applied sciences

MDPI

Article

Effect of Camera Choice on Image-Classification Inference

Jason Brown [1,*], Andy Nguyen [1] and Nawin Raj [2]

[1] School of Engineering, University of Southern Queensland, Brisbane 4300, Australia; andy.nguyen@unisq.edu.au

[2] School of Mathematics, Physics and Computing, University of Southern Queensland, Brisbane 4300, Australia; nawin.raj@unisq.edu.au

[*] Correspondence: jason.brown2@unisq.edu.au; Tel.: +61-7-3470-4026

Featured Application: The research presented in this paper has significant implications for many engineering applications that make use of image classification, including industrial inspection, medical diagnosis, and autonomous vehicle operation. This is because the specific camera used to capture an image can affect the top prediction of object class, particularly in scenarios with a more complex background.

Abstract: The field of image classification using Convolutional Neural Networks (CNNs) to predict the principal object in an image has seen many recent innovations. One aspect that has not been extensively explored is the effect of the camera employed to acquire images for inference. We investigate this by capturing comparable images of five drinking vessels using six cameras in various scenarios. We examine the classification ranking of object classes when these images are input to an independently pretrained Resnet-18 model based on the ImageNet-1k dataset. We find that the camera used can affect the top prediction of object class, particularly in scenarios with a more complex background. This is the case even when the cameras have similar fields of view. We also introduce a metric called selectivity, defined as the mean absolute difference between prediction probabilities of similar relevant object classes (such as cups and mugs). We show that the effect of the camera is largest when the selectivity of the pretrained model between these object classes is small. The effect of camera choice is also demonstrated quantitatively by examining Cohen's Kappa (κ) statistic. Finally, we make recommendations on mitigating the effect of the camera on image-classification inference.

Keywords: image classification; computer vision; inference; prediction; camera

Academic Editor: Andrea Prati

Received: 12 November 2024
Revised: 22 December 2024
Accepted: 27 December 2024
Published: 30 December 2024

Citation: Brown, J.; Nguyen, A.; Raj, N. Effect of Camera Choice on Image-Classification Inference. *Appl. Sci.* **2025**, *15*, 246. https://doi.org/10.3390/app15010246

1. Introduction

There has been much progress in the use of Convolutional Neural Networks (CNNs) as deep learning models for image classification and other computer vision techniques over the past few years [1–3]. Models such as AlexNet [4], VGG-16 [5], Resnet [6], and MobileNetV2 [7] have been progressively developed to increase the performance and efficiency of image classification.

The network models are typically trained and tested on large datasets of clean images. For example, the ImageNet-1k [8,9] database is composed of 1,281,167 training images, 50,000 validation images and 100,000 test images of 1000 distinct object classes such as furniture, drinking vessels, and various animals and birds.

The original metadata for the dataset images, and in particular the source camera and settings (such as exposure time and white balance), are usually unknown or at least unpublished. This is potentially significant because no two camera types are the same; they have unique characteristics (e.g., field of view, supported resolutions) and unique imperfections (e.g., lens distortion, chromatic aberration) [10] which filter into the trained image classifier. Therefore, there is no guarantee that the datasets are representative of the full range of camera types available, currently or in the future. This raises the question of how significantly the choice of camera type affects the deployment of trained image-classification models in the field. More specifically, is it possible or even likely that the same trained image classifier will give different predictions when using two different camera types? This is the primary objective of investigation in this paper, i.e., to determine whether and to what extent the camera employed to capture an image affects the image-classification output. The answer to this question has very significant implications for many fields, including industrial inspection [11–13], medical diagnosis [14,15], and autonomous vehicle operations [16,17]. For example, consider the consequences of a hypothetical situation where two different camera types are used to acquire comparable images for the diagnosis of a serious human health condition, and the same pretrained CNN model classifies images from one camera as positive for the health condition, but classifies images from the second camera as negative for the health condition. Compared to the analogous topic of examining the robustness of speech recognition in the face of multiple spoken accents [18,19], the issue of characterizing the robustness of image classification to multiple camera types, each with its own properties, has received relatively little attention.

As illustrated in Figure 1, a typical digital camera comprises a lens, image sensor (with Bayer filter), and Image Signal Processor (ISP) [20]. The lens focuses the image on the image sensor, the sensor forms a raw pixelated image, and the ISP provides various processing functions such as de-mosaicing, noise reduction, and compression.

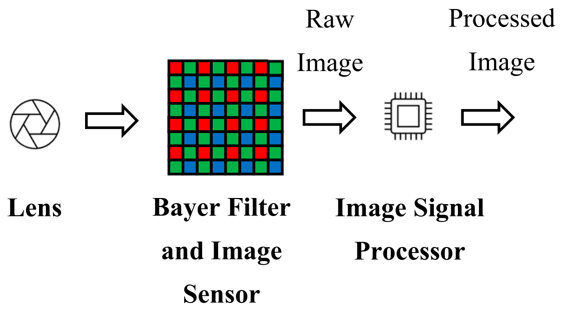


Figure 1. Typical components and data flows of a digital camera.

Each of these main components of the camera system may have unique properties and imperfections that affect the processed image output. Although one purpose of the ISP is to correct for some of the imperfections of the earlier components, such as lens distortion, the exact capabilities and performance of the ISP will still vary from one camera type to the next.

Furthermore, there are properties of a digital camera that depend upon multiple components. One of the most important of these is the Angle of View (AOV) [20], sometimes known as the Field of View (FOV). The AOV describes the angle of observability of the camera and depends both upon the lens and the physical size of the image sensor. A camera with a larger AOV will capture more content than a camera with a smaller AOV. The consequence of this is that the same target object captured by a camera with a larger AOV will appear smaller in the image than when captured by a camera with a smaller AOV, assuming all other parameters (e.g., distance to the object) are kept constant. An equivalent way of expressing this is to state that the AOV affects the perceived size of objects within images. These concepts are illustrated in Figure 2.

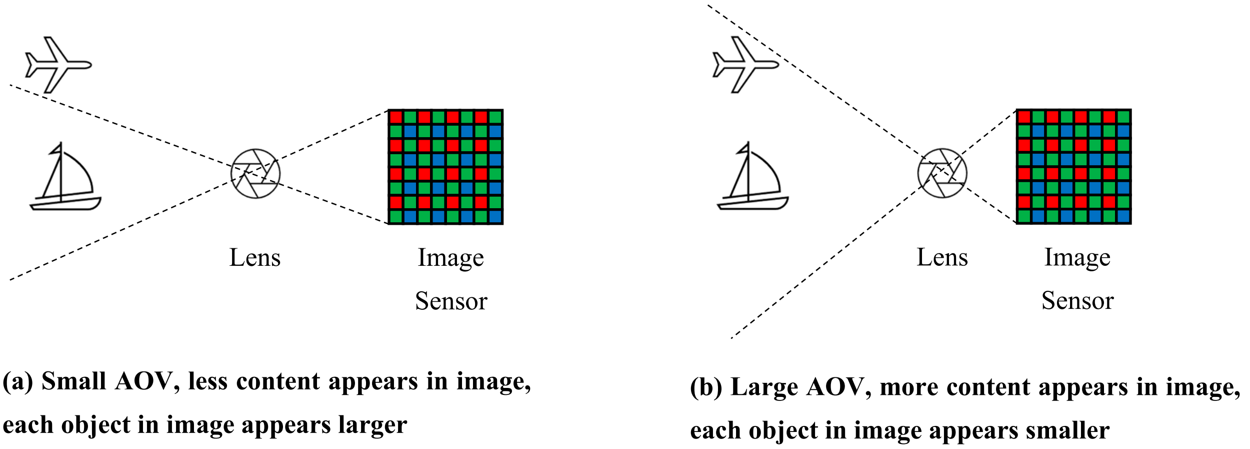

Lens
Image Sensor
(a) Small AOV, less content appears in image, each object in image appears larger
Lens
Image Sensor
(b) Large AOV, more content appears in image, each object in image appears smaller

Figure 2. Simplified demonstration of the effect of camera Angle of View (AOV) on image capture (assumes pinhole lens).

In this paper, we investigate the effect of the specific camera used to capture an image on classification inference. To do this, we capture images of certain objects that can be recognized by an established, independently pretrained image-classification model using different cameras while keeping the environmental conditions (e.g., lighting) as constant as possible. We then feed these images into the pretrained image-classification model to observe any differences in the classification output. It is important to note that the main target object in the images is easily recognizable by a human observer; we do not attempt to intentionally confuse the pretrained model in any way. In particular, the lighting is set to facilitate easy identification of the main target object in the images by a human observer. We expect the classification to depend on the camera AOV because this impacts how dominant the target object is in the captured image and how many other peripheral objects are captured in the image. Nevertheless, for cameras with similar or even identical AOV, it is of great interest to explore whether comparable images taken with these cameras can result in different classification outputs when using the same specific pretrained model. Note that we do not perform any model training in this paper; rather, the images we capture are used for inference purposes only in concert with a pretrained model.

To our knowledge, there is no similar publicly available study on the effect of the specific camera used to capture an image on classification inference, despite the potentially significant implications for many fields, including industrial inspection. Although this study was exploratory, and there was always the possibility from the outset that the actual observed differences between cameras would transpire not to be particularly significant, we were, in fact, able to demonstrate that comparable images taken with different camera types (even those with a similar AOV) can result in different classification outputs when using the same specific pretrained CNN model.

The remainder of this paper is organized as follows. In Section 2, we discuss related work, particularly the recently growing field of investigating the robustness of computer vision to natural image degradations, of which differences in camera types are an important yet neglected area. Section 3 addresses the materials and methods we adopted for acquiring comparable images from different camera types and classifying them according to a specific pretrained CNN model. Results from these experiments are depicted graphically in Section 4 and analyzed according to their significance using Cohen's Kappa (κ) statistic. We demonstrate that comparable images taken with different camera types (even those with a similar AOV) can result in different classification outputs when using the same specific pretrained CNN model. Conclusions and directions for further research are discussed in Section 5.

2. Related Work

A large-scale investigation into the effect of image quality on image classification was conducted in [21]. The authors used a subset of the validation images from the ImageNet-1k database and digitally introduced varying levels of quality distortions, including blur, noise, contrast, and compression. They then classified the manipulated images using pretrained Caffe reference [22], VGG-CNN-S [23], VGG-16 [5], and GoogleNet [24] models, with the model weights taken from the Caffe library. The image-classification performance was found to be most sensitive to the quality distortions of blur and noise for all models. The authors noted that this extended range of quality distortions may be found in adversarial images [25], which are images that are purposely manipulated to fool an image classifier but are unlikely to be found in non-manipulated images. However, the normal image acquisition process can introduce both blur (e.g., when an image is taken out of focus) and noise (e.g., due to a low-quality image sensor). This motivates the current study discussed in this paper, where we attempt to find differences in the image-classification output from images of the same item captured by different cameras under the same environmental conditions.

