGURE 3. RMS wavefront error over the field. Higher wavefront error corresponds to more blur at a particular normalized distance d from the imager's center. Different defocus coefficient offsets represent defocusing the lens. The RMS at -1.5 (red) starts at a lower value compared to $+1.5$ (blue) and then constantly increases towards the imager's diagonal. Conversely, the blue curve flattens at medium field positions and then decreases towards the imager's diagonal. The objects within Fig. 2 are located at $d = 0.48$ (dotted line). Thus, the cause of the differently blurred pedestrians in Fig. 2 can be understood by the $\pm 1.5\lambda$ (blue and red) curves.

Furthermore, the refocusing sometimes yields less blurred regions compared to the nominal position with no defocus offset: The -0.5λ curve (orange) in Fig. 3 proceeds slightly lower than the nominal curve (grey) in a region between approximately 0.2 and 0.5 of the normalized distance from the imager's center. The nominal setting has lower sharpness than the -0.5λ offset lens within this range. However, this is expected behavior because the overall wavefront error of the nominal curve is still lower than the -0.5λ offset curve.

C. DEEP LEARNING MODEL SELECTION

This study investigates six state-of-the-art object detection algorithms as outlined in Table 1. This selection is a representative sample of the latest architectures, including single-single, two-stage, and transformer-based architectures. Representing single-stage algorithms are YOLOv8m and FCOS with a ResNet50 FPN backbone. Two predominant two-stage algorithms tested are Faster RCNN with a ResNet50 FPN backbone and Cascade RCNN with a ResNet50 FPN backbone. Transformer-based algorithms utilized in this study are Deformable Detection Transformer (DETR) and Faster RCNN with a Swin Tiny FPN backbone.

The initial deep learning architectures for object detection included a localization step, in which bounding boxes were predicted that contained objects, and then a classification step, in which an object classifier predicted the class within each bounding box, known as a two-stage architecture. In 2014, Girshick et al. [41] introduced Regions with CNN features (RCNN), which utilized Selective Search [42] for detection and Convolutional Neural Networks (CNNs) paired with support vector machines (SVMs) for classification. Subsequent work on RCNN led to the development of Fast RCNN [43] and Faster RCNN [44] with iterative improvements in inference speed and performance. This study evaluates multiple versions of Faster RCNN, which replaces Selective Search with a CNN-based Region Proposal Network (RPN) to obtain bounding box coordinates.

In this study, we evaluate *Faster RCNN with the ResNet50 FPN backbone*. The Residual Network (ResNet) [45] combined with a Feature Pyramid Network (FPN) [46] is a multi-scale feature extraction backbone that combines the strengths of both ResNet and FPN. ResNet can learn very deep feature representations by using skip connections to bypass a few layers at a time, which helps mitigate the problem of vanishing gradients. FPN, on the other hand, is a network architecture that allows for the detection of objects at multiple scales by combining feature maps of different resolutions. FPNs use a top-down pathway that upsamples feature maps from higher layers and fuses them with feature maps from lower layers to create a feature pyramid. Combining the strengths of both ResNet and FPN, the ResNet FPN backbone in Faster RCNN can generate high-quality multi-scale features necessary for accurate object detection.

Cascade RCNN is another two-stage object detection algorithm under investigation in this study that also uses the ResNet50 FPN backbone. It was introduced in 2018 by Cai et al. [47] and builds upon the Faster RCNN framework by introducing a cascade of classifiers to refine the bounding box proposals. Cascade RCNN uses a series of increasingly complex classifiers to filter out false positives at each stage of the detection process. The first stage uses a relatively simple classifier to filter out the most apparent background regions, while the subsequent stages use increasingly complex classifiers to refine the remaining proposals. By using a cascade of classifiers, Cascade RCNN achieves higher performance than Faster RCNN. Cascade RCNN also introduces a novel IoU (Intersection over Union) balancing technique to address the class imbalance issue that occurs when the number of negative samples far exceeds the number of positive samples, leading to biased classifiers. The IoU balancing technique used in Cascade RCNN helps mitigate this issue by assigning different IoU thresholds to positive and negative samples.

Fully Convolutional One-Stage (FCOS) algorithm [48], introduced in 2019, is a single-stage object detection algorithm used in this study. It aims to balance the speed of single-stage detectors, like YOLO, and the accuracy of two-stage detectors, like Faster RCNN. The FCOS architecture in this study has the same ResNet50 FPN backbone as Faster RCNN and Cascade RCNN. However, unlike all previous algorithms that predict bounding boxes at predefined anchor locations, FCOS uses an object-centric approach where each location on the feature map predicts the center of an object, its size, and its class. This approach enables FCOS to handle objects of different scales and aspect ratios more effectively. FCOS also uses a focal loss function that assigns

higher weights to hard examples during training, which helps address the class imbalance issue in object detection.

In 2016, Redmond et al. [49] developed the You Only Look Once (YOLO) architecture, the first widely recognized object detection architecture to employ a single-stage design. In YOLO, the localization and classification tasks were unified and solved by a single CNN, allowing for bounding box predictions and class labels to be generated in a single pass using the entire input image. YOLO's single-stage approach resulted in faster inference times compared to two-stage models like Faster RCNN, although it did come at a slight performance cost. The development of YOLO has been conducted by multiple groups, increasing performance and speed [50], [51], [52], [53], [54], [55] and for this study, we utilize *YOLOv8* [55] which was released in 2023. YOLOv8 is the first YOLO architecture since the original version to have an object-centric prediction output, similar to FCOS, with neither utilizing anchor boxes.

Another object detection algorithm that does not utilize anchor boxes is *Deformable Detection Transformer (DETR)* [56]. Deformable DETR is an end-to-end transformer-based object detection algorithm introduced in 2020 by Facebook AI Research as a follow-on from their original DETR transformer-based architecture [57]. This architecture is a two-stage design, as the image is initially passed through a CNN backbone, in this case, a modified ResNet50 backbone, to generate a feature map. The feature map is then passed through a set of transformer encoder layers that use self-attention mechanisms to encode the spatial information of the image. It uses a set of learned object queries to attend to different regions of the image. The transformer decoder takes the feature map, and the object queries as inputs and generates a set of bounding box predictions and class probabilities for each object query. Deformable DETR uses deformable convolutional layers in the ResNet backbone and a deformable attention mechanism to handle the variability of object shapes and sizes.

Deformable DETR is an end-to-end transformer-based object detection model; however, transformer-based backbones have also been created. In 2021, Liu et al. created the Swin Transformer [58] vision backbone, which is a type of CNN architecture that uses a hierarchical vision transformer to extract features from input images. It divides the input image into patches of different sizes and applies convolutional operations on each patch. The patches are then processed hierarchically, with the features from one stage being passed to the next for further processing and aggregation with features from other scales. It also uses a technique called shifted windows to reduce the computation cost of processing multi-scale windows, allowing it to be more computationally efficient than ResNet. In this study, we investigate a *Faster RCNN model with a Swin Tiny FPN backbone*. Swin Tiny, the smallest version of Swin, is used in this study as it has a similar complexity to ResNet50, as seen in Table 1.

TABLE 1. Object detection model breakdown.

Investigated Object Detection Models				
Model	Backbone	Architecture	Params (M)	Year
YOLOv8m	CSP	Single-Stage	25.9	2023
FCOS	ResNet50 FPN	Single-Stage	32	2019
Cascade RCNN	ResNet50 FPN	Two-Stage	41	2017
Faster RCNN	ResNet50 FPN	Two-Stage	41.5	2016
Deformable DETR	ResNet50	Transformer	41	2020
Faster RCNN	Swin Tiny FPN	Transformer	41	2021

The six investigated object detection models represent the latest innovations in object detection. Each object detection algorithm was trained on the COCO object detection dataset, with an image size of 640×480, to keep the training constant.

