**Article:**
APPIAH, KOFI ESSUMING (2023) Naturalistic Scene Modelling:deep Learning with Insights from Biology. Signal Processing Systems. 1153–1165.

https://doi.org/10.1007/s11265-023-01894-4

Journal of Signal Processing Systems (2023) 95:1153–1165
https://doi.org/10.1007/s11265-023-01894-4

# Naturalistic Scene Modelling: Deep Learning with Insights from Biology

Kofi Appiah[1] · Zhiyong Jin[2,3] · Lei Shi[4] · Sze Chai Kwok[2,3,5]

Received: 17 January 2023 / Revised: 5 September 2023 / Accepted: 24 September 2023 / Published online: 14 October 2023

## Abstract

Advances in machine learning coupled with the abundances of training data has facilitated the deep learning era, which has demonstrated its ability and effectiveness in solving complex detection and recognition problems. In general application areas with elements of machine learning have seen exponential growth with promising new and sophisticated solutions to complex learning problems. In computer vision, the challenge related to the detection of known objects in a scene is a thing of the past. With the tremendous increase in detection accuracies, some close to that of human detection, there are several areas still lagging in computer vision and machine learning where improvements may call for more architectural designs. In this paper, we propose a physiologically inspired model for scene understanding that encodes three key components: object location, size and category. Our aim is to develop an energy efficient artificial intelligent model for naturalistic scene understanding capable of deploying on a low power neuromorphic hardware. We have reviewed recent advances in deep learning architecture that have taken inspiration from human or primate learning systems and provided direct to future advancement on deep learning with inspiration from physiological experiments. Upon a review of areas that have benefitted from deep learning, we provide recommendations for enhancing those areas that might have stalled or grinded to a halt with little or no significant improvement.

**Keywords** Scene understanding · Machine learning · Deep learning · Physiologically inspired models

## 1 Introduction

The past decade has seen a revival in the area of Artificial Intelligence (AI) and machine learning, mainly driven by the impressive reported performance of deep learning. Computer vision with its sister counterpart natural language processing have improved significantly and contributed to several key application areas [1]. The likes of Amazon Echo, Google Home and various other internet of things devices we find around in our homes generate interesting data that can be used to train and recognise various activities using deep learning.

Zhiyong Jin, Lei Shi and Sze Chai Kwok contributed equally to this work.

✉ Kofi Appiah
kofi.appiah@york.ac.uk

Zhiyong Jin
jinz.y950713@gmail.com

Lei Shi
sl1012002322@126.com

Sze Chai Kwok
sze-chai.kwok@st-hughs.oxon.org

[1] Department of Computer Science, University of York, Deramore Lane, York YO10 5GH, Yorkshire, England

[2] Phylo-Cognition Laboratory, Division of Natural and Applied Sciences, Data Science Research Center, Duke Kunshan University, Duke Institute for Brain Sciences, Kunshan 215316, Jiangsu, China

[3] Shanghai Key Laboratory of Brain Functional Genomics, Key Laboratory of Brain Functional Genomics (Ministry of Education), Affiliated Mental Health Center (ECNU), School of Psychology and Cognitive Science, East China Normal University, Shanghai, China

[4] Department of Neurosurgery, China Medical University, The First People's Hospital of Kunshan, Suzhou 215300, Jiangsu, China

[5] Shanghai Changning Mental Health Center, Shanghai, China

Image recognition and classification systems benefit enormously from deep learning and have contributed to several consumer applications like the Apple photo organiser [2], capable of grouping similar pictures with a common theme. Facebook's face recognition [3] system is another classical example of how image recognition systems have been incorporated into social media. At industrial level, the use of AI has contributed in various way. For example, Amazon's warehouses are heavily reliant on robots for moving shelves with load [4]. The robots have sense of coordination and can navigate around the warehouse, avoiding collisions with other robots as well as many other stationary objects. In healthcare, medical image analysis contributes to non-invasive diagnosis [5, 6]. The number of useful applications for AI and computer vision increases on daily basis and it is all around us [7–10]. In agriculture, satellite imagery has contributed in various ways to estimate crop yield [11–14]; in sports like football (or soccer) it has been used in making decisions like goal-line technology [15]. Various manufacturing companies use computers as part of their production line to identify defective items [16]. Similarly, the success of driverless cars is heavily reliant on the use of computer vision to identify object in the scene [17, 18]. What is really missing is the ability for machines (robots) to see, recognise and react to their immediate surroundings just like humans with their most complicated cognitive ability: vision. Thus, the use of perception to generate knowledge.