The research in [26] aimed at training robust image classifiers to handle failures in the image acquisition process. These failures were classified as internal (e.g., black pixels due to a poor or failed image sensor), external (e.g., scratched lens), or environmental (e.g., ice, rain, or condensation on the lens) [27,28]. The authors augmented three well-known Traffic Signal Recognition (TSR) datasets with additional images corresponding to a total of 13 specific visual camera failures, then trained and tested AlexNet [4], MobileNetV2 [7], and Inceptionv3 [29] classifiers based on the augmented dataset. The results demonstrated significantly improved classification accuracy using this approach. Our study is somewhat different in scope in that we are examining the effect of the camera type employed to acquire an image on the image classifier output based upon the fact that the different cameras have different characteristics. We do not consider visual camera failures per se, and we do not train models (rather, we use a pretrained classifier corresponding to the ImageNet-1k database). Nevertheless, it may be that retraining of classifiers based upon the data augmentation proposed in [26] may help to mitigate differences in image-classification output when different cameras are used to capture images. However, this is a topic for further research.

The study discussed in [30] confirmed that degradations in image quality have a significant impact on image-classification performance. Building on some previous research [31–33], the authors examined nine specific degradations, including hazy images, motion blur, out-of-focus blur, underwater images, fish-eye camera images, very low-resolution images, and salt-and-pepper noise using AlexNet [4], VGG-16 [5], and Resnet [6] image classifiers. The salient aspect of this research was that an attempt was made to remove the degradations using accepted image processing techniques to restore the clean images prior to image classification. However, it was discovered this only marginally mitigated the drop in image-classification performance.

A recent systematic review of research into the robustness of computer vision to natural image degradations rather than intentional adversarial image degradations is provided in [10]. This highlights that natural variations in image quality (for example, as produced by different camera types) have received relatively little interest compared to adversarial perturbations. This is the case even though studies [28,34–37] show naturally degraded images can result in a 30–40% decrease in image-classification accuracy, which is very significant when computer vision is integrated into safety critical systems such as autonomous vehicles. The review paper [38] summarizes state-of-the-art research for analyzing and mitigating the environmental and camera effects on IoT images.

There has also been some work on the simulation of cameras to enable the generation of images with realistic camera degradations. The study in [39] reported that some imperfections, such as distortion, chromatic aberration, and vignetting, could be simulated relatively successfully, whereas lens flare was too specific to individual camera types to be simulated realistically. Other studies with a similar aim include [40] for airborne computer vision and [41] which uses a neural network to generate images with programmable exposure time, light sensitivity level, and aperture size. The research in [42] introduces the Image Systems Evaluation Toolkit (ISET), which comprises software routines to simulate the capture and processing of visual scenes. In contrast, our study examines the effect of image-classification performance on real images taken with a variety of camera types.

3. Materials and Methods

3.1. Camera Selection

The cameras employed in this study are listed in Table 1 with their important properties. We focused exclusively on webcams firstly because they are small and, therefore, easy to position accurately. Secondly, they can be controlled from a common software application (in this study, the Microsoft Windows Camera application), so there is a consistent means of capturing and managing images from the different cameras. There is significant cross-support for certain resolutions among these webcams, which facilitates a valid comparison between them. In addition, we selected webcams from different vendors to increase the diversity of features and implementations. Note that some properties, such as the image sensor size and the aperture f-value, are not included in Table 1 because not all webcam vendors publish these values in their data sheets.

Table 1. Cameras and Their Properties (All cameras sourced in Australia).

Property	**Angetube 962A**	**EMEET SmartCam C960**	**Logitech C270**	**Logitech C920e**	**Microsoft LifeCam Cinema**	**Razer Kiyo X**
Field of View [a]	78°	90°	55°	78°	73°	82°
Focus Method	Autofocus	Fixed Focus	Fixed Focus	Autofocus	Autofocus	Autofocus
Light Correction	Yes	Yes	No	Yes	No	No
Number of supported resolutions [b]	13	7	16	13	9	4
Supported resolutions [b] (16:9 resolutions in green, 4:3 resolutions in blue, other aspect ratios in black)	1920 × 1080	1920 × 1080		1920 × 1080		1920 × 1080
	1600 × 896			1600 × 896		
	1280 × 720	1280 × 720	1280 × 720	1280 × 720	1280 × 720	1280 × 720
	1024 × 576	1024 × 576	1024 × 576	1024 × 576		
	800 × 448		800 × 448	800 × 448	800 × 448	
	640 × 360	640 × 360	640 × 360	640 × 360	640 × 360	640 × 360
		1280 × 960	1280 × 960			
	1024 × 768					
			960 × 720	960 × 720		
		800 × 600	800 × 600	800 × 600	800 × 600	
	640 × 480	640 × 480	640 × 480	640 × 480	640 × 480	640 × 480
	320 × 240		320 × 240	320 × 240	320 × 240	
	352 × 288		352 × 288	352 × 288	352 × 288	
	960 × 544		960 × 544		960 × 544	
			1184 × 656			
	848 × 480					
	864 × 480		864 × 480	864 × 480		
			752 × 416			
			544 × 288			
			432 × 240	432 × 240		
					424 × 240	

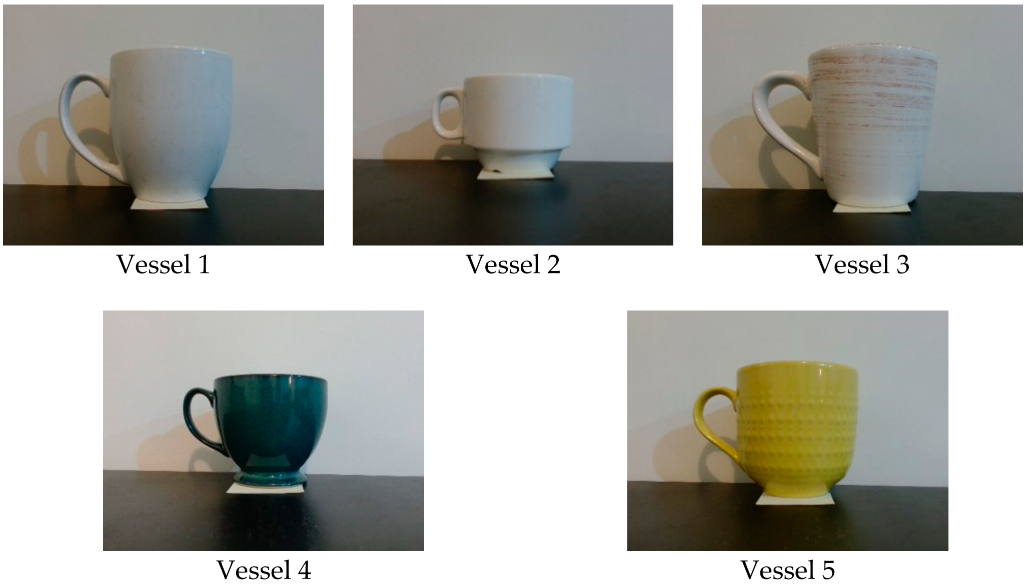
[a] As reported by the manufacturer in the datasheet or technical specifications of the camera. [b] As reported by the camera to the Microsoft Camera application; resolutions smaller than 224 × 224 have been omitted since they were not used in data acquisition.

3.2. Image Targets

In this study, we employed a pretrained off-the-shelf image-classification model (to be described later), for which the training dataset was the well-known ImageNet-1k dataset [8,9], which comprises 1000 classes. Therefore, the images captured for inference purposes were required to contain objects that are represented in these 1000 classes. We chose the classes of "cups" and "coffee mugs" for this study for the following reasons:

- They are inanimate and, therefore, can easily be photographed under controlled conditions.
- They are ubiquitous and exist in many different forms.
- They are very similar to each other. Therefore, there is the possibility of the model inferring an image of a particular drinking vessel taken with one camera is most likely a "cup", while inferring an image of the same drinking vessel taken under the same environmental conditions with a different camera is most likely a "coffee mug". This would clearly show the impact of the specific camera used to capture the image on the inference.

The five drinking vessels acting as image targets are depicted in Figure 3. These vessels differ in several areas: size, shape, color, surface texture, and reflectivity. There is no need to declare whether the ground truth of each vessel is a cup or a coffee mug, as this would, in fact, be subjective. Instead, we are interested in how these image targets might be perceived differently by the model when images of them are taken by different cameras under the same environmental conditions. For this reason, we refer to these objects as "vessels" in the remainder of this paper.



Vessel 1

Vessel 2

Vessel 3

Vessel 4

Vessel 5

Figure 3. Drinking vessels used in the study (camera: Logitech C920e, distance: 20 cm, scenario: Scenario 1, image resolution: 320 × 240).

It should be noted that the ImageNet-1k dataset comprises some other classes which are reasonably like cups and coffee mugs, including coffeepots, teapots, and pitchers, and sometimes we might expect the vessels to be classified by the model as these other classes.

3.3. Image Capture

As illustrated in Figure 4, images were captured at distances of 20 cm, 30 cm, 40 cm, 50 cm, and 60 cm between the camera lens and the frontmost point of the vessel. This set of distances was chosen because images are often blurred for a distance less than 20 cm, and the image-classification model employed will usually fail to recognize the target vessel correctly at a distance greater than 60 cm, at least for scenarios where there are other background objects of interest in the image.

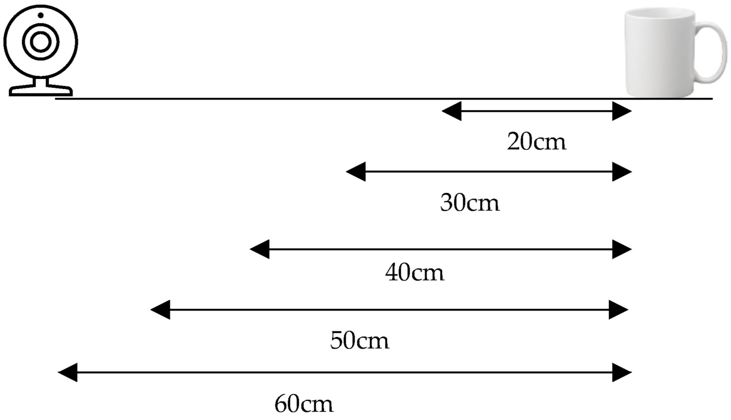


Figure 4. Distance values used for image capture.

Images of Vessel 1 taken at different distances are illustrated in Figure 5. At 60 cm, Vessel 1, which is a relatively large drinking vessel, only occupies a relatively small section of the image space, although it is still clearly identifiable as a drinking vessel by a human observer.

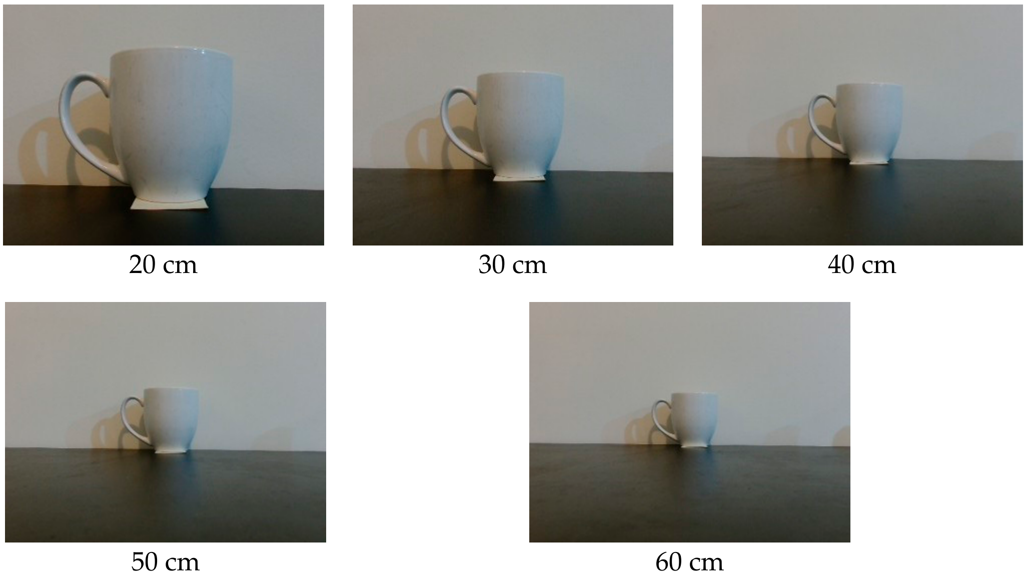


Figure 5. Images of Vessel 1 at different distances (camera: Logitech C920e, scenario: Scenario 1, image resolution: 320 × 240).