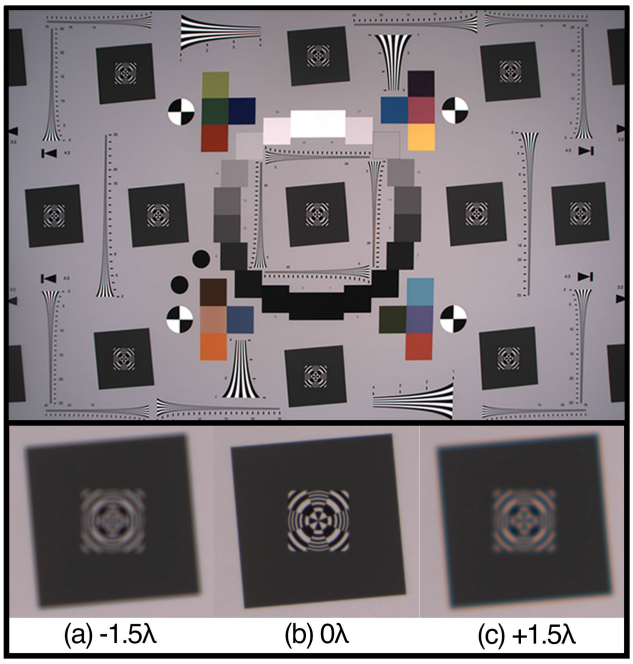


FIGURE 4. Original image of Imatest test chart utilized to measure MTF with example levels of blur shown beneath. (a) shows the middle slanted edge feature blurred to the -1.5λ blur level, (b) illustrates the middle slanted edge with the nominal lens model applied, with no defocus offset, and (c) refers to the +1.5λ blur level.

D. METRICS

To evaluate the overall performance of an object detection algorithm on a set of images, the predictions made by that algorithm are compared with a manually annotated ground truth. Within the PASCAL VOC [59] object detection benchmark released in 2012, the standard metric is average precision (AP) with a static IoU of 50%. This metric, called AP50, is the area under the precision-recall curve. The curve is generated by adjusting the confidence threshold over a set of predictions. At a low confidence threshold, there are fewer missed objects but more predictions for objects that do not exist, leading to a higher recall but lower precision.

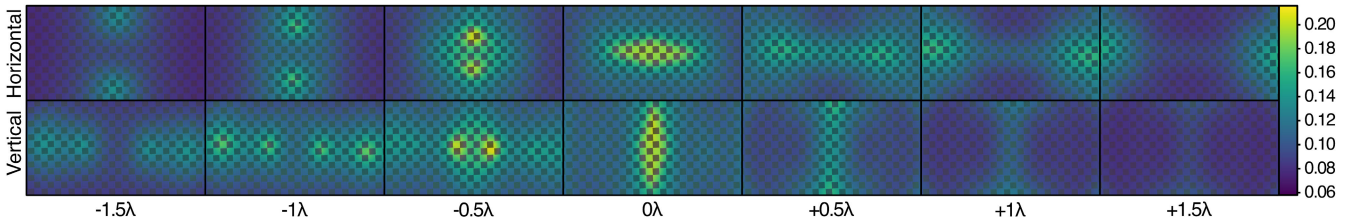


FIGURE 5. Heatmap of MTF50 values interpolated throughout the image. The image background shows the custom slanted edge chart used to measure the MTF50 values for each horizontal and vertical edge. The top row shows the MTF50 values of the horizontal slanted edge interpolated throughout the imager. The bottom row shows the MTF50 values of the vertical slanted edge interpolated throughout the imager.

Increasing the confidence threshold inverts this relationship with more missed objects but fewer predictions of objects that do not exist, leading to a higher precision but a lower recall. The precision and recall values associated with varying confidence thresholds are then plotted and interpolated, and the area under the curve is AP50. A true positive, for AP50, is defined as any prediction with an IoU greater than 50% with a ground truth bounding box. The primary metric in the COCO [11] object detection benchmark, published in 2015, is AP5095. To calculate AP5095, the IoU threshold varies from 50% to 95% in 5% increments, and the results are averaged. Given the varying IoU threshold, AP5095 scores more accurate bounding box localization positively. AP5095 is the standard metric to benchmark object detection performance because it encompasses true positives, false positives, false negatives, and IoU and is the primary metric used in this study. To evaluate the robustness of the detection models to different levels of lens blur, ΔAP is utilized. ΔAP is the absolute difference between the highest AP5095 and the lowest AP5095 measured across all blurred datasets.

Given the spatial non-uniformity associated with the lens model in this study, spatially dependent metrics such as the Spatial Recall Index (SRI) [17] and the Spatial Precision Index (SPI) [18] could be employed to directly analyze the lens's impact on object detection performance. However, these metrics are most robust on large datasets with hundreds of object samples at each pixel location, so they have not been evaluated in this study.

In previous works, where the impact of blur on deep learning models was investigated, blur level was generally related to the kernel size of a Gaussian blur, limiting the reproducibility of these studies in the real world. MTF is used to measure the ability of an imaging system to capture spatial frequencies [60]. ISO standard 12233:2017 [61] provides a methodology for determining the resolution and MTF of digital cameras and specifies a test chart used to measure these parameters. The standard also provides guidelines for interpreting MTF measurements, specifying key parameters such as MTF50, the spatial frequency at which MTF is 50% of its maximum value. MTF50 is the most commonly used metric from an MTF curve, as it correlates well with perceived sharpness. In this study, the Imatest eSFR test chart [62] is used to measure the MTF at different regions throughout the imager at various levels of blur. An image of this test chart was captured in a controlled environment, as seen in Fig. 4, after which the same lens model and defocus parameters were employed to degrade the image. While an official eSFR Imatest test chart is utilized to characterize the MTF of the various levels of blur, in this study, MTF50 is also approximated at a pixel level. To approximate MTF50 at a pixel level, a test chart was generated with a 13×7 grid of slanted edges that were then measured according to ISO12233:2017. The MTF50 measurements taken from each slanted edge were bilinearly interpolated throughout the imager for each pixel. The MTF50 heatmaps are shown in Fig. 5 with the generated test chart visible in the background.

IV. RESULTS AND DISCUSSION

A. MODULATION TRANSFER FUNCTION

To relate the blur level within the images to MTF, an image of an Imatest [62] eSFR test chart image was captured in a controlled environment, and the physically realistic lens blur was applied with the same defocus parameters, as can be seen in Fig. 4. The figure also shows that negatively defocused slanted edge (a) and positively defocused slanted edge (c) do not have the same defocus despite being defocused by the same magnitude, illustrating that negative and positive defocus do not result in the same blurred image, as is the case also for real lenses. While the nominal model, applied to the slanted edge (b), shows minimal chromatic aberrations, aberrations can be seen in (a) and (c), although with different angles and magnitudes.