This work explores the failures in existing machine learning models for scene understanding and proposes new directions for modelling the scene with inspiration from bio-inspired vision as well as experimental results to demonstrate the effectiveness of this approach. A key contribution in this exploratory work is how attention mechanisms studied by psychologist have demonstrated improvement in modelling global descriptors to fully understand a visual scene.

## 2 Current State of Affair

Machine learning (a subset of artificial intelligence) has rapidly evolved in the past decade and continues to advance. Several key areas have seen significant progress and have become popular research topics and the very few related to this work have been listed below:

- Deep Learning: Deep learning has been at the forefront of machine learning advancements. Neural networks with multiple layers (deep neural networks) have achieved remarkable results in various domains, including image recognition, natural language processing, and speech synthesis. Techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been extensively explored and improved upon. In [19], deep learning has been used in monitoring a construction site to identify any safety concerns as well as worker behaviour.
- Transfer Learning and Pre-trained Models: Transfer learning has gained prominence, allowing models to take advantage of knowledge from pre-training on large datasets. Pre-trained models like Bidirectional Encoder Representations from Transformers (BERT) [20], Generative Pre-trained Transformer (GPT) [21], and ImageNet-trained CNNs have demonstrated excellent performance across various tasks. By fine-tuning these models on specific data, researchers have achieved state-of-the-art results with smaller labelled datasets. Transfer learning has been used effectively in areas like medical imaging [22] where labelled data may not be readily available.
- Reinforcement Learning: Reinforcement learning has made significant strides, especially in the field of game-playing agents. Algorithms like Deep Q-Networks (DQN) [23], have achieved superhuman performance in games like Go, Chess, and Dota 2. Reinforcement learning has also been applied to robotics, control systems, and recommendation systems [24].

### 2.1 The Key Factors for AI Acceleration

Standard neural networks which would normally consist of many simple neurons that may produce a sequence of real-valued activation have been around for many decades [25]. Purely supervised neural networks improved significantly during the 1990 s and 2000 s, which has contributed to the success of deep learning and artificial intelligence in general. Three main factors have contributed to the acceleration of AI, making it possible to incorporate AI into various application areas. The key and driving factors for the acceleration of AI are three folds:

1. Better algorithms from supervised and unsupervised learning [26, 27]
2. Huge volumes of data available from multiple sources on the internet [28–30] and
3. Improved computational power from heterogenous architectures [31–34].

We have moved from the days of predefined rules (Expert Systems) and have better algorithms that learn from examples [25]. Neural network and machine learning algorithms have contributed to the success of deep learning in various ways. Deep learning architecture mainly rely on cascading several neural networks with various machine learning based classifiers. The inception of internet and social media has contributed to the volume of useful data generated every second. It is estimated that approximately 300 h of video are uploaded onto YouTube every minute [35]; that contributes to the number of useful training data. Similarly, a white paper published by Facebook reported that its users upload approximately 350 million new photos each day [36]. These huge volumes of data generated by various people around the world have positive impact of datasets needed to train supervised learning algorithms like deep learning. This has been made possible by the fast internet access we enjoy these days and the volume of data is expected to increase with ubiquitous internet of things devices [27].

Thanks to Moore's law, the computing power has constantly been increasing for the past three decades. The ability to improve the performance with fast processors has reduced, making way for processing speed to increase with concurrent and parallel execution [37]. The dominance of Graphics Processing Units (GPU) and Field Programmable Gate Arrays (FPGA) for applications like image processing that naturally benefit from parallel execution has also contributed to the AI acceleration [32, 34].