For each image taken, the camera was positioned such that the target vessel was in the center of the field of view with the plane of the handle perpendicular to the notional line between the camera and vessel. This ensured that the target vessel was clearly identifiable by a human observer as a drinking vessel and as the principal object in the image. We did not make any changes to the configuration settings of the cameras (apart from exercising different capture resolutions) and relied on default out-of-box settings. When capturing an image, the camera autofocus mechanism, if supported by the specific camera, was allowed to settle before the image was taken, thus ensuring crisp, non-blurred images.

For each camera, images were captured for each combination of the following:

- The supported resolutions of the camera (see Table 1);
- The five target vessels;

- The five distances of 20 cm, 30 cm, 40 cm, 50 cm, and 60 cm;
- Three scenarios, where each scenario comprises a different background to the target vessel.

Therefore, with reference to the resolutions supported by each camera in Table 1, the complete dataset to be used for image-classification inference comprised (13 + 7 + 16 + 13 + 9 + 4) × 5 × 5 × 3 = 4650 images.

The three different scenarios/backgrounds are illustrated in Figure 6. The use of different scenarios facilitated a larger overall dataset for analysis and also allowed an investigation into whether the image backgrounds affect the image classification to a significant degree. Scenario 1 has the quietest background, and Scenario 3 has the busiest background, but for all scenarios, it is clear to a human observer that the foreground drinking vessel is the primary object of interest in the image.

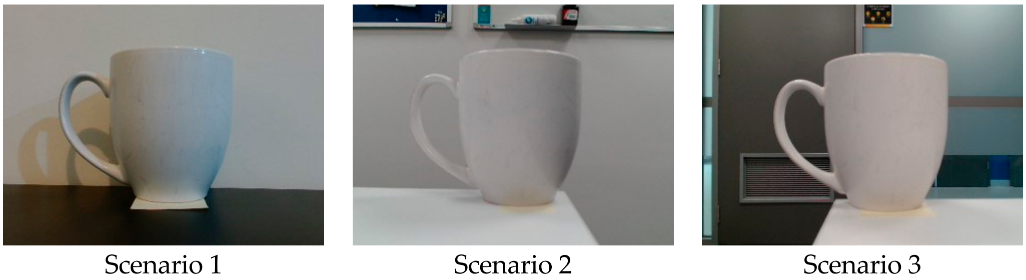


Figure 6. Images of Vessel 1 in different scenarios (camera: Logitech C920e, distance: 20 cm, image resolution: 320 × 240).

The three scenarios are all indoors since it was easier to maintain consistent environmental conditions, particularly with respect to lighting, while capturing the complete image dataset. However, in order to guard against any residual changes to environmental conditions and the consequent impact on assessing the effect of the camera type on image-classification inference, the following methodology was adopted:

- For each scenario, vessel, and distance combination, a camera was used to capture images of the vessel for all supported resolutions and then swapped out for another camera.
- For each scenario, vessel, and distance combination, the order in which cameras were used was randomized.

This frequent cycling of cameras to guard against residual changes to environmental conditions is illustrated in Figure 7.

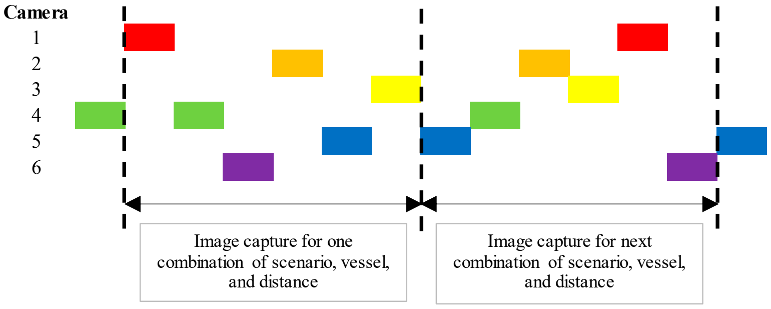

Camera
1
2
3
4
5
6
Image capture for one combination of scenario, vessel, and distance
Image capture for next combination of scenario, vessel, and distance

Figure 7. Image-capture methodology involving frequent cycling of cameras and randomized ordering of camera use.

The complete image-capture process is illustrated in the flow chart of Figure 8.

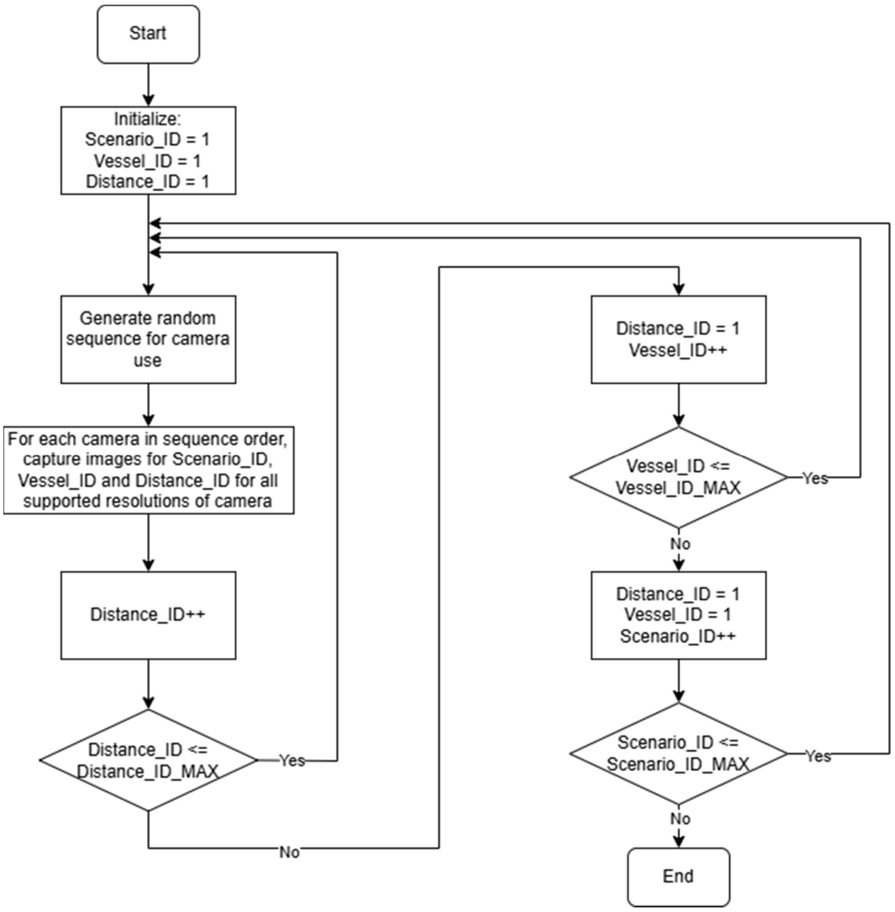

Start
Initialize:
Scenario_ID = 1
Vessel_ID = 1
Distance_ID = 1
Generate random sequence for camera use
For each camera in sequence order, capture images for Scenario_ID, Vessel_ID and Distance_ID for all supported resolutions of camera
Distance_ID++
Distance_ID <= Distance_ID_MAX
Yes
No
Distance_ID = 1
Vessel_ID++
Vessel_ID <= Vessel_ID_MAX
Yes
No
Distance_ID = 1
Vessel_ID = 1
Scenario_ID++
Scenario_ID <= Scenario_ID_MAX
Yes
No
End

Figure 8. Flow chart for overall image-capture process.

3.4. Model Inference

The 4650 captured images were used as input to a Resnet-18 [6] image-classification model, which had been independently trained on the ImageNet-1k [8,9] database. The specific pretrained Resnet-18 model used was supplied with the PyTorch framework [43] and employed the ResNet18_Weights.IMAGENET1K_V1 weights. This model can provide, for each image, a prediction probability for each of the 1000 object classes represented in the ImageNet-1k database. For this study, we are primarily interested in the object class with the highest prediction probability and, in some contexts, the object classes with the five highest prediction probabilities.

A Resnet-18 model was selected for this analysis because it is a modern, lightweight, high-performance, and efficient image-classification CNN model that can be used in a variety of image-classification applications. The efficiency of the model was of interest because the overall time required to make inferences on 4650 images can be quite large. Although other image-classification models are of interest for future research, it transpired that using Resnet-18 alone was sufficient to meet the objectives of this research with respect to showing how comparable images of the same object taken using different cameras can lead to differences in the prediction of the most likely object class when using the same pretrained image-classification model.

We now provide some more details on Resnet-18. The input image is reduced to a resolution of 224 × 224 before being processed by the model. The architecture consists of 18 convolutional and/or fully connected layers, as shown in Figure 9. A typical convolution layer has a number of parameters of interest. For example, considering the convolutional layer with the label “3 × 3, Conv, 256,/2”, this comprises 256 filters with a window size of

3 × 3 and a stride of 2. The curved arrows are skip connections, which reduce the chance of overfitting the model to the data during training. For further information, the reader is referred to [6].