Measuring MTF as outlined in ISO12233:2017 yields the results shown in Fig. 6. The top row shows the MTF curves for the slanted edge at the imager's top middle under three defocus levels: 1.5λ, 0λ, and $+1.5\lambda$. The bottom row shows the MTF curves for the slanted edge in the middle of the imager under the same three conditions. The solid lines represent the MTF for the vertical slanted edge, with dashed lines representing the horizontal slanted edge. As expected, the middle nominal MTF curve achieves the best result, with an MTF50 of $0.12098 cy/px$. In the middle nominal MTF curve, the vertical edge is slightly sharper than the horizontal edge, with the green channel being the sharpest, followed by red and then blue. The top nominal MTF, however, illustrates a vital aspect of a physically realistic lens model, astigmatism. The vertical lines are much sharper than the horizontal lines, even without any defocus applied

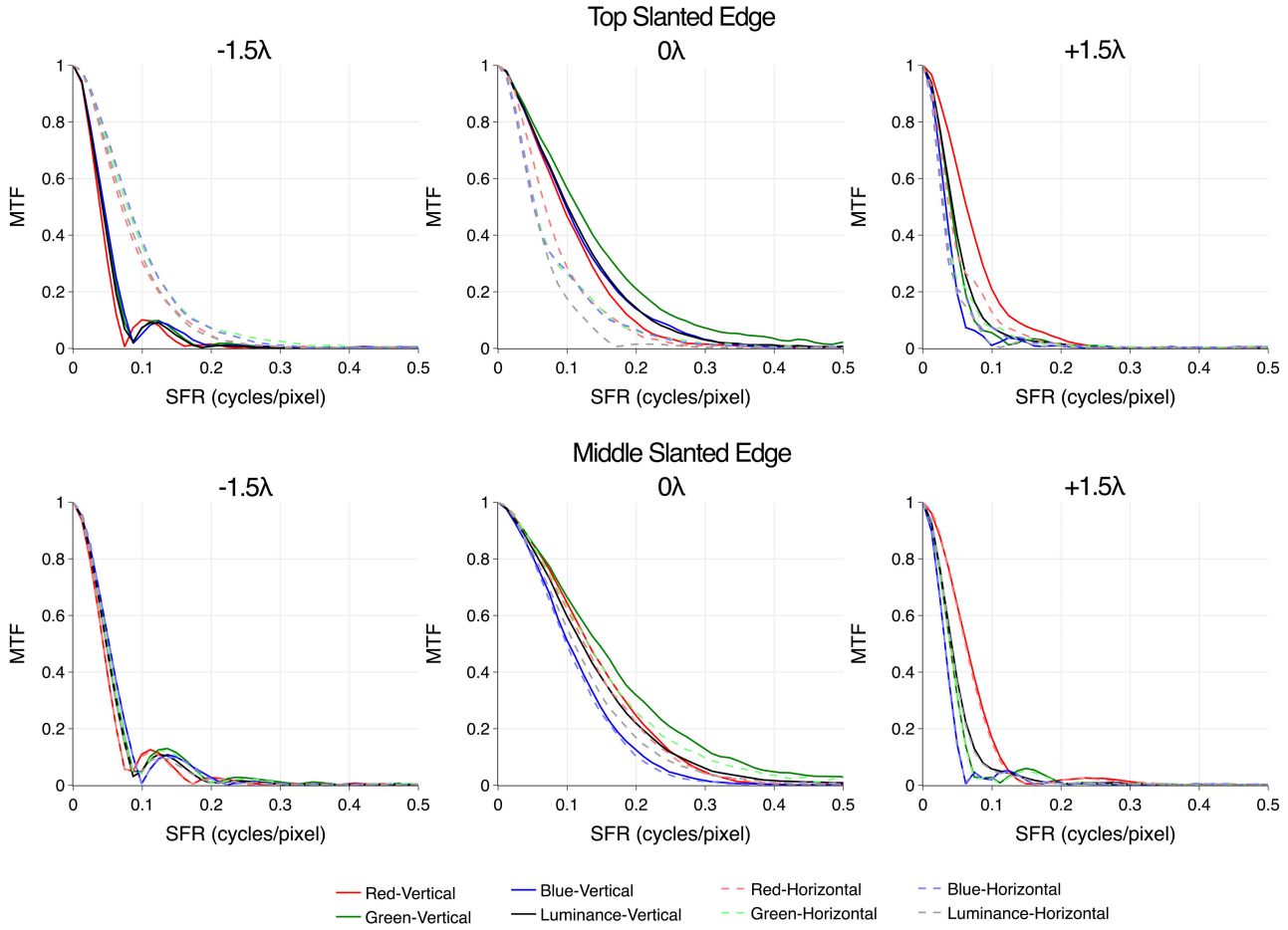


FIGURE 6. MTF plots, measured at the top of the image plane (top row) and the middle of the image plane (bottom row). The nominal lens model is in the center column, with 0λ defocus offset. The left column shows the -1.5λ defocus, and the right column shows the +1.5λ defocus.

due to the inherent astigmatism of the Cooke triplet lens model at the top of the imager. The observed astigmatism inverts when offsetting the defocus parameter to -1.5λ as the horizontal slanted edge is much sharper than the vertical. The aspect ratio of the PSF for the top middle region of the imager shifts from predominantly vertically orientated at the nominal position to horizontally orientated. Wavelength-dependent chromatic aberration can also be seen when comparing the nominal MTF curves to the $+1.5\lambda$ MTF curves because the sharpest channel in the nominal, green, gets replaced by red in the positively defocused MTF curves.

The MTF50 results, shown in Table 2, illustrate the sharpness of the image given the varying conditions. For the middle edge position, the highest MTF50 was observed at 0λ defocus for both the vertical and horizontal slanted edges. As the defocus deviates from 0λ in the positive and negative direction, the MTF50 lowers for both vertical and horizontal slanted edges. Looking solely at the vertical slanted edge results, the nominal, or 0λ, defocus produces an MTF50 result of $0.12098cy/px$. At $+1.5\lambda$ defocus, the MTF50 is reduced to $0.041287cy/px$; however, the -1.5λ defocus sees marginally less of a reduction, down at $0.047575cy/px$. This asymmetry in sharpness degradation is even more apparent at the 0.5λ defocus positions, with the positive defocus at $0.084495cy/px$, and the negative defocus at $0.11cy/px$. At the middle position of the imager, the MTF50 results from the vertical and horizontal slanted

TABLE 2. MTF50 values for different levels of defocus at different positions around the imager.

Defocus	Edge Position	MTF50 Vertical (cy/px)	MTF50 Horizontal (cy/px)
+1.λ	Middle	0.041287	0.040637
+1λ	Middle	0.056236	0.054987
+0.5λ	Middle	0.084495	0.079849
0λ	Middle	**0.12098**	**0.11006**
-0.5λ	Middle	0.11	0.10296
-1λ	Middle	0.069764	0.068073
-1.5λ	Middle	0.047575	0.047399
+1.5λ	Top	0.04276	0.030552
+1λ	Top	0.057035	0.034844
+0.5λ	Top	0.081225	0.042687
0λ	Top	**0.10096**	0.053692
-0.5λ	Top	0.083931	0.069822
-1λ	Top	0.057023	**0.079151**
-1.5λ	Top	0.041982	0.073814