## 2.2 Current Machine Learning Achievements

There have been significant and key achievements in computer vision that are worth pointing out. The rich nature of current computer vision algorithms, thanks to deep learning, have made it possible to perform visual recognition with accuracies as high as that of humans and even outperform humans in certain instances [38]. These accuracies are really limited to recognition and detection - perception rather than understanding of visual scenes - knowledge. Zhang et al. [39] proposed to leverage emerging deep reinforcement learning techniques for enabling model-free driverless vehicles control, and present a novel and highly effective control framework, which utilises the powerful convolutional neural network for feature extraction of the necessary information (including traffic flow), then makes decisions under the guidance of the network. How reinforcement learning is deployed in the field of decision making is illustrated in these recent works [40, 41]. Tremendous achievements have also been made in medical imaging, where deep learning approaches have helped with the early detection of tumours in images taken from the brain [42].

There have been reported cases of the use of deep learning for the detection of leakage in 3D blood vessel [42], leaks that were missed by medical professions were easily detected by machine learning based systems. The list in medical imaging continues with capabilities of lesion detection in the eye as well as various cancer cells [42]. Detection of known objects in an image has matured to a level that computer vision techniques have the capabilities of detecting multiple objects in a single image even if they are partially occluded, with very high accuracy levels [43]. Work on how to construct meaningful sentences from images have also produced impressive results [44]. Use of multiple images for the reconstruction of 3D environments have also been achieved to an acceptable level. The ability to stitch together multiple 2D images to form a 3D panoramic view has also been reported by Song et al. [45]. He et al. [46] presented Mask regional convolutional neural network (R-CNN) which is conceptually simple and aims to segment or separate each occurrence of any object in an image. Faster R-CNN [43] has two outputs for each candidate object, a class label and a bounding-box offset; to this, [46] added a third branch that outputs the object mask. Mask R-CNN [46] is thus a natural and intuitive idea, which basically combines two state-of-the-art models (a region proposal network and a binary mask classifier). But the additional mask output is distinct from the class and box outputs, which requires extraction of much finer spatial layout of an object, making it possible for such a model to be used innovatively in medical imaging.

There are also a number of challenges one would have to consider when trying to deploy a deep learning model in a real-world application. When the domain of application has limited training data available, deep learning may not be the best choice. Similarly, the training process makes accurate predictions based on statistical associations and hence application that relies heavily on causality rather than correlation may not be the best fit. These are only a few of the challenges associated with the deployment of deep learning and the next section will further highlight some of these problems in the real-world.

## 2.3 Areas where Boosting would be Needed

The hype around computer vision grows exponentially and it is worth pointing out that, even though computer vision and AI in general have made significant progress and increasingly solving many complex problems [7–10], there are several shortfalls [38]. Thinking of the application areas where computer vision can be used, the very few areas that they have dominated [1–5] shouldn't overshadow instances where they underperform and would still need significant improvement [47]; if not a complete overhauling of the existing techniques.

In March 18, 2018, Uber's autonomous car hit and killed a 49-year-old as she was walking her bike across the street at night in Tempe, Ariz [48, 49]. With its 360-degree cameras and sensors, the car should have been able to detect someone crossing in front of it, even at night. Safety reports released by Uber in November 2018 said the software that detects obstacles on the road and processes that information was too slow to act. In the same month, on March 23, 2018, a Tesla vehicle in the autopilot mode slammed into a concrete road divider killing its driver [50]. With all the sensors including cameras around today's autonomous cars, one can be fooled in thinking these cars can really operate autonomously. What is really missing is unfamiliar scenes or combination of objects that these autonomous cars have not been trained with or can't interpret correctly [48] as well as real-time processing needs. There is no doubt that the way in which computer vision techniques have been used in these systems are novel, but we need to note that, the systems are not yet perfect [38].

Another example to support this is the fact that a robot trained to open doors struggled to open a significantly different door [51]. This goes to confirm that when such systems are challenged with very different scenarios their behaviour may not necessarily conform to what we expect. In another incidence on 12th July 2016, a Stanford mall security robot collided with a 16-month old toddler and nearly run him over [52]. Programmed to predict crimes in schools, businesses, and neighbourhoods, the K5 robot uses video cameras, thermal imaging sensors, a laser range finder, radar, air quality sensors and a microphone to detect irregularities in the area. If it detects any abnormal noise, temperature change or even appearance of known criminals, it will notify authorities. It turns out that the robot did not detect the young boy.

To compare systems trained to understand a scene using deep learning and what is proposed in this work, we use Table 1 to summarise the difference.