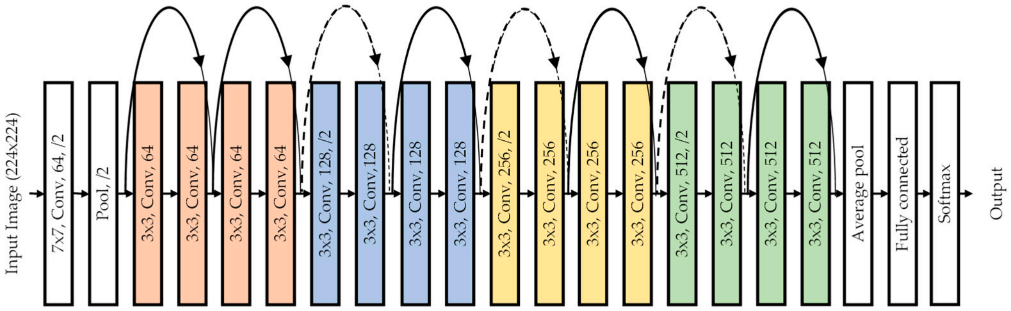
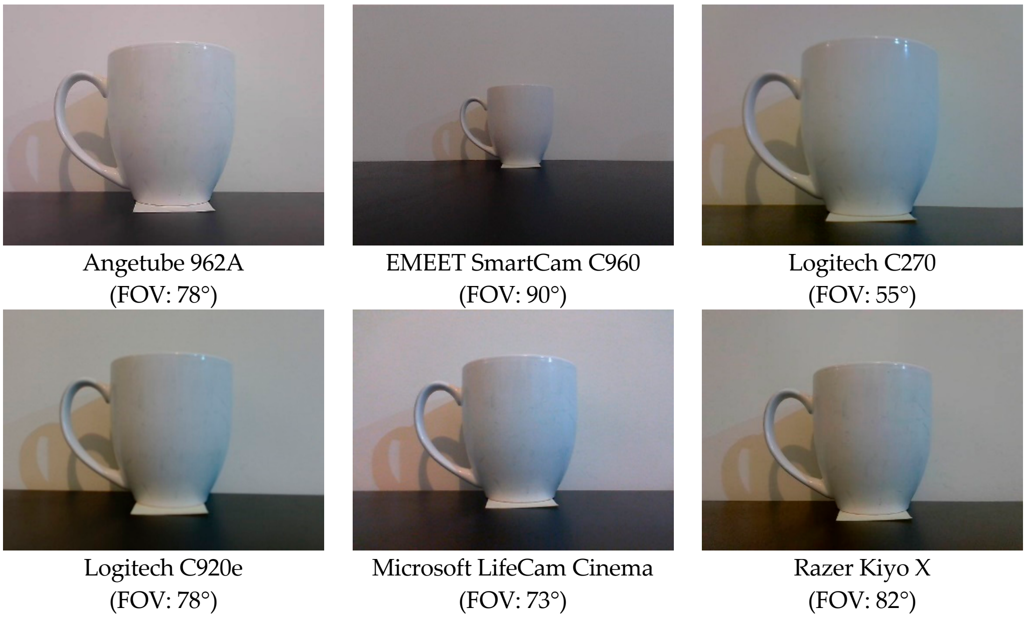


Figure 9. Structure of the Resnet-18 model.

4. Results and Discussion

4.1. Impact of FOV

As discussed in the Introduction to the paper, the FOV of the camera affects the perceived size of objects within images. Therefore, the interpretation of the results of the model inference must take FOV into account. To set the scene for this, Figure 10 illustrates images of Vessel 1 taken in the same scenario and at the same distance by the different cameras employed in the study.


Angetube 962A
(FOV: 78°)

EMEET SmartCam C960
(FOV: 90°)

Logitech C270
(FOV: 55°)

Logitech C920e
(FOV: 78°)

Microsoft LifeCam Cinema
(FOV: 73°)

Razer Kiyo X
(FOV: 82°)

Figure 10. Images of Vessel 1 for different cameras (scenario: Scenario 1, distance: 20 cm, image resolution: 640 × 480).

The apparent size of Vessel 1 in the image is very similar for the Angetube 962A, Logitech C920e, Microsoft LifeCam Cinema, and Razer Kiyo X, on account of their similar FOV values. However, it can be seen there are other differences in these images, e.g., differences in the rendered colors. At the extremes, the apparent size of Vessel 1 in the image produced by the Logitech C270 is significantly larger due to the smaller FOV of the camera, and the apparent size of Vessel 1 in the image produced by the EMEET SmartCam C960 is significantly smaller due to the larger FOV of the camera.

4.2. Scenario 1

We begin the results with Scenario 1 because it involves a very plain background, so there is less chance that an image-classification model might be confused between the foreground object (i.e., the drinking vessel) and background objects. As discussed previously, we do not declare whether the ground truth of each vessel is a cup or a coffee mug, as this would be subjective. Instead, we are interested in whether the pretrained Resnet-18 model classifies each image as a cup, coffee mug, or some other class as a function of the camera used to acquire the image.

Figures 11 and 12 illustrate stacked bar charts showing the proportion of images classified as a cup or a coffee mug by the pretrained Resnet-18 model as a function of distance for different combinations of drinking vessel and camera. In Figure 11, only the top prediction class of an image is considered; in Figure 12, the top five prediction classes are considered such that if a cup and/or coffee mug appears in the top five, the highest ranking of these two classes is selected.

It is clear from Figure 11 that the proportion of images classified as a cup or a coffee mug decreases with distance, especially for Vessel 2. Although this is to be expected from the perspective that it is harder to recognize an object with increasing distance, it is important to note that a human observer would be able to easily recognize vessels in all the images taken for this study. We also see that the proportion of images classified as a cup or a coffee mug generally decreases with distance more quickly for the EMEET SmartCam C960 due to its relatively large FOV and less quickly for the Logitech C270 due to its relatively small FOV. On the contrary, the proportion of images classified as a cup or a coffee mug for the Logitech C270 is quite low for the larger vessels (i.e., Vessel 1 and Vessel 3) at a distance of 20 cm; this is because of the relatively small FOV of this camera, which increases the apparent size of objects, sometimes making them blurred and only just able to fit in the frame at small distances.

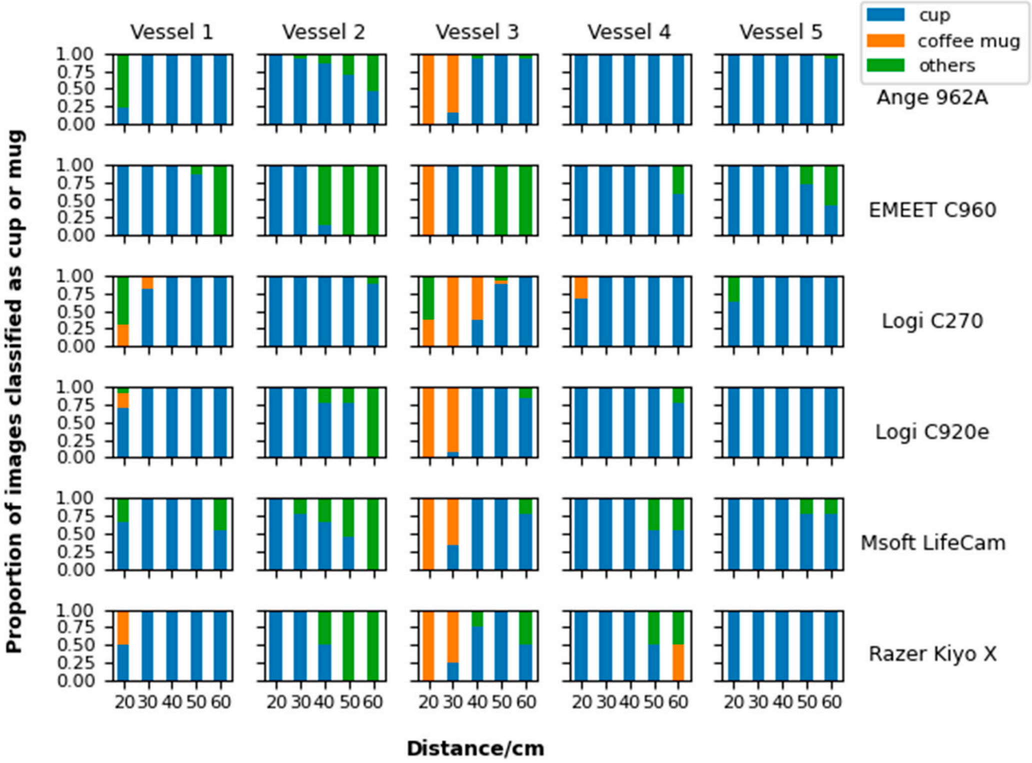


Figure 11. Proportion of images classified as a cup or a coffee mug in the top prediction as a function of distance for different drinking vessels and camera combinations (scenario: Scenario 1).

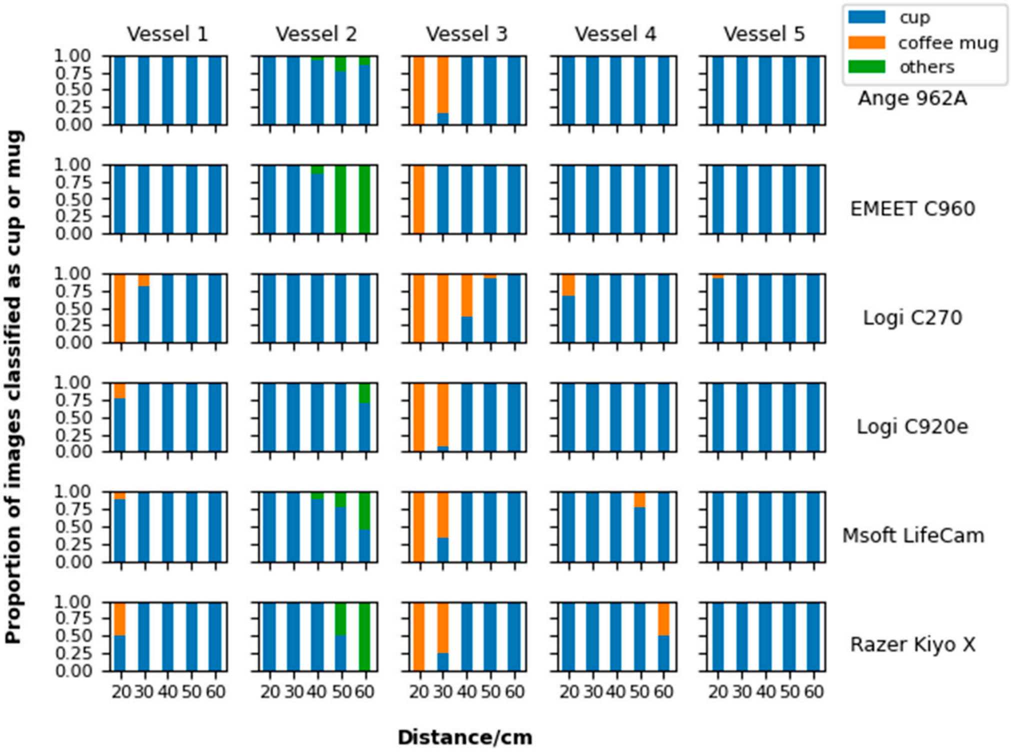


Figure 12. Proportion of images classified as a cup or a coffee mug in the top five predictions as a function of distance for different drinking vessels and camera combinations (scenario: Scenario 1).

The bar charts for Vessel 3 (and, to a lesser extent, Vessel 1) are interesting from the perspective that the pretrained Resnet-18 model tends to classify this vessel as a coffee mug at small distances but then transitions to classifying it as a cup at larger distances. We refer to this effect as transitioning in the remainder of this paper. One possible explanation is that at larger distances, the increasing pixelation of the image section that contains the vessel leads to a difference in classification for this vessel's shape, texture, and/or color. Another possible explanation is unintentional bias in the training data labeling within ImageNet, i.e., training images with drinking vessels in the distance may appear more like cups than mugs to a casual observer due to their small apparent size.

For Scenario 1, the classification output of the pretrained Resnet-18 model does not appear to depend significantly on the specific camera used to capture an image, apart from with respect to the expected differences due to the different FOVs of the cameras. In particular, the pretrained Resnet-18 model seems to classify each vessel relatively consistently as a cup or a coffee mug (for a specific distance), irrespective of the camera. This is not entirely surprising because Scenario 1 is very simple with an extremely plain background.

However, looking in greater detail at Figures 11 and 12, the proportion of images classified as a cup or a coffee mug decreases with distance more quickly for the Microsoft Lifecam than for the Angetube962 and Logitech C920e, especially for Vessel 2, even though the three cameras have a very similar FOV. To gain more insight into this observation, we define a metric to quantify the ability of the pretrained Resnet-18 model to distinguish between the two similar classes of cup and coffee mug.