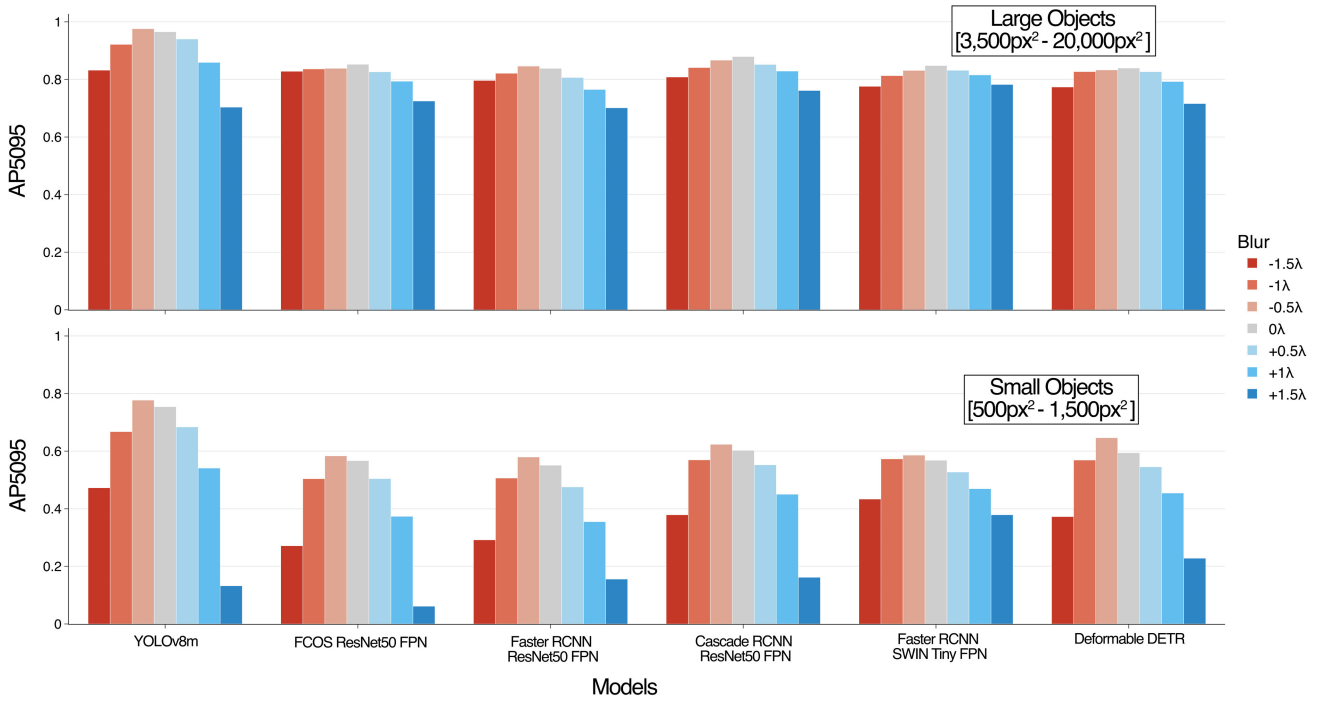


FIGURE 7. AP5095 of each model for the person class. Models are represented along the x-axis, with each blur level depicted with varying colors, outlined in the legend. The top figure contains large objects with areas between $3,500px^2$ and $20,000px^2$ corresponding to a person 15m to 30m from the camera. The bottom figure contains small person objects with areas between $1,250px^2$ and $3,500px^2$ corresponding to 50m to 100m from the camera.

edges share the same relationship, with the positive defocus causing more degradation than the negative defocus. This asymmetry in sharpness degradation illustrates that positive and negative defocus parameters must be evaluated individually.

Given a spatially uniform lens model, the MTF50 results would remain constant when looking at the different edge positions; however, due to the field curvature of the lens model used in this study, the top edge position results are significantly different from the middle slanted edge results. The MTF50 for the nominal defocus at the top edge position is $0.10096cy/px$ vertically and $0.053692cy/px$ horizontally. The variation between orientations highlights the astigmatism of the lens model, as seen in Fig. 6. The sharpness degradation at the top edge position, when varying the defocus parameter, is more symmetric surrounding the nominal defocus when looking at the vertical MTF50 results. The lowest vertical MTF50 between the top and middle edge positions are very similar. However, the top position horizontal MTF50 results are significantly reduced compared to the middle position. The highest MTF50 result from the top horizontal slanted edge is $0.079151cy/px$ at -1λ defocus, at 28% reduction in MTF50 when compared to the maximum value at the middle slanted edge. The chromatic aberration and astigmatism visible in these results are not replicated with previous, simpler blur models, such as a uniform Gaussian disk kernel, in [16], which highlights the importance of modeling lens blur utilizing PSFs.

To characterize the relationship between object detection performance and MTF, it was necessary to obtain an approximation of MTF for each pixel, and this was completed by generating a custom slanted edge chart with 91 slanted edges, yielding 364 MTF measurements, which were then bilinearly interpolated across the entire image. Fig. 5 illustrates the MTF50 values for horizontal edges in the top row and vertical edges in the bottom row for each level of defocus. As seen in Fig. 5, different regions within the image become sharper as the field curvature of the lens intersects the image plane caused by the defocus offset. At a defocus of -0.5λ, the area with the highest MTF50 on the horizontal slanted edge is between the middle and top of the image, a region that contains many of the objects in our controlled dataset.

B. OBJECT DETECTION PERFORMANCE

The six representative object detection models were evaluated on the seven degraded datasets with various defocus parameters, and the results for the person class can be seen in Fig. 7. Overall performance ranged from 6.15% AP5095 to 97.59% across all models, blur levels, and object sizes. In Fig. 7, the top row relates to model performance on large objects, between $3,500px^2$ and $20,000px^2$, corresponding to 15m to 30m away from the camera. The bottom row relates to the model's performance on small objects between $500px^2$ and $1,500px^2$, corresponding to 50m to 100m from the camera. Across all models and levels of blur, the average performance on the large objects is 82.42% AP5095 and 47.85% on the

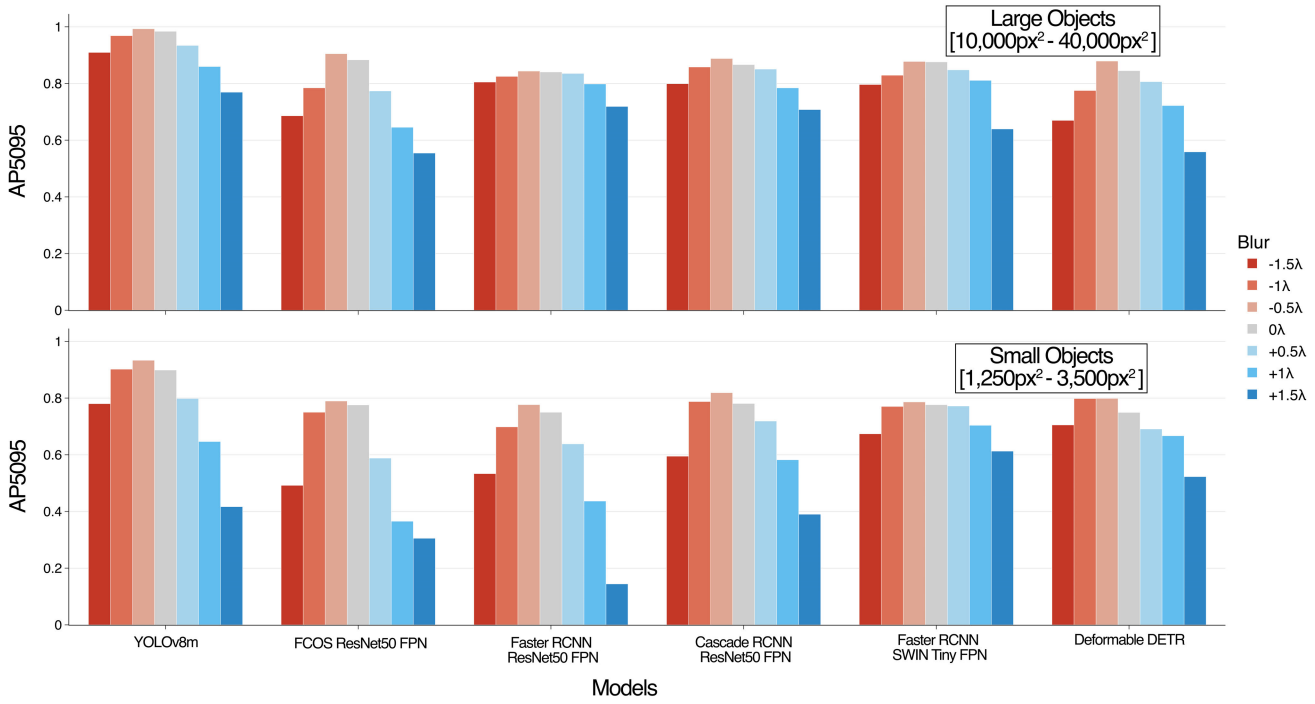


FIGURE 8. AP5095 of each model for the car class. Models are represented along the x-axis, with each blur level depicted with varying colors, outlined in the legend. The top figure contains large objects with areas between $10,000px^2$ and $40,000px^2$ corresponding to a car 15m to 30m from the camera. The bottom figure contains small objects with areas between $1,250px^2$ and $3,500px^2$ corresponding to 50m to 100m from the camera.

small objects, a substantial performance reduction, indicating that the small object dataset is a significantly more difficult task for the object detection models.