## 3 Unwrapping the Failures

These failures point to the fact that AI and computer vision techniques have performed exceptionally in some areas (like image categorisation [2] and industrial packaging [4]) but doesn't mean that new and emerging areas (like autonomous vehicles [38]) will enjoy the same benefits with simple tweaks. For autonomous cars, the state-of-the-art in computer vision is good and provides bounding boxes around objects in the scene. But detected objects are normally pre-trained and the system manages to recognise variant of such objects individually, with a high degree of accuracy. What is missing or becomes challenging is putting together the individual objects to give a global interpretation of the scene. State-of-the-art recognition systems have no contextual understanding of the scene at a global level; hence such systems have little ability to make acceptable and reasonable scene-level decisions [53].

For example, a navigational robot will generally be able to move from one point to the other, avoiding collisions to arrive at the destination in the most optimal way [54]. When it becomes challenging is when the robot will have to decide based on the scene rather than localised objects to take a safer route, which might not necessarily be the shortest path. Such decisions are made by humans intuitively, as they understand how objects interact in a scene [55]. However, in the case of a robot the decision taken might be optimal but not necessarily feasible or safe. Modelling a crowded scene to infer interaction as well as unusual situations with little or no data poses minimal problems to humans; but it can be incredibly hard for state-of-the-art AI systems to handle. Generally, things that humans find intuitive (such as dealing with complex scenes and walking), tend to be very hard for artificially modelled systems. This goes a long way to explain why machine learning techniques with inspiration from biological systems tend to outperform pure statistical models [56]. Deep learning models like convolutional networks used in computer vision and other related application areas represent candidate models for the computations performed in mammalian visual systems [56]. The convolution layers found in Convolutional Neural Networks (CNNs) mimic the effect of the brain to calculate the information from visual inputs. Visual object recognition framework has gained renewed interest with the success of deep neural network models trained to "recognise" objects: these hierarchical feed-forward networks show similarities to human visual cortex, including categorical separability [57]. Deep neural networks (DNN) like convolutional neural networks and recurrent neural networks (RNN) have existed since 1990, but only improved to an acceptable performance level in the past decade, thanks to all the three factors that have accelerated AI in general [58].

## 4 Pre-DNN Detection and Classification

Prior to the huge drive for the use of DNN in computer vision, standard feature detectors like scalar-invariant feature transform (SIFT) [59], histogram of oriented gradients

**Table 1** Some differences between a deep learning trained systems for scene understanding and what this work proposes.

| DL-based scene understanding | The proposed scene understanding |
|---|---|
| With minimal training data it fails to understand the scene | It is trained to understand what is normal with minimal data |
| When the new data is significantly different from the training data results is unpredictable | The learning is more explainable and expected to handle new data logically |
| When there are multiple interpretations it is likely to fail | Ambiguity is avoided by training to understand the entire scene not just combined features |

(HOG) [60] and local binary patterns (LBP) [61] had been the dominant hand-crafted features in many computer vision tasks and neuroscientific studies [62]. Compared to CNN where features are learnt and stacked in a hierarchical structure, features are hand-crafted in SIFT, LBP and HOG. Rather than the use of different algorithms for object detection and categorisation, with CNN the same algorithm is adapted for the same purpose at the expense of requiring large volume of training data [59]. A standard neural network (NN) consists of many simple, connected neurons, each producing a sequence of real-valued activations, and may suffer from the curse of dimensionality [63] or might not scale very well. Input neurons get activated through sensors perceiving the environment, while other neurons get activated through weighted connections from previously active neurons [25]. Unlike NN or shallow NN [25], deep neural networks have three distinguishing factors that make it possible to minimise the common problems associated with standard neural networks. There are more neurons with varying width, height and depth in DNN compared to shallow NN. Also, DNNs enforce local connectivity between adjacent neurons and replicate each filter across the entire visual field to allow translational invariance.