The selectivity σ is defined formally in Equation (1) and is the mean absolute difference between the prediction probabilities for the cup and coffee mug classes.

$$\sigma(s,v,c,d) = \frac{1}{N}\sum_{j=1}^{N}\left|P_{cup}(s,v,c,d,j) - P_{mug}(s,v,c,d,j)\right| \tag{1}$$

where

- P_{cup} is the prediction probability for the cup class provided by the pretrained Resnet-18 model for image j corresponding to the tuple (s, v, c, d);
- P_{mug} is the prediction probability for the coffee mug class provided by the pretrained Resnet-18 model for image j corresponding to the tuple (s, v, c, d);
- s is the scenario which defines the image background;
- v is the vessel that is the image target;
- c is the camera acquiring the image;
- d is the distance between camera and vessel;
- N is the number of images in the dataset corresponding to the tuple (s, v, c, d).

Figure 13 illustrates bar charts showing the selectivity of the pretrained Resnet-18 model as a function of distance for different combinations of drinking vessel and camera. To examine the differences in selectivity specifically between the Angetube962, Logitech C920e, and Microsoft Lifecam as three cameras with similar FOVs, we reformat the graphs of Figure 13 as grouped bar charts in Figure 14. It is clear from Figure 14 that the selectivity of the image classifier to distinguish between a cup and a coffee mug is usually higher (and usually significantly higher) for images acquired by the Angetube 962A than the Microsoft LifeCam. This demonstrates a significant dependence on the performance of the image classifier in Scenario 1 on the camera used to acquire images. This dependence of selectivity on the camera does not, for this scenario, translate into a corresponding dependence of the classification output in terms of the top one and top five predictions (illustrated in Figures 11 and 12, respectively) on the camera; this is primarily because, although the selectivity values are significantly different for the different cameras, they are usually quite high, which means the image classifier is easily able to distinguish between a cup and coffee mug for this scenario.

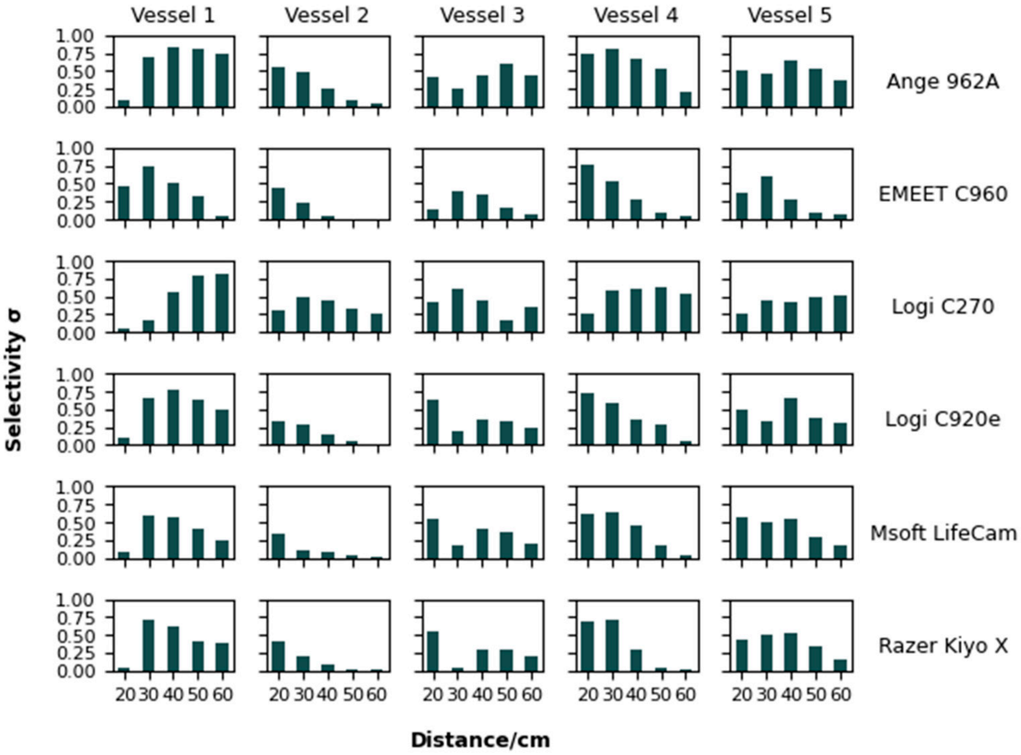


Figure 13. Selectivity σ (i.e., mean absolute difference between prediction probabilities for cup and mug classes) as a function of distance for different drinking vessel and camera combinations (scenario: Scenario 1).

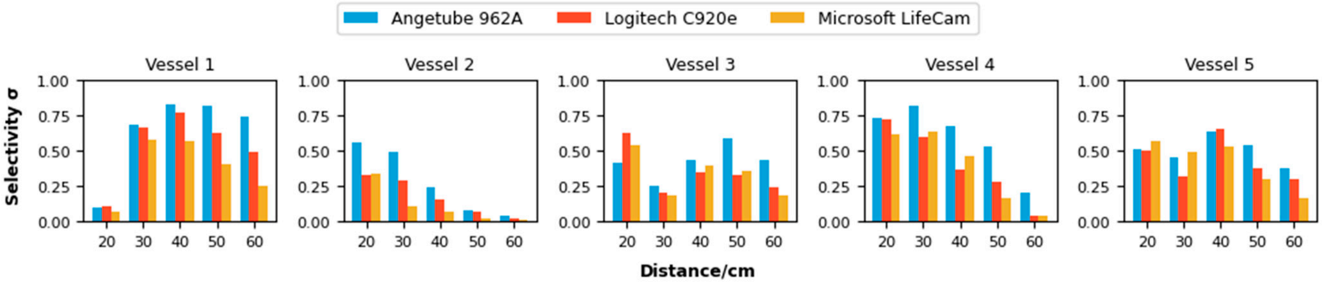


Figure 14. Selectivity σ (i.e., the mean absolute difference between prediction probabilities for cup and mug classes) comparison for cameras with similar FOV (scenario: Scenario 1).

Some other interesting observations can be made about Figure 14. The selectivity is very low for Vessel 1 at 20 cm for all three cameras. This may reflect the fact that Vessel 1 is relatively large and, therefore, occupies a large section of the image space at this distance. It can be seen from Figure 13 that the selectivity for Vessel 1 at 20 cm is significantly higher when acquired with the EMEET SmartCam C960, for which the larger FOV makes the apparent size of the vessel smaller. Separately, for Vessel 3, the selectivity is minimum for all three cameras at the intermediate distance of 30 cm; this corresponds to the distance at which the image classifier predictions are transitioning from "coffee mug" to "cup" for these three cameras in Figure 11.

4.3. Scenario 2

Scenario 2 involves a more complex background than Scenario 1, in which there are a small number of less prominent objects (such as a board eraser, as illustrated in Figure 6) in addition to the foreground drinking vessel. Figures 15 and 16 illustrate stacked bar charts showing the proportion of images classified as a cup or a coffee mug by the pretrained Resnet-18 model as a function of distance for different combinations of drinking vessel and camera. In Figure 15, only the top prediction class of an image is considered; in Figure 16, the top five prediction classes are considered such that if a cup and/or coffee mug appears in the top five, the highest ranking of these two classes is selected. The pretrained Resnet-18 model is much more likely to classify the drinking vessels as a coffee mug in this scenario than in the previous scenario.

The following subplots of Figures 15 and 16 are shaded with a gray face color to designate that they are of interest with respect to demonstrating a dependence of the classification on the camera:

- Microsoft LifeCam and Vessel 1: the image classifier shows a propensity for classification of the vessel as a coffee mug, whereas with other cameras, the same classifier is more likely to select a cup.
- EMEET SmartCam C960 and Vessel 2: the image classifier shows a propensity for classification of the vessel as a coffee mug, whereas with other cameras, the same classifier is much more likely to select a cup.
- Angetube 962A and Vessel 4: the image classifier shows a propensity for classification of the vessel as a cup, whereas with other cameras (except for the Logitech C270), the same classifier is more likely to select a coffee mug.

It is of interest to determine whether these observed differences in the image-classification output are related to the selectivity metric. Figure 17 compares the selectivity of the pretrained Resnet-18 model as a function of distance for the Angetube962, Logitech C920e, and Microsoft Lifecam. It is immediately apparent that all the selectivity values are very small for Vessel 1 and Vessel 4 in Scenario 2, which was not the case in Scenario 1. This may explain the differences in classification by the image classifier highlighted above for these vessels. Specifically, the low selectivity means the image classifier is not easily able to distinguish between a cup and a coffee mug, so any small difference in the images produced by different cameras may be sufficient to change the top prediction.

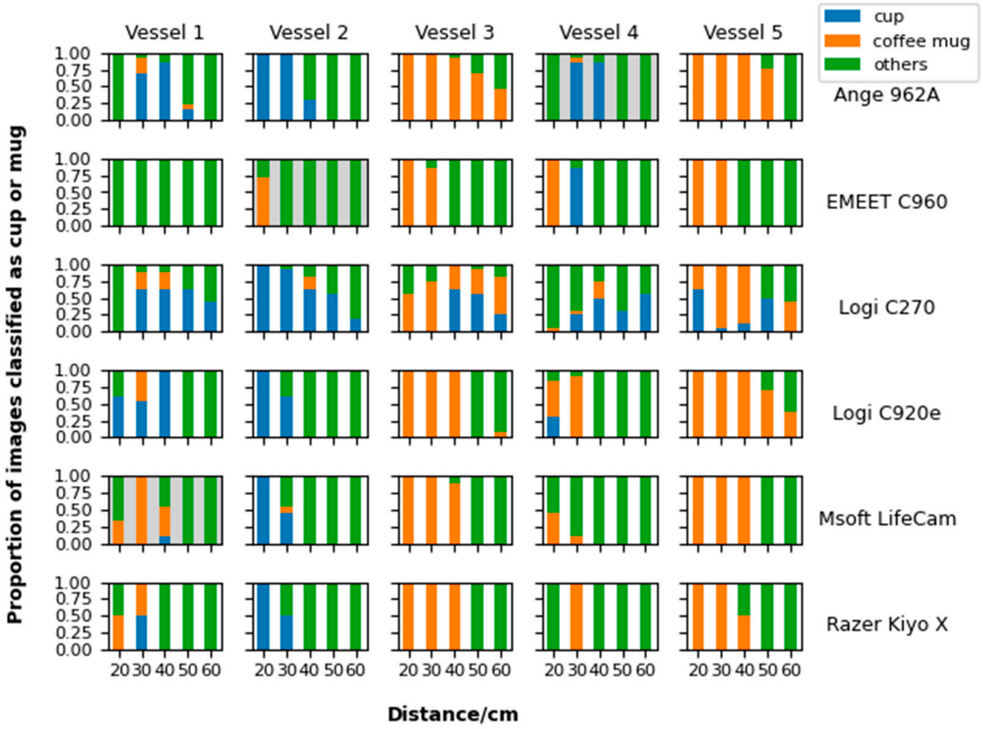


Figure 15. Proportion of images classified as a cup or a coffee mug in the top prediction as a function of distance for different drinking vessel and camera combinations (scenario: Scenario 2)—subplots of interest highlighted with a gray face color.