The percentage performance difference between large and small objects, across all models and blur levels, is a 41.94% reduction; however, when analyzing just the nominal defocus dataset, there is only a 30.34% reduction. The additional performance degradation between the large and small datasets, when including the various blur levels, proves that image blur disproportionately impacts small objects. For the large object dataset, the models performed best with a nominal defocus; however, -0.5λ defocus closely follows the nominal and is even the top-performing blur level for YOLOv8m and Faster RCNN ResNet50 FPN models with the large object dataset. Objects within the large object person dataset are approximately centered or slightly above the center of the image. Fig. 5 shows that the 0λ defocus position has the highest MTF50 value for objects in the center of the imager, with the -0.5λ having the next highest; however, the -0.5λ peaks in MTF50 slightly above the imager's center. Within our ITS infrastructure node perspective dataset, the camera is angled at 20° towards the ground, resulting in more distant objects appearing higher up on the image plane. For the small person object dataset, the positions of most objects are between the center and top of the imager, aligning more closely with the MTF50 hotspot in the -0.5λ defocus position as seen in Fig. 5. The alignment of objects in the small person object dataset, within the higher MTF50 region of the -0.5λ defocus, likely explains why the models achieve a higher AP5095 on the -0.5λ compared to the 0λ across each model. For this dataset, the models performed best when blurred with a defocus offset due to object positions aligning with the lens' field curvature, highlighting the importance of validating a safety-critical object detection system utilizing a physically realistic lens model alongside spatially dependent performance metrics such as SRI and SPI.

TABLE 3. The difference in AP5095, or ΔAP, between the maximum and minimum performance for each model on the large and small person object datasets. A lower ΔAP represents higher robustness to image blurring.

Model	Large Object ΔAP	Small Object ΔAP
YOLOv8m	27.24%	64.46%
FCOS ResNet50 FPN	12.74%	52.2%
Faster RCNN ResNet50 FPN	14.48%	42.4%
Cascade RCNN ResNet50 FPN	11.79%	46.18%
Faster RCNN Swin Tiny FPN	**7.22%**	**20.72%**
Deformable DETR ResNet50	12.33%	41.84%

The model with leading performance is YOLOv8m at 97.59% AP5095 on the large person object dataset and 77.72% on the small person object dataset. The next top performer in the large object dataset is Cascade RCNN ResNet50 FPN at 87.94% AP5095, and in the small object dataset is Deformable DETR with 64.65% AP5095. While YOLOv8m significantly outperforms other models on average, and with almost half the number of parameters, its performance on the +1.5λ defocused dataset drops beneath most other models. The robustness of each model to blur is evaluated based on the

difference between its best and worst AP5095 performance (ΔAP) over the seven degraded datasets, as shown in Table 3. In the person dataset, the Swin-based Faster RCNN model is the most robust model for image blurring, losing just 7.22% AP5095 on the large object dataset and 20.72% on the small object dataset. This model's high robustness is achieved mainly by retaining a significant amount of AP5095 on the dataset blurred with the $+1.5\lambda$ defocus parameter as this defocus parameter has the lowest MTF50 values, as seen in Fig.5 and is the defocus parameter causing the highest performance degradation. The transformer-based Swin Tiny FPN backbone appears to be more robust to blurring than the ResNet50 FPN backbone with the main difference between the two backbones being that ResNet50 utilizes a CNNs architecture and Swin utilizes a transformer-based architecture. Across all blur levels and object sizes, the Faster RCNN model with the Swin Tiny FPN backbone achieved 69.97% AP5095 compared to just 60.65% AP5095 for the ResNet50 FPN backbone. The Faster RCNN model with a ResNet50 FPN backbone, Cascade RCNN model, and Deformable DETR model all appear similarly robust with a ΔAP of 11.79 - 14.48% on the large object dataset and 41.84 - 46.18% on the small object dataset. YOLOv8m is the least robust model averaged across both the small and large object datasets; however, while robustness is the worst, it outperforms the other models in AP5095 on all blurred datasets except the dataset with a $+1.5\lambda$ defocus applied that severely affects the model's performance. YOLOv8m's low performance on the $+1.5\lambda$ defocus data could be due to its smaller number of parameters, single-stage design, or not utilizing anchor boxes. FCOS shares many of these attributes with a single-stage design, no anchor boxes, and fewer parameters compared to the two-stage and transformer-based models, but with more parameters than YOLOv8m. FCOS' robustness to image blur is average on the large object dataset, with a 12.74% ΔAP, but much worse on the small object dataset, with a 52.5% ΔAP. Deformable DETR also does not use anchor boxes, but it has similar blur robustness to the models that do, suggesting that the fewer parameters and single-stage design of YOLOv8m and FCOS could be the cause of the worse blur robustness.

TABLE 4. The difference in AP5095, or ΔAP, between the maximum and minimum performance for each model on the large and small car object datasets. A lower ΔAP represents higher robustness to image blurring.

Model	Large Object ΔAP	Small Object ΔAP
YOLOv8m	22.37%	51.65%
FCOS ResNet50 FPN	35.06%	48.43%
Faster RCNN ResNet50 FPN	**12.48%**	63.23%
Cascade RCNN ResNet50 FPN	18.02%	42.89%
Faster RCNN Swin Tiny FPN	23.81%	**17.36%**
Deformable DETR ResNet50	32.03%	27.61%

In Fig. 8, the top row relates to the performance of models on large objects, between 10,000px^2 and 40,000px^2, corresponding to cars that are 15m to 30m away from the camera. The bottom row shows performance on small objects between 1,250px^2 and 3,500px^2, corresponding to 50m to 100m from the camera. Across all models and levels of blur, the average performance on large objects is 80.97% AP5095 for the car class, a similar result to the large-person object dataset. The average performance for the small car object dataset is 66.96% AP5095, an increase of 19.11% AP5095 over the average performance on the small person object dataset due to cars being physically larger than people, thus are represented by more pixels within the normalized distance threshold. When analyzing the car dataset, the percentage performance difference between the large and small objects, over all the models and blur levels, is a 17.3% reduction; however, when analyzing just the nominal defocus dataset, there is only a 10.65% reduction. Comparing the large and small car datasets, it is clear that the small dataset is more difficult for the models and even more so when there is image blurring. The performance reduction due to blurring is slightly less with the car object dataset than with the person object dataset. This is likely due to the higher number of pixels representing the car objects and their wider aspect ratios. The best-performing model on the car object dataset was YOLOv8m for both large and small object datasets. As seen in Table 4, the most robust model for the small car object set is the same as for the small person object dataset, Faster RCNN Swin Tiny FPN, and when the robustness score is averaged between the large and small object sets, this model is the overall most robust. Unexpectedly, the most robust object detection model for the large object dataset is Faster RCNN with the ResNet50 FPN backbone at 12.48% ΔAP while having the worst small object dataset robustness at 63.23%. The high robustness of the Faster RCNN ResNet50 FPN model could be due to the larger pixel area or the more horizontal aspect ratio of the car objects. Due to lens astigmatism, the aspect ratio of objects within a safety-critical object detection system will likely influence the object detection system's performance on that object, given that the vertical or horizontal features may have a different level of sharpness.