Convolutional neural network was first proposed by Kunihiko Fukushima in 1982 [1], whose work was inspired by an article published in 1962 by Hubel and Wiesel [64], on revealing the mechanism of the visual system. Since then, there has been a lot of research in this area with the most significant ones being that of LeCun in 1989 [65] and Krizhevsky in 2009 [3]. Deep convolutional networks became illustrious in 2012 when Krizhevsky et al. [28] used CNNs to win the annual computer vision challenge with an impressive 15% error rate compared to 26% in the previous year. Deep convolutional networks belong to a class of deep, feedforward artificial neural networks with backpropagation to transmit signals backward for training. In CNN, the weights of the convolutional layers are used for feature extraction and the weights of the fully-connected layers are used for classification; these can be determined through the training process. The success stories about the rise of CNNs and their capabilities of learning high-level features in object recognition have increased steadily since 2012 [34] and keep increasing due to the availability of large datasets like ImageNet [30]. Given that deep learning architectures can classify objects in an image with near-human-level performance, other studies have revealed some of the shortfalls of CNNs in computer vision [38, 66]. Nguyen et al. [38] have demonstrated that discriminative DNN models are easily fooled to classify many unrecognizable images with very high certainty as members of a recognisable class.

Here we describe the shortfalls of CNNs using the three key stages of their design: training data requirements, the actual training process and finally the recognition or testing phase. To properly train a CNN architecture, the training data is expected to be large enough to cover most variations. Typically, to train the architecture to detect birds, the training data is expected to have all kinds of bird orientations, various image resolutions as well as all possible actions that a bird may perform [25]. This is clearly not the case for humans as they can easily recognise a bird in any state after learning about them in a single state [67]. The training phase as well as internal architecture of CNNs have long been assumed to be black-boxes, however because they are computer programs one can easily step through to understand how input images are represented in each stage. Turner et al. [68] used various input images to visualise the output of every single layer of a CNN architecture (Visual Geometry Group VGG-19 network). In [68] it becomes clear that the internal output of the various layers of a CNN may not necessarily say much about the input image. Other forms of visualisations have also been used to analyse the internals of CNN architectures, these are the Activation Maximization [66], Network Inversion [69], DeconvNet [70] and Network Dissection [71]. These visualisation techniques do not only show the low-level features but also explain the working mechanisms of CNNs in general. It must be noted that the existing visualisation tools do have their limitations compared to the capabilities of humans as reported by neuroscientists on the mechanisms associated with the visual system [72]. Another problem associated with CNN training, is the demand for huge computational time and power to detect and classify an object [73], compared to the visual systems in which an object can be recalled in few seconds by using minimal resources in brain [74]. Also, CNN models require a large search space (including the depth, the number of feature maps, interconnection patterns, window sizes for convolution and pooling layers), making them impractical to discover an optimal network structure with any systematic approaches [75].

Finally, the recognition or test phase of CNN still have some level of errors, especially in multi-object recognition and classifying tasks [28]. With all the huge training data, CNNs are not able to produce recognition without errors. Even though the error rates are very minimal, they are still not acceptable for application areas like autonomous or driverless cars. Compared to human recognition system it would be odd to find a sound and healthy individual not being able distinguish between two different items like an apples and bananas. The reported accidents [48–50] related to self-driving vehicles go a long way to confirm that the current state of deep convolutional networks aren't able to handle complex situation or the intriguing fact that they can be easily fooled to misinterpret unrecognisable object with high confidence [38].

## 5 Summary of Recommendations

Some of the problems with current artificial systems have been identified and it would take some time to resolve [76], if they can ever be solved at all. Computer scientists have long been working with other key subject areas like physics, engineering and mathematics to solve challenging computer vision problems. What is missing and might be crucial to address most of the challenges in computer vision is the combined expertise from biology and psychology. To succeed in taking computer vision to the next level of robust scene understanding, the most appropriate research direction should be multi-disciplinary involving neuroscientists, psychologists, and physiologists. Thus, the use of biological and physiological data collected from experiments can be used to inform the design of models that mimic human vision and interpretation of a scene. There have been some successful cross-fertilization examples in visual cognitive neuroscience and CNN that provide a rationale for multi-disciplinary work for robust scene understanding. Greene et al. [77] presented a model for visual scene categorisation that reflects functions or actions that can be performed within a scene. The model in [77] is much closer to human scene categorisation and outperformed alternative models like object-based distance and visual features from CNN. Another study by Groen et al. [78] determined the contributions of the models tested in [77] to neural representations in scene-selective cortex by disentangling different types of information in the visual cortex. Besides, how strongly a simple property of the visual encoding of an image and its population response magnitude correlates with its memorability demonstrate how memory is shaped by visual context is presented in [79]. It is however worth noting that comparisons of deep network models with empirical electrophysiological, functional magnetic resonance imaging, and behavioral data do not invariably only show similarities between brains and models [80], but also at times discrepancies [76].