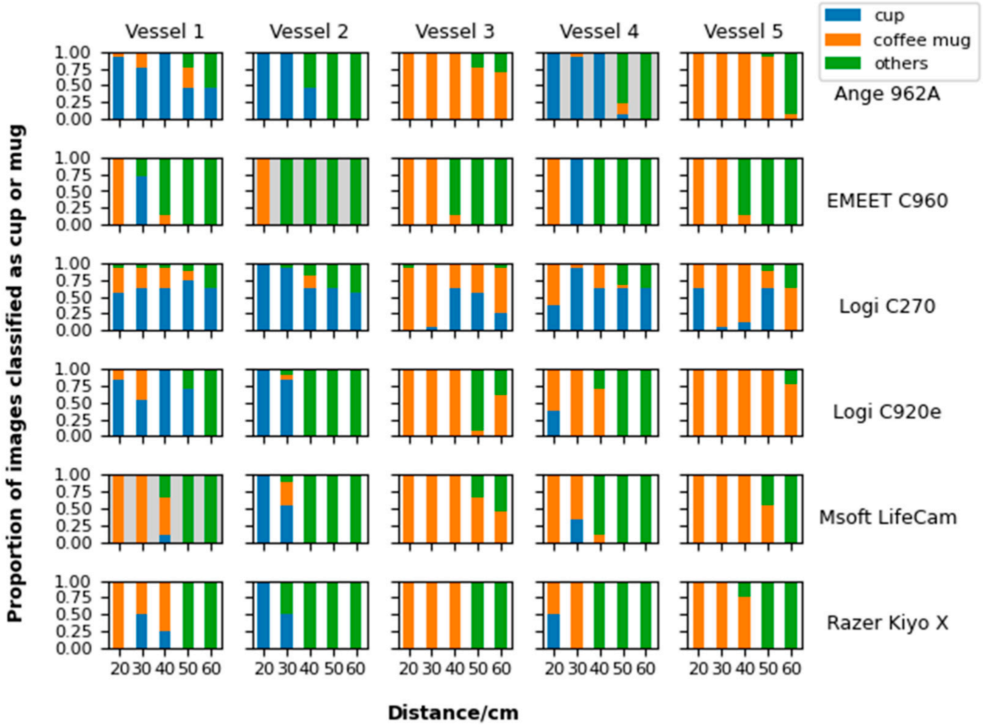


Figure 16. Proportion of images classified as a cup or a coffee mug in the top five predictions as a function of distance for different drinking vessel and camera combinations (scenario: Scenario 2)—subplots of interest highlighted with a gray face color.

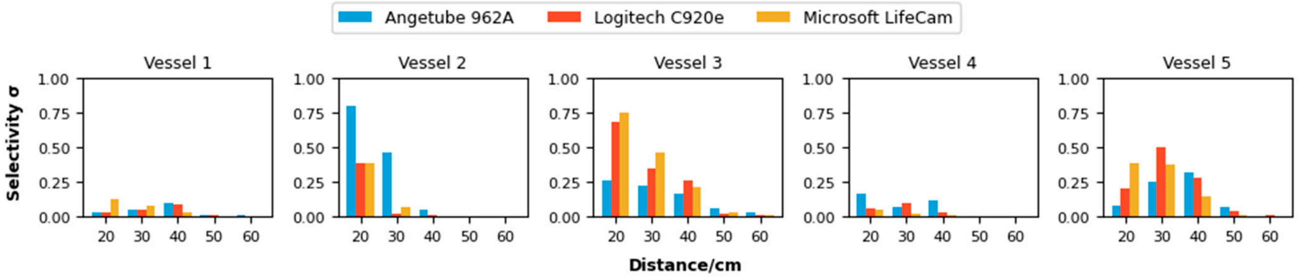


Figure 17. Selectivity σ (i.e., the mean absolute difference between prediction probabilities for cup and mug classes) comparison for cameras with similar FOV (scenario: Scenario 2).

There are some other interesting observations that can be made. With reference to Figures 15 and 16, the effect of transitioning from a prediction of coffee mug to cup with increasing distance for Vessel 3 in Scenario 1 is not replicated (at least as significantly) in Scenario 2. In fact, the pretrained Resnet-18 model is much more likely to classify Vessel 3 as a coffee mug rather than a cup in Scenario 2. In addition, with reference to Figure 17, the selectivity values for Vessel 3 in Scenario 2 are usually significantly higher for images captured by the Microsoft Lifecam than for images captured by the Angetube 962A; this is completely different from the case in Scenario 1.

These observations from Scenarios 1 and 2 jointly demonstrate that the output of the image classifier depends in a complex way on both the camera used to capture the image and the scenario (including such aspects as the background to the foreground objects and lighting). It may be that differences between cameras are not as evident when the background is plain but become increasingly evident when the background becomes more involved. This is explored more in the results for Scenario 3.

4.4. Scenario 3

Scenario 3 involves the most complex background of the three scenarios (see Figure 6). Figures 18 and 19 illustrate stacked bar charts showing the proportion of images classified as a cup or a coffee mug by the pretrained Resnet-18 model as a function of distance for different combinations of drinking vessel and camera. In Figure 18, only the top prediction class of an image is considered; in Figure 19, the top five prediction classes are considered such that if a cup and/or coffee mug appears in the top five, the highest ranking of these two classes is selected. The following subplots of Figures 18 and 19 are shaded with a gray face color to designate that they are of interest with respect to demonstrating a dependence of the classification on camera:

- Microsoft LifeCam and Vessel 2: the image classifier shows a propensity for classification of the vessel as a coffee mug, whereas with other cameras, the same classifier is more likely to select a cup.
- Angetube 962A and Vessel 5: the image classifier shows a propensity for classification of the vessel as a cup, whereas with other cameras, the same classifier is more likely to select a coffee mug.

In fact, these observed differences in the classification output are more pronounced than those observed in Scenario 2 and demonstrate a very significant effect of the camera used on the image classifier. Figure 20 compares the selectivity of the pretrained Resnet-18 model as a function of distance for the Angetube962, Logitech C920e, and Microsoft Lifecam. It is apparent that most of the selectivity values are very small for Vessel 2 and Vessel 5 in Scenario 3, which was not the case in Scenario 1 or Scenario 2. Again, this may explain the differences in classification by the image classifier highlighted above for these vessels. Specifically, the low selectivity means the image classifier is not easily able to distinguish between a cup and a coffee mug, so any small difference in the images produced by different cameras may be sufficient to change the top prediction.

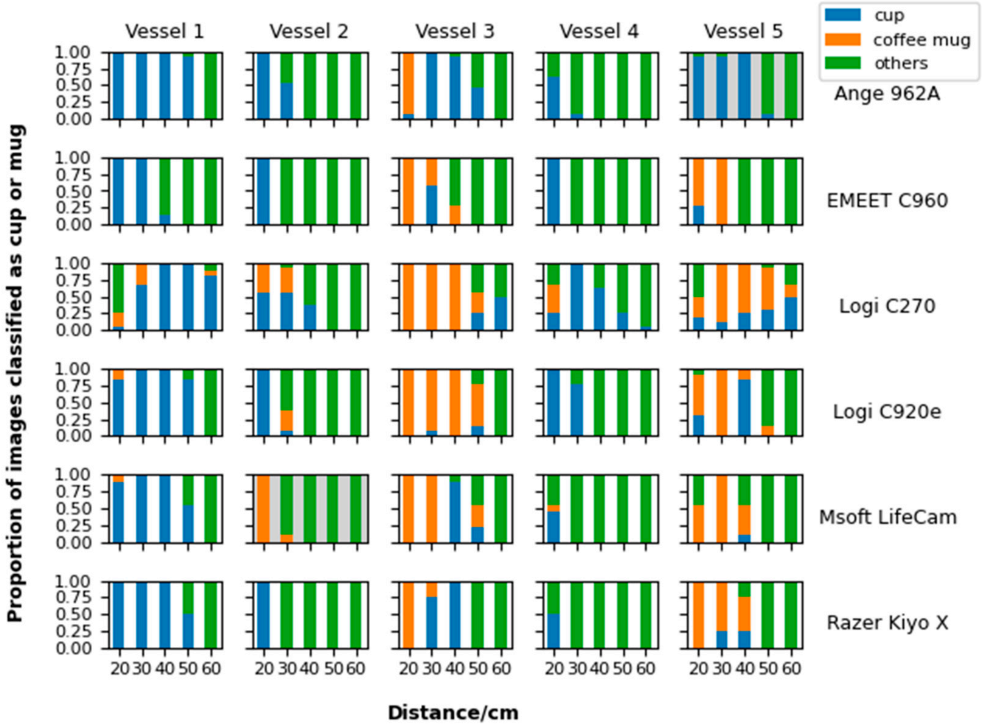


Figure 18. Proportion of images classified as a cup or a coffee mug in the top prediction as a function of distance for different drinking vessel and camera combinations (scenario: Scenario 3)—subplots of interest highlighted with a gray face color.

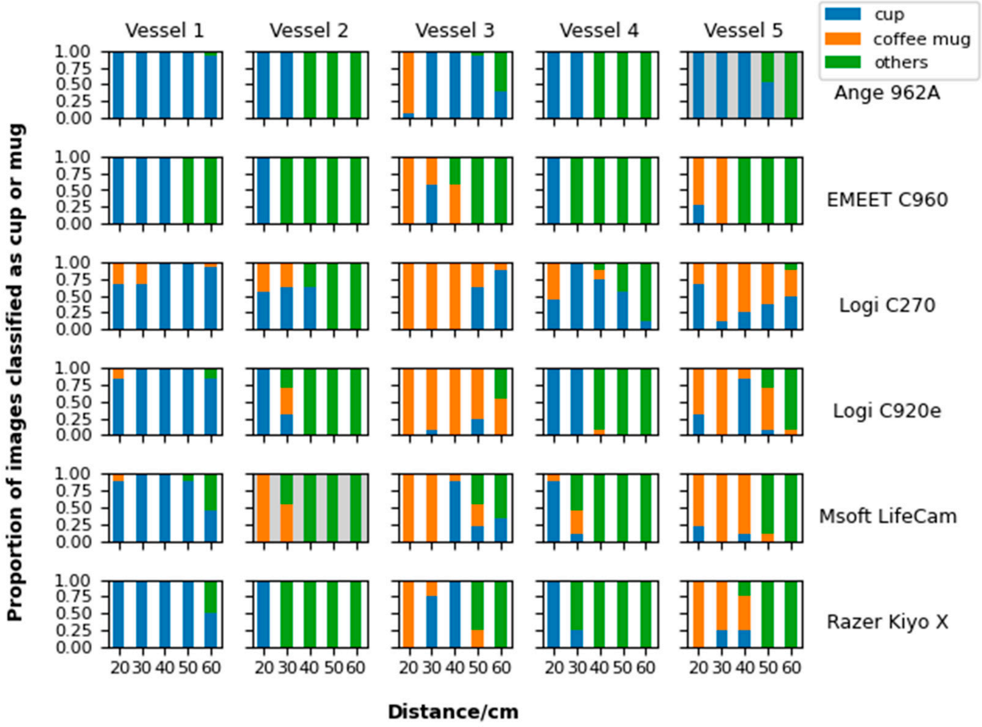


Figure 19. Proportion of images classified as a cup or a coffee mug in the top five predictions as a function of distance for different drinking vessel and camera combinations (scenario: Scenario 3)—subplots of interest highlighted with a gray face color.