C. OBJECT DETECTION PERFORMANCE VS. MTF

The MTF measurements captured throughout the imager, as discussed in Section III-D, are utilized to populate the MTF50 heatmaps seen in Fig. 5. This method allows us to approximate MTF50 at each pixel location. The average MTF50 for the overall dataset, including person and car classes with small and large objects, is measured by getting an MTF50 value associated with each object within the ground truth annotations and averaging them. Each dataset, blurred with a different defocus parameter, has an associated MTF50 value and AP5095, averaged over the person and car classes, utilizing both the large and small datasets, making it possible to evaluate the correlation between MTF50 and object detection performance. Fig. 9 shows the relationship between MTF50 and each model's AP5095 performance on our dataset. There appears to be a strong non-linear correlation

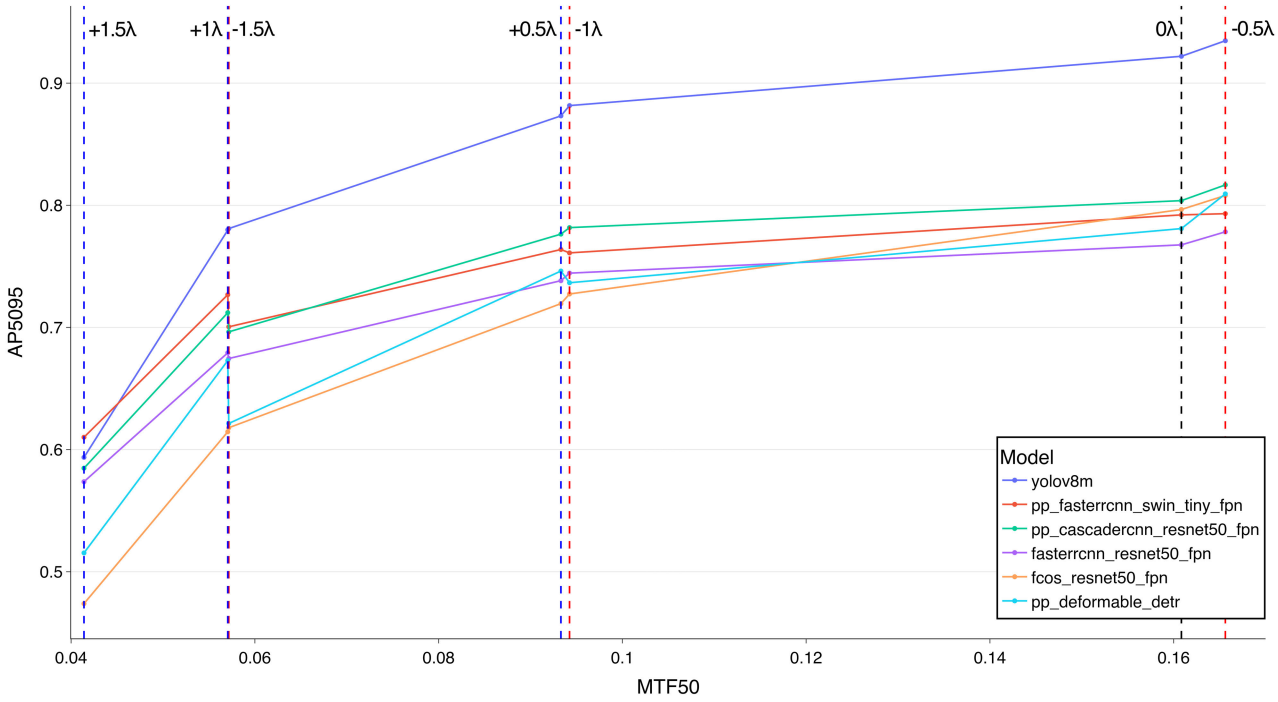


FIGURE 9. The lens model, with a 0λ defocus offset, three positive and three negative defocus offsets were applied to our dataset, and a slanted edge chart. MTF50 values for each defocus offset were measured from the test chart over the imager and bilinearly interpolated to approximate an MTF50 value for each pixel within the imager. The average MTF50 was measured based on the MTF50 associated with each object location. Each object detection model predicted objects within the blurred datasets. The object detection model's AP5095 metrics are plotted against the average MTF50 metric.

between MTF50 and AP5095, proving that sharpness directly impacts object detection performance. While it is clear that there is a correlation, the positive defocus parameters do not align with the negative defocus parameters, as seen by the figure's annotations, causing an oscillation between the two. The difference between the positive and negative defocus parameters is likely due to the wavelength-dependent nature of the lens model, causing different chromatic aberrations that may improve or degrade object detection performance for the severely blurred datasets. An overall limitation of the methodology is that this lens model, containing just three lens elements, produces MTF50 results that are relatively low at a maximum of $0.22cy/px$, whereas typical wide-angle lenses generally achieve an MTF50 of$0.3cy/px$ with a six-element lens.

V. CONCLUSION

In this study, a controlled raw image dataset was acquired and annotated, and a physically realistic lens model was utilized to blur the images to varying amounts by offsetting the defocus coefficient of the model's Zernike polynomials expansion. The sharpness of blurred images was then characterized utilizing MTF measurements that were taken with the same camera as the raw dataset in a controlled environment, according to ISO12233:2017. The RMS wavefront error and MTF50 heatmaps were generated to illustrate the non-uniform sharpness present in the blurred dataset. The performance of six state-of-the-art object detection models was benchmarked on the blurred datasets. The AP5095 performance of the models was evaluated against the differing defocus parameters for person and car classes and split into large and small objects. While YOLOv8m performed best overall in AP5095 performance across the various datasets, the Faster RCNN Swin Tiny FPN was the most robust model against image blurring. Each model was significantly degraded at extreme levels of blur; however, this degradation was inconsistent across each model. For a safety-critical object detection system, the same methodology utilized in this study is needed to determine the minimum MTF50 allowable to achieve the minimum object detection performance required to meet a safety standard. Having characterized the MTF50 and object detection performance associated with the various defocus parameters, the relationship between MTF50 and object detection performance was evaluated, making it possible to see that MTF50 and AP5095 are strongly correlated. We have shown, in this study, that a physically realistic lens model must be utilized when validating mass-produced safety-critical object detection systems because lens field curvature causes the image sharpness to be spatially non-uniform, leading to different levels of object detection performance within a single image frame.

The next step in investigating the relationship between MTF and object detection performance is to utilize a

larger dataset with many object samples covering the entire image frame. More defocus parameters and object detection models would provide more data points to characterize the relationship fully. Spatial metrics such as SRI and SPI must be utilized to measure the spatially non-uniform object detection performance caused by the lens field curvature. An investigation is necessary into how robust an object detection model can become by using the physically realistic lens model defocus parameter as a data augmentation step. The robustness of the trained models can also be quantified by utilizing the same methodology in this study.