The main reason why DNN have managed to achieve state of the art performance has been linked to the human visual system by way of how they learn uninterpretable solutions. This has been reiterated in [75] with the development of a CNN model which borrows biological guidance from the human visual cortex and capable of determining critical design choices with simple calculations. The model in [75] simulates the V1, V2, V4 and inferotemporal cortex (IT) layers of the human ventral stream, uses convolutional layers with varied sizes and complexities and increased the use of concurrency for improved processing speed. The design presented by Zhang et al. [75] outperformed seven other CNN techniques to achieve state-of-the-art performances on four widely used benchmark datasets: CIFAR-10, CIFAR-100, SVHN and MNIST. These views are also advocated by others that the brain's innate structures such as connectivity and mechanisms will inform deep learning network models and steer them toward more authentic human-like learning [81]. Rajalingham et al. [76] demonstrate that state-of-the-art deep convolutional neural network models cannot account for the image-level behavioural patterns of primates (humans and macaque monkeys) and made a strong case for the design of new models that precisely capture the neural mechanisms underlying primate object vision. The experiments conducted in [76] confirm that the failure of current DNN models to accurately capture the image-level signatures of primates cannot easily be rectified by simply modifying the existing architectures, but rather a complete overhaul of the models and architectures.

Redmon et al. [82] presented a new approach for object detection, based on the fact that humans' glance at an image and instantly know what objects are in the image, where they are, and how they interact. The same analogy has been used in [76] that primates, including humans, can typically recognise objects in visual images at a glance despite naturally occurring identity-preserving image transformations. The model presented in [82] reframe object detection as a single regression problem, straight from image pixels to bounding box coordinates and class probabilities. What makes the YOLO model [82] different from other DNN models is the fact that it does not rely on sliding window but rather implicitly encodes contextual information about classes as well as their appearance. The model [82] has twice outperformed other state-of-the-art DNN models like Deformable Part Models (DPM) and Regional CNN using ImageNet and COCO datasets; a confirmation of the observations in [76].

### 5.1 Scene Understanding

To understand the scene, Zhou et al. [71] combined various local and global features, used CNN to learn deep features from the scene to present categories like humans, with the assumption that in an image set, high density is equivalent to the fact that images have in general similar neighbours. Semantic segmentation is a challenging task in computer vision which assigns a category label to each pixel of an image; a fundamental but challenging task in computer vision research [83]. There are other related feature descriptors that are combined in various ways to model and understand a scene using computer vision techniques. Table 2 provides a list of some of the common feature descriptors used in scene understanding.

There has been great advances in the research into the modelling of scenes for segmentation and detection using convolutional neural networks [84]. However, similar to [71] and related CNN based scene understanding techniques, segmentation-based convolutional neural networks require

**Table 2** Some key scene understanding features and their importance.

| Descriptor | Importance |
|---|---|
| Object Detector | This involves the identification and classification of various objects in the scene |
| Semantic Segmentation | This divides the image into meaningful regions represents a single entity |
| Motion Estimator | This is the apparent motion of objects in-between two subsequent image frames |
| Depth Estimator | The ability to estimate the distance of an already detected object from the imaging sensor |
| Local Descriptor | These are feature descriptors local and specific to an object like the texture, the edges, the corners, shape and orientation |
| Global Descriptor | These are the contextual information relating to the surrounding elements like scene geometry and spatial information |

tremendous computational power and are not optimal for future autonomous vehicles. To address the scene understanding and power consumption, Gaurav et al. [85] presented a deep spiking neural model that translates a conventional CNN model into it's spiking equivalence. The work demonstrate the capabilities of spiking neural architectures as well as the energy efficiency of neuromorphic hardware architectures like the Intel's Loihi [86].