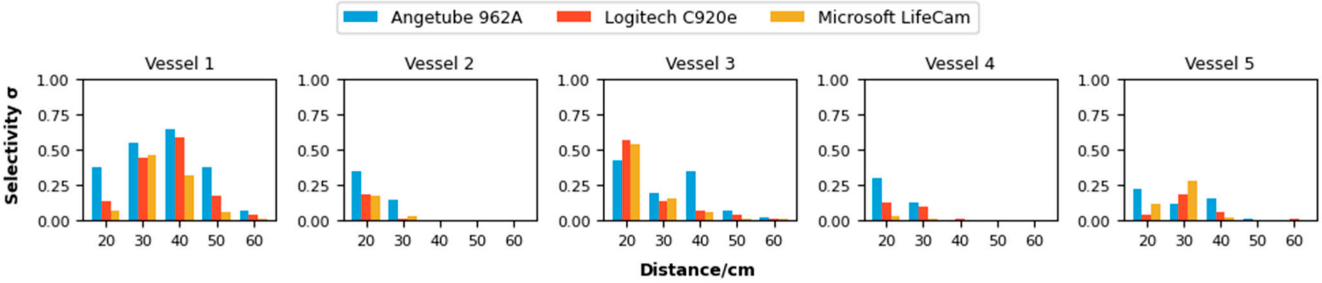


Figure 20. Selectivity σ (i.e., the mean absolute difference between prediction probabilities for cup and mug classes) comparison for cameras with similar FOV (scenario: Scenario 3).

Interestingly, with reference to Figures 18 and 19, the effect of transitioning from a prediction of coffee mug to cup with increasing distance for Vessel 3 in Scenario 1 is replicated in Scenario 3, but it was not apparent in Scenario 2. Again, this highlights the complex relationship between image-classification output, the camera used to acquire the image, and the exact scenario.

4.5. Visual Summary of Effect of Camera Choice

Figure 21 illustrates stacked bar charts showing the proportion of images classified as a cup or a coffee mug by the pretrained Resnet-18 model as a function of distance for all scenarios, Vessel 2 and Vessel 5, and the Angetube 962A, Logitech C920e, and Microsoft LifeCam cameras. Only the top prediction class of an image is considered. The three cameras have a very similar FOV.

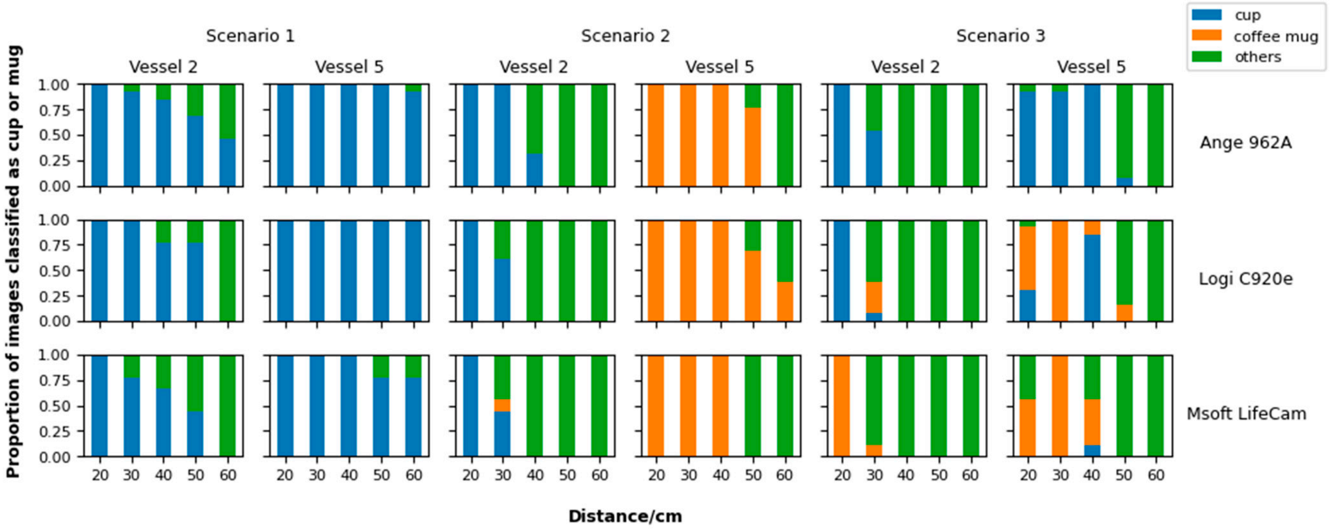


Figure 21. Proportion of images classified as a cup or a coffee mug as the top prediction as a function of distance for different scenario, drinking vessel, and camera combinations.

The pretrained Resnet-18 model is much more likely to classify the images of Vessel 2 as a cup rather than a coffee mug in both Scenarios 1 and 2, so there is no clear classification dependence on camera or scenario in this case. In contrast, the pretrained Resnet-18 model is much more likely to classify the images of Vessel 5 as a cup rather than a coffee mug in Scenario 1 and as a coffee mug rather than a cup in Scenario 2, irrespective of the source camera. Therefore, the source camera has no clear effect on the top prediction made by the model in these scenarios, but the classification ranking is highly dependent upon the scenario.

However, in Scenario 3, the model classifies images as follows:

- Images of Vessel 2 and Vessel 5 acquired by the Angetube 962A are much more likely to be classified as cups rather than coffee mugs.
- Images of Vessel 2 and Vessel 5 acquired by the Microsoft LifeCam are much more likely to be classified as coffee mugs rather than cups.
- Images of Vessel 2 and Vessel 5 acquired by the Logitech C920e are sometimes classified as cups and sometimes classified as mugs.

These findings are confirmed in Table 2, which shows the top prediction proportions for cups, coffee mugs, and all other classes for different scenario, drinking vessel, and camera combinations summarized over all distances. Hence, the source camera has a very significant effect on the top prediction made by the model in Scenario 3. This illustrates the complex relationship between model, camera, and scenario for image-classification inference.

Table 2. Top prediction proportions for different scenario, drinking vessel, and camera combinations (summarized over all distances).

Camera	**Scenario 1**				**Scenario 2**				**Scenario 3**			
	Vessel 2		**Vessel 5**		**Vessel 2**		**Vessel 5**		**Vessel 2**		**Vessel 5**	
Angetube 962A	Cup	0.78	Cup	0.98	Cup	0.46	Cup	0.00	Cup	0.31	Cup	0.58
	Mug	0.00	Mug	0.00	Mug	0.00	Mug	0.75	Mug	0.00	Mug	0.00
	Others	0.22	Others	0.02	Others	0.54	Others	0.25	Others	0.69	Others	0.42
Logitech C920e	Cup	0.71	Cup	1.00	Cup	0.32	Cup	0.00	Cup	0.22	Cup	0.23
	Mug	0.00	Mug	0.00	Mug	0.00	Mug	0.82	Mug	0.06	Mug	0.38
	Others	0.29	Others	0.00	Others	0.68	Others	0.18	Others	0.72	Others	0.38
Microsoft LifeCam	Cup	0.58	Cup	0.91	Cup	0.29	Cup	0.00	Cup	0.00	Cup	0.02
	Mug	0.00	Mug	0.00	Mug	0.02	Mug	0.60	Mug	0.22	Mug	0.40
	Others	0.42	Others	0.09	Others	0.69	Others	0.40	Others	0.78	Others	0.58

4.6. Quantitative Effect of Camera Choice

To assess the effect of camera choice on inference quantitatively, we examine the agreement of the model inference in terms of the top prediction class for comparable images taken by different cameras. As in the previous section, we limit attention to the Angetube 962A, Logitech C920e, and Microsoft LifeCam cameras because these three cameras have a very similar FOV, and they have a relatively large number of supported resolutions in common (see Table 1). For each scenario and camera, there are 150 comparable images corresponding to the following:

- Five target vessels;
- Five distances (20 cm, 30 cm, 40 cm, 50 cm, and 60 cm);
- Six common image resolutions (1280 × 720, 800 × 448, 640 × 360, 640 × 480, 320 × 240, and 352 × 288).

For each camera pairing and scenario, the most basic agreement statistic is the raw observed agreement proportion p_o between the comparable images of the two cameras, defined as follows:

$$p_o = \frac{n_{agree}}{N} \tag{2}$$

where

- n_{agree} is the number of compatible images for the two cameras for which the top model prediction is the same;
- N is the total number of comparable images for the two cameras (N = 150 per scenario in this case).

p_o has a range of between 0 and 1, with higher values indicating a greater level of agreement. $p_o = 0$ indicates zero agreement, and $p_o = 1$ indicates full agreement.

However, a more complex agreement statistic is Cohen's Kappa κ [44,45], which has a value in the range $-1 < \kappa \leq +1$ and compensates for the fact that an observed agreement in the top model prediction may have occurred by random chance. Formally, Cohen's Kappa κ is defined as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3}$$

where

- p_e is the theoretical probability of agreement by random chance and is calculated using the observed data as follows:

$$p_e = \frac{1}{N^2} \sum_c n_{ci} n_{cj} \tag{4}$$

where

- c is the index of the categories or classes that can be chosen (ranges from 1 to 1000 in this case, given that the ImageNet-1k dataset contains 1000 classes);
- n_{ci} is the observed number of times that the classifier predicted category c (out of N images) for images captured by one camera in the camera pairing;
- n_{cj} is the observed number of times that the classifier predicted category c (out of N images) for images captured by the other camera in the camera pairing.

Table 3 illustrates the raw observed agreement proportion p_o and Cohen's Kappa κ as a function of scenario and camera pairing. It is clear that Cohen's Kappa κ is always significantly smaller than the raw agreement proportion p_o. Although the model is trained on the ImageNet-1k database and can therefore make predictions on 1000 classes, it is usually classifying the images in this study as cups or mugs, and even when it makes different classifications, only a small percentage of the 1000 image classes are represented in the inference results. For this reason, p_e tends to be relatively large, so Cohen's Kappa κ is smaller than p_o.

Table 3. Top prediction agreement statistics for camera pairings and scenarios.

Camera Pairing	Scenario 1		Scenario 2		Scenario 3	
	Raw Agreement Proportion p_o	Cohen's Kappa κ	Raw Agreement Proportion p_o	Cohen's Kappa κ	Raw Agreement Proportion p_o	Cohen's Kappa κ
Angetube 962A/ Logitech C920e	0.86	0.56	0.52	0.44	0.37	0.25
Angetube 962A/ Logitech C920e	0.81	0.50	0.43	0.35	0.37	0.28
Logitech C920e/ Microsoft LifeCam	0.82	0.49	0.59	0.50	0.51	0.43

Table 3 demonstrates clearly that the best agreement in model predictions for different camera pairs is achieved in Scenario 1, while the worst agreement in model predictions occurs in Scenario 3. This is to be expected because Scenario 1 is the least cluttered scenario and so we would expect that differences between comparable images acquired by different cameras would have less bearing in the top model prediction.

It is also clear that the agreement in top model predictions for the Logitech C920e/Microsoft LifeCam pairing is significantly better than the other two camera pairings in Scenarios 2 and 3 (although not in Scenario 1). This quantitatively demonstrates that the chosen camera model can have a significant effect on the model inference output.