REFERENCES

[1] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[2] H. Zhu, H. Wei, B. Li, X. Yuan, and N. Kehtarnavaz, "A review of video object detection: Datasets, metrics and methods," *Appl. Sci.*, vol. 10, no. 21, p. 7834, Nov. 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/21/7834

[3] Y. Chen, Y. Xie, L. Song, F. Chen, and T. Tang, "A survey of accelerator architectures for deep neural networks," *Engineering*, vol. 6, no. 3, pp. 264–274, Mar. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2095809919306356

[4] A. Mhalla, T. Chateau, S. Gazzah, and N. E. B. Amara, "An embedded computer-vision system for multi-object detection in traffic surveillance," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 11, pp. 4006–4018, Nov. 2019.

[5] J. Clancy, D. Mullins, B. Deegan, J. Horgan, E. Ward, P. Denny, C. Eising, E. Jones, and M. Glavin, "Feasibility study of V2X communications in initial 5G NR deployments," *IEEE Access*, vol. 11, pp. 75269–75284, 2023.

[6] J.-B. Tomas-Gabarron, E. Egea-Lopez, and J. Garcia-Haro, "Vehicular trajectory optimization for cooperative collision avoidance at high speeds," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1930–1941, Dec. 2013.

[7] A. S. Tigadi, N. Changappa, S. Singhal, and S. Kulkarni, *Autonomous Vehicles: Present Technological Traits and Scope for Future Innovation*. Cham, Switzerland: Springer, 2021, pp. 115–143, doi: 10.1007/978-3-030-59897-6_7.

[8] M. M. Antony and R. Whenish, *Advanced Driver Assistance Systems (ADAS)*. Cham, Switzerland: Springer, 2021, pp. 165–181, doi: 10.1007/978-3-030-59897-6_9.

[9] P. Li and H. Zhao, "Monocular 3D object detection using dual quadric for autonomous driving," *Neurocomputing*, vol. 441, pp. 151–160, Jun. 2021.

[10] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," Oct. 2020, *arXiv:1808.08750*.

[11] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollar, "Microsoft COCO: Common objects in context," Feb. 2015, *arXiv:1405.0312*.

[12] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2633–2642.

[13] O. van Zwanenberg, S. Triantaphillidou, R. B. Jenkin, and A. Psarrou, "Estimation of ISO12233 edge spatial frequency response from natural scene derived step-edge data," *J. Imag. Sci. Technol.*, vol. 65, no. 6, Nov. 2021, Art. no. 060402.

[14] Q. Jueqin. *Qiujueqin/Fast-Openisp: Fast-Openisp: A Faster Re-Implementation of Openisp*. Accessed: Jan. 3, 2024. [Online]. Available: https://github.com/QiuJueqin/fast-openISP

[15] P. Müller and A. Braun, "Simulating optical properties to access novel metrological parameter ranges and the impact of different model approximations," in *Proc. IEEE Int. Workshop Metrology Automot. (MetroAutomotive)*, Jul. 2022, pp. 133–138.

[16] P. Müller, A. Braun, and M. Keuper, "Impact of realistic properties of the point spread function on classification tasks to reveal a possible distribution shift," in *Proc. NIPS*, 2022, pp. 1–15.

[17] P. Müller, M. Brummel, and A. Braun, "Spatial recall index for machine learning algorithms," *London Imag. Meeting*, vol. 2021, no. 1, pp. 58–62, 2021.

[18] M. Brummel, P. Müller, and A. Braun, "Spatial precision and recall indices to assess the performance of instance segmentation algorithms," *Electron. Imag.*, vol. 34, no. 16, pp. 1–6, Jan. 2022. [Online]. Available: https://library.imaging.org/ei/articles/34/16/AVM-101

[19] *Opticstudio—Optical, Illumination & Laser System Design Software*. Accessed: Jan. 3, 2024. [Online]. Available: https://www.zemax.com/pages/opticstudio

[20] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2001, pp. 511–518.

[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[22] S. Karahan, M. Kilinc Yildirum, K. Kirtac, F. S. Rende, G. Butun, and H. K. Ekenel, "How image degradations affect deep CNN-based face recognition?" in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2016, pp. 1–5.

[23] Z. Pezzementi, T. Tabor, P. Hu, J. K. Chang, D. Ramanan, C. Wellington, B. P. W. Babu, and H. Herman, "Comparing apples and oranges: Off-road pedestrian detection on the NREC agricultural person-detection dataset," Oct. 2017, *arXiv:1707.07169*.

[24] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0278364913491297

[25] S. Gilroy, E. Jones, and M. Glavin, "Overcoming occlusion in the automotive environment—A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 1, pp. 23–35, Jan. 2021.

[26] S. Gilroy, M. Glavin, E. Jones, and D. Mullins, "Pedestrian occlusion level classification using keypoint detection and 2D body surface area estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Montreal, BC, Canada, Oct. 2021, pp. 3826–3832. [Online]. Available: https://ieeexplore.ieee.org/document/9607551/

[27] H. Nada, V. A. Sindagi, H. Zhang, and V. M. Patel, "Pushing the limits of unconstrained face detection: A challenge dataset and baseline results," Aug. 2018, *arXiv:1804.10275*.

[28] D. Temel, M.-H. Chen, and G. AlRegib, "Traffic sign detection under challenging conditions: A deeper look into performance variations and spectral characteristics," 2019, *arXiv:1908.11262*.

[29] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," 2019, *arXiv:1903.12261*.

[30] T. Brophy, D. Mullins, A. Parsi, J. Horgan, E. Ward, P. Denny, C. Eising, B. Deegan, M. Glavin, and E. Jones, "A review of the impact of rain on camera-based perception in automated driving systems," *IEEE Access*, vol. 11, pp. 67040–67057, 2023.

[31] M. Aqqa, P. Mantini, and S. Shah, "Understanding how video quality affects object detection algorithms," in *Proc. 14th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, Prague, Czech Republic: SCITEPRESS-Science and Technology Publications, 2019, pp. 96–104.

[32] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," 2016, *arXiv:1604.04004*.

[33] Y. Zhou, S. Song, and N.-M. Cheung, "On classification of distorted images with deep convolutional neural networks," 2017, *arXiv:1701.01924*.

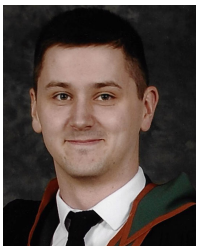
[34] B. RichardWebster, S. E. Anthony, and W. J. Scheirer, "PsyPhy: A psychophysics driven evaluation framework for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2280–2286, Sep. 2019.

[35] I. Vasiljevic, A. Chakrabarti, and G. Shakhnarovich, "Examining the impact of blur on recognition by convolutional networks," May 2017, *arXiv:1611.05760*.

[36] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," 2018, *arXiv:1811.12231*.

[37] L. Yahiaoui, J. Horgan, B. Deegan, S. Yogamani, C. Hughes, and P. Denny, "Overview and empirical analysis of ISP parameter tuning for visual perception in autonomous driving," *J. Imag.*, vol. 5, no. 10, p. 78, Sep. 2019. [Online]. Available: https://www.mdpi.com/2313-433X/5/10/78

[38] B. Sekachev, N. Manovich, M. Zhiltsov, A. Zhavoronkov, D. Kalinin, B. Hoff, D. Kruchinin, A. Zankevich, M. Markelov, M. Chenuet, A. Melnikov, J. Kim, L. Ilouz, N. Glazov, R. Tehrani, S. Jeong, V. Skubriev, S. Yonekura, and T. Truong, "OpenCV/CVAT: V1.1.0," Ultralytics, Los Angeles, CA, USA, Tech. Rep. 6.1, 2020, doi: 10.5281/zenodo.4009388.
[39] J. W. Goodman, *Introduction to Fourier Optics*, 4th ed. New York, NY, USA: MaCmillan Learning, 2017.
[40] J. G. Nagy and D. P. O'Leary, "Fast iterative image restoration with a spatially varying PSF," *Proc. SPIE*, vol. 3162, pp. 388-399, Oct. 1997. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.279513
[41] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2013, *arXiv:1311.2524*.
[42] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013. [Online]. Available: http://link.springer.com/10.1007/s11263-013-0620-5
[43] R. Girshick, "Fast R-CNN," 2015, *arXiv:1504.08083*.
[44] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.
[46] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016, *arXiv:1612.03144*.
[47] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," 2017, *arXiv:1712.00726*.
[48] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.
[49] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788.
[50] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 6517–6525. [Online]. Available: http://ieeexplore.ieee.org/document/8100173/
[51] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," Univ. Washington, Seattle, WA, USA, Tech. Rep. 1, 2018.
[52] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
[53] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, J. Fang, K. Michael, D. Montes, J. Nadar, P. Skalski, Z. Wang, A. Hogan, C. Fati, L. Mammana, D. Patel, D. Yiwei, F. You, J. Hajek, L. Diaconu, and M. T. Minh, "Ultralytics/YOLOV5: V6.1—TensorRT, TensorFlow edge TPU and OpenVINO export and inference," Tech. Rep., Feb. 2022, doi: 10.5281/zenodo.6222936.
[54] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.
[55] G. Jocher, A. Chaurasia, and J. Qiu. (2023). *YOLO by Ultralytics*. [Online]. Available: https://github.com/ultralytics/ultralytics
[56] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.
[57] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020, *arXiv:2005.12872*.
[58] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.
[59] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. (2012). *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.htm
[60] H. Nasse, "How to read MTF curves," Carl Zeiss AG, Oberkochen, Germany, Tech. Rep. 1, 2008.
[61] *Photography—Electronic Still Picture Imaging—Resolution and Spatial Frequency Responses*, Standard ISO12233:2017, International Organization for Standardization, Geneva, CH, USA, 2017.
[62] *Using ESFR ISO*. Accessed: Jan. 3, 2024. [Online]. Available: https://www.imatest.com/docs/esfriso_instructions/



DARA MOLLOY received the B.Eng. (Hons.) degree from the University of Galway, in 2018, where he is currently pursuing the Ph.D. degree. He is a member of the Connaught Automotive Research (CAR) Group under the supervision of Prof. Martin Glavin and Prof. Edward Jones. His research interest includes computer vision and sensor availability within an autonomous vehicle context.

PATRICK MÜLLER received the M.Sc. degree from Hochschule Düsseldorf, in 2018. He is currently pursuing the Ph.D. degree with the University of Siegen and Hochschule Düsseldorf. His research interests include the impact of different aspects of the point spread function (PSF) on the robustness of computer vision algorithms and space-variant evaluation.

BRIAN DEEGAN received the bachelor's degree in computer engineering and the M.Sc. degree in biomedical engineering from the University of Limerick, in 2004 and 2005, respectively, and the Ph.D. degree in biomedical engineering from the University of Galway, in 2011. The focus of the research was the relationship between blood pressure and cerebral blood flow in humans. From 2011 to 2022, he was with Valeo as a Vision Research Engineer, focusing on image quality. In 2022, he joined the Department of Electrical and Electronic Engineering, University of Galway, as a Lecturer and a Researcher. His research interests include high dynamic range imaging, LED flicker, topview harmonization algorithms, and the relationship between image quality and machine vision.

DARRAGH MULLINS received the B.E. degree in energy systems engineering and the Ph.D. degree in electronic engineering from the University of Galway, in 2013 and 2018, respectively. His Ph.D. research topic involved the application of imaging sensors and signal processing to wastewater treatment plant performance sensing. From 2018 to 2022, he was a Postdoctoral Research Fellow with the University of Galway, where he managed a research program and co-supervised six Ph.D. student projects involving sensors and V2X communication systems for pedestrian and vehicle monitoring from both vehicle and fixed infrastructure point-of-view. He is currently a Senior Technical Officer and an Adjunct Lecturer with the School of Engineering, University of Galway. He was awarded the Lero Director's Prize for Education and Public Engagement, in 2020.

JONATHAN HORGAN is currently the Computer Vision and Deep Learning Architecture Manager and a Senior Expert with Valeo. He has worked in the field of computer vision for more than 16 years, with a focus over the last ten years on automotive computer vision for advanced driver assistance systems (ADAS), automated parking, and automated driving. He is currently working on next-generation advanced computer vision and deep learning with the ultimate goal of achieving fully autonomous driving and parking. He has 25 publications in peer-reviewed conferences and journals and more than 100 patents published in the field of automotive computer vision.

ENDA WARD received the B.E. and master's (by Research) degrees in electronic engineering from the University of Galway, in 1999 and 2002, respectively, with a focus on biomedical electronics. He is responsible for defining the camera product roadmap for surround and automated driving applications within Valeo and has worked with key technology experts across the supply chain and within OEMs to define optimal system architectures. Lecturing for a number of years in the areas of electronics and computing systems with Atlantic Technological University, Ireland, later he moved into industry, working in the biomedical space. He has spent the last 16 years in automotive ADAS design. He holds several patents in the area of automotive vision.

EDWARD JONES (Senior Member, IEEE) received the B.E. and Ph.D. degrees in electronic engineering from the University of Galway, Ireland. His Ph.D. research topic was on the development of computational auditory models for speech processing. From 2009 to 2010, he was a Visiting Researcher with the Department of Electrical Engineering, Columbia University, New York, NY, USA; and a Visiting Fellow with the School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, Australia. He is currently a Professor of electrical and electronic engineering with the School of Engineering, University of Galway. He has several years of industrial experience in senior positions, in both start-ups and multinational companies, including Toucan Technology Ltd., PMC-Sierra Inc., Innovada Ltd., and Duolog Technologies Ltd. He also represented Toucan Technology and PMC-Sierra on international standardization groups ANSI T1E1.4 and ETSI TM6. His current research interests include DSP algorithm development and embedded implementation for applications in biomedical engineering, speech and audio processing, and image processing. He is a Chartered Engineer and a fellow of the Institution of Engineers of Ireland.

ALEXANDER BRAUN received the Diploma degree in physics from the University of Göttingen, in 2001, with a focus on laser fluorescent spectroscopy, the Ph.D. degree in quantum optics from the University of Hamburg, and the Postdoctoral degree from the University of Siegen, in 2007. He was an Optical Designer for camera-based ADAS with the company Kostal and became a Professor of physics with Hochschule Düsseldorf, in 2013, where he is currently researching optical metrology and optical models for simulation in the context of autonomous driving. He is a member of DPG, SPIE, and IS&T, participating in norming efforts at IEEE (P2020) and VDI (FA 8.13), and currently serves on the advisory board for the AutoSens Conference.

MARTIN GLAVIN (Member, IEEE) received the B.E. degree in electronic engineering and the Ph.D. degree in algorithms and architectures for high-speed data communications systems from the University of Galway, Ireland, in 1997 and 2004, respectively, and the Higher Diploma degree in third level education, in 2007. He was a lecturer (fixed term contract), from September 1999 to December 2003, and became a permanent member of the academic staff, in January 2004. He is currently the Joint Director of the Connaught Automotive Research (CAR) Group, University of Galway. He is also a Funded Investigator of Lero, the Irish Software Research Centre. He has a number of Ph.D. students and postdoctoral researchers in collaboration with industry in the areas of signal processing and embedded systems for automotive and agricultural applications.