Although somewhat arbitrary, there is an interpretation attached to certain ranges of Cohen's Kappa κ [46], as illustrated in Table 4. From this table, the agreement between the top model predictions for comparable images taken with different cameras is moderate in Scenario 1 but generally only fair in Scenario 3. Again, this demonstrates the complex relationship between model, camera, and scenario for image-classification inference.

Table 4. Strength of agreement for ranges of Cohen's Kappa κ.

Range of Cohen's Kappa κ	Strength of Agreement
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

5. Conclusions

The primary objective of this research [47] was to determine whether and to what extent the camera employed to capture an image affects the image-classification output. The study employed a custom dataset of 4650 acquired images from six different cameras to demonstrate that the output of an image classifier during inference can depend significantly upon the specific camera that is used to acquire the image. This was shown both using graphical methods and by calculating Cohen's Kappa κ on the top predictions made by the same model on compatible images acquired by different cameras. This observed dependency is the case even when the foreground principal object in a scene is easily recognizable by a human observer and when the cameras employed have a similar FOV.

The extent of the dependence of the output of an image classifier upon the camera for a similar FOV is a function of the scenario and in particular, the background to the foreground object and the ambient lighting. In this exploratory study, we found that the dependence upon the camera is relatively small when there is a plain background to the foreground object. That is, the object class with the highest prediction probability is usually not affected by the specific camera used to acquire the image, but the prediction probabilities of the various object classes in the classification ranking do show some dependence on the camera. However, when the background is more complex, the object class with the highest prediction probability can sometimes be different depending upon the specific camera used to acquire the image.

It should be stressed that these results are more significant given that there are some elements of the image-classification ecosystem that inherently suppress the effect of different cameras:

- One function of the camera ISP is to correct for camera-specific degradations such as lens distortion. Of course, the ISP can never fulfill this function perfectly, and the quality of the ISP correction will depend upon the price of the camera, but the ISP still regularizes the image acquired at the source to some extent.
- During model training, validation, and testing, images from many different cameras will typically be used, so there is already an in-built level of data augmentation and diversity as far as cameras are concerned. This is certainly the case when models trained with the ImageNet-1k database are used for inference, as in this study, although the exact cameras represented are unknown or at least publicly unavailable.

Since there are many important applications of image-classification inference, such as industrial inspection, medical diagnosis, emotion recognition [48], and autonomous vehicle operation, there is a question about how to minimize the effect of the specific camera employed on the inference in the future. There are possible solutions both in how images are acquired during model training and during inference. For model training, a systematic plan can be employed to acquire training images from a vast array of cameras rather than relying on the current incidental diversity approach. However, this may not be realistic, as acquiring sufficient training images and training models is already a very lengthy and expensive exercise. In addition, this would not safeguard against the use of new cameras for inference that were not available at the time the model was trained. As an alternative, a generic trained model can be customized for use with a specific camera by capturing a relatively small number of training images using the camera and then optimizing the generic model with further training. In terms of model inference, a solution is to capture inference images not with one camera but simultaneously with multiple cameras of different types (and possibly with different FOVs) and then employ an algorithm to combine the individual camera classification rankings into a single overall classification result.

6. Limitations and Future Directions

This study on the effect of camera choice on image-classification inference was conducted using an image dataset captured with six distinct cameras, five distinct drinking vessels, and three distinct scenarios involving different image backgrounds. A larger dataset could be acquired by expanding the number of cameras, using more drinking vessels (or even different types of image targets such as vehicles, sports balls, etc.), and considering different image backgrounds. The difficulty with this is that the image-capture process is time-consuming because the capture of each image must be precisely staged (e.g., using a prescribed distance between the camera and the image target). However, acquiring a larger dataset for inference will be useful to confirm and expand on the results of this paper.

Another limitation of the study is that it was mostly conducted using a pretrained Resnet-18 model. The next phase of our research will involve characterizing the dependence of various pretrained image-classification models (in addition to the pretrained Resnet-18 model used in this paper) on the camera employed to acquire images for inference. It may be that other models are dependent to a greater or lesser extent on the specific camera used for image acquisition during inference.

In this study, the lighting was set to facilitate easy identification of the main target object in the images by a human observer. A related area of research is to vary the lighting conditions from dim to bright and observe the effect on image-classification inference.

Author Contributions: Conceptualization, J.B., A.N. and N.R.; methodology, J.B., A.N. and N.R.; software, J.B.; validation, J.B.; formal analysis, J.B.; investigation, J.B.; resources, J.B.; data curation, J.B.; writing—original draft preparation, J.B.; writing—review and editing, J.B., A.N. and N.R.; visualization, J.B.; supervision, J.B.; project administration, J.B., A.N. and N.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [CrossRef] [PubMed]
2. Shrestha, A.; Mahmood, A. Review of deep learning algorithms and architectures. *IEEE Access* **2019**, *7*, 53040–53065. [CrossRef]
3. Zhao, Z.-Q.; Zheng, P.; Xu, S.-T.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef] [PubMed]
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
5. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
6. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
7. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
8. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
9. ImageNet. Available online: https://www.image-net.org/ (accessed on 17 June 2023).
10. Drenkow, N.; Sani, N.; Shpitser, I.; Unberath, M. A systematic review of robustness in deep learning for computer vision: Mind the gap? *arXiv* **2021**, arXiv:2112.00639.
11. Dung, C.V. Autonomous concrete crack detection using deep fully convolutional neural network. *Autom. Constr.* **2019**, *99*, 52–58. [CrossRef]
12. Czimmermann, T.; Ciuti, G.; Milazzo, M.; Chiurazzi, M.; Roccella, S.; Oddo, C.M.; Dario, P. Visual-based defect detection and classification approaches for industrial applications—A survey. *Sensors* **2020**, *20*, 1459. [CrossRef] [PubMed]
13. Ren, R.; Hung, T.; Tan, K.C. A generic deep-learning-based approach for automated surface inspection. *IEEE Trans. Cybern.* **2017**, *48*, 929–940. [CrossRef]
14. Cai, L.; Gao, J.; Zhao, D. A review of the application of deep learning in medical image classification and segmentation. *Ann. Transl. Med.* **2020**, *8*, 713. [CrossRef] [PubMed]
15. Yadav, S.S.; Jadhav, S.M. Deep convolutional neural network based medical image classification for disease diagnosis. *J. Big Data* **2019**, *6*, 113. [CrossRef]
16. Fujiyoshi, H.; Hirakawa, T.; Yamashita, T. Deep learning-based image recognition for autonomous driving. *IATSS Res.* **2019**, *43*, 244–252. [CrossRef]
17. Turay, T.; Vladimirova, T. Towards performing image classification and object detection with convolutional neural networks in autonomous driving systems: A survey. *IEEE Access* **2022**, *10*, 14076–14119. [CrossRef]
18. Shi, X.; Yu, F.; Lu, Y.; Liang, Y.; Feng, Q.; Wang, D.; Qian, Y.; Xie, L. The accented english speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6918–6922.
19. Kardava, I.; Antidze, J.; Gulua, N. Solving the problem of the accents for speech recognition systems. *Int. J. Signal Process. Syst.* **2016**, *4*, 235–238. [CrossRef]
20. Maître, H. *From Photon to Pixel: The Digital Camera Handbook*, 2nd ed.; Digital signal and image processing series; Wiley: Hoboken, NJ, USA, 2017.
21. Dodge, S.; Karam, L. Understanding how image quality affects deep neural networks. In Proceedings of the 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, Portugal, 6–8 June 2016; pp. 1–6. [CrossRef]
22. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.
23. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv* **2014**, arXiv:1405.3531.
24. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
25. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
26. Atif, M.; Ceccarelli, A.; Zoppi, T.; Bondavalli, A. Tolerate Failures of the Visual Camera with Robust Image Classifiers. *IEEE Access* **2023**, *11*, 5132–5143. [CrossRef]
27. Ceccarelli, A.; Secci, F. RGB cameras failures and their effects in autonomous driving applications. *IEEE Trans. Dependable Secur. Comput.* **2022**, *20*, 2731–2745. [CrossRef]

28. Hendrycks, D.; Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv* **2019**, arXiv:1903.12261.
29. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
30. Pei, Y.; Huang, Y.; Zou, Q.; Zhang, X.; Wang, S. Effects of Image Degradation and Degradation Removal to CNN-Based Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 1239–1253. [CrossRef] [PubMed]
31. Karahan, S.; Yildirum, M.K.; Kirtac, K.; Rende, F.S.; Butun, G.; Ekenel, H.K. How image degradations affect deep cnn-based face recognition? In Proceedings of the 2016 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 21–23 September 2016; pp. 1–5.
32. Wang, Z.; Chang, S.; Yang, Y.; Liu, D.; Huang, T.S. Studying very low resolution recognition using deep networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4792–4800.
33. Liu, D.; Cheng, B.; Wang, Z.; Zhang, H.; Huang, T.S. Enhance visual recognition under adverse conditions via deep networks. *IEEE Trans. Image Process.* **2019**, *28*, 4401–4412. [CrossRef]
34. Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 8340–8349.
35. Djolonga, J.; Yung, J.; Tschannen, M.; Romijnders, R.; Beyer, L.; Kolesnikov, A.; Puigcerver, J.; Minderer, M.; D'Amour, A.; Moldovan, D.; et al. On robustness and transferability of convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16458–16468.
36. Taori, R.; Dave, A.; Shankar, V.; Carlini, N.; Recht, B.; Schmidt, L. Measuring robustness to natural distribution shifts in image classification. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18583–18599.
37. HajiRassouliha, A.; Richardson, S.P.; Taberner, A.J.; Nash, M.P.; Nielsen, P.M. The effect of camera settings on image noise and accuracy of subpixel image registration. *Mach. Vis. Appl.* **2021**, *32*, 95. [CrossRef]
38. Kaur, R.; Karmakar, G.; Xia, F.; Imran, M. Deep learning: Survey of environmental and camera impacts on internet of things images. *Artif. Intell. Rev.* **2023**, *56*, 9605–9638. [CrossRef] [PubMed]
39. Kuciš, M.; Zemcık, P. Simulation of camera features. In Proceedings of the 16th Central European Seminar on Computer Graphics, Smolenice, Slovakia, 29 April–1 May 2012; pp. 117–123.
40. Mahmoudi, A.; Sabzehparvar, M.; Mortazavi, M. A virtual environment for evaluation of computer vision algorithms under general airborne camera imperfections. *J. Navig.* **2021**, *74*, 801–821. [CrossRef]
41. Ouyang, H.; Shi, Z.; Lei, C.; Law, K.L.; Chen, Q. Neural camera simulators. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7700–7709.
42. Farrell, J.E.; Xiao, F.; Catrysse, P.B.; Wandell, B.A. A simulation tool for evaluating digital camera image quality. *Image Qual. Syst. Perform.* **2003**, *5294*, 124–131.
43. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
44. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
45. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
46. Richard, L.J.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [CrossRef]
47. Khan, M.; Gueaieb, W.; Elsaddik, A.; Masi, G.D.; Karray, F. Graph-Based Knowledge Driven Approach for Violence Detection. In *IEEE Consumer Electronics Magazine*; IEEE: Piscataway, NJ, USA, 2024. [CrossRef]
48. Khan, M.; Saddik, A.E.; Deriche, M.; Gueaieb, W. STT-Net: Simplified Temporal Transformer for Emotion Recognition. *IEEE Access* **2024**, *12*, 86220–86231. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.