### 5.2 Object-Scene Appearance Modelling

The strong need for a more robust scene understanding model suitable for application areas like autonomous or self-driving cars has motivated our proposed model for scene understanding with the use of physiological data. Our work has been motivated by the conclusions drawn from Eckstein et al. [87], which emphasised that missing giant targets is a functional brain strategy to discount distractors. The work in [87] demonstrates that search is guided toward target sizes consistent with the scene and thus, if targets are scaled to be larger but inconsistent in size with the scene, it would be missed more often during visual search. To utilise the results from [87] in our model, we will conduct further experiments (cf. [87]) to understand how humans and primates recognise key objects in a scene. Rather than using synthetic scenes, we will combine natural and synthetic scenes as part of our experiments.

Similar to the approach used by Izadinia et al. [88], we argue that the type of scene can be determined by the objects and their sizes, as well as their distribution. For example, in a kitchen we expect to see a worktop, cabinets and possibly a kettle or microwave around the worktop. We also aim to avoid the place category as in Zhou et al. [89] and provide a list of objects in the scene with associated spatial relationships, considering their relative sizes. Deep convolutional networks take full advantage of the ubiquitous and improved computational power from heterogenous architectures, that resource will be utilised in generating our exhaustive list and relationship between scene objects. The key is to establish how primates, especially humans, use spatial relationships between known objects during visual search, beyond those attempts relying solely on salience and contextual cues [90]. This would also be different from those studies which have focused on passive free viewing by humans and monkeys [91, 92]. The proposed model will incorporate global and local descriptors. The need to build a model that utilises spatial relationship between objects, the aspect ratio of the objects and a pair-wise relationship between objects is the driving factor and novelty behind our model described in the following two sections.

### 5.3 Eye Tracking Paradigm on Humans

This section details an eye tracking paradigm on humans that will be used to compare results from our computational model. Clues will also be taken from the experimental design to inform and make our model more biologically inspired. Experimental design: 20 participants will perform a search-identification task. The design contains six conditions by crossing two factors: the number of objects (in this case 2, 3 or 5) and time limit allowed for targets search/identification (limited time or self-paced). We will use eye tracking to record eye movements throughout the task and the eye movement data will be incorporated into the computational model. Each trial is composed of three phases (Fig. 1).

In terms of experimental material, we will generate 120 scenes with a gameplay (The Sims 4, Maxis Software, Electronic Arts). These scenes are unique and made to depict various types of indoor scenes including those of kitchens, living rooms, bedrooms, classrooms, and offices. Overall, 960 objects are extracted from these scenes removing any information associated with the original scenes in which they have been taken from (8 objects are extracted from each scene). The objects are made to be of comparable dimension when presented to the participants as part of the trial (second phase as shown in Fig. 1). To track and record the eye movement we use EyeLink 1000 Plus [93].

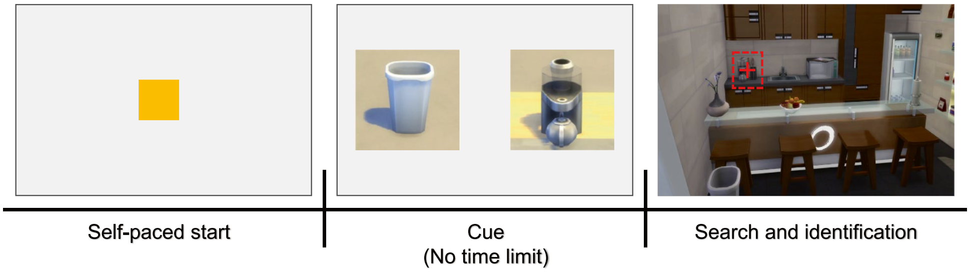

**Figure 1** An example trial. Participants initiate self-paced start with a central cue. A number of objects (two objects are shown as example) will then be presented as search cue. Participants will then search for and identify the targets (presented in search cue stage) within a scene with both eye-fixation and keyboard-press responses.

**Figure 2** Three representative scenes (using kitchens as example) taken from our simulated scenes.

## 5.4 Eye Tracking Paradigm on Non-human Primates

In addition to comparison with humans, we provided a further example to illustrate how eye tracking could be applied on non-human primates. In this study, we trained macaque monkeys to view sets of still images and after a delay, to choose from three choices the one that they had viewed previously [94]. In the current context, we would then analyse the physiological responses such as saccadic scan-paths, fixations, and even pupil dilation when the animals view and process these still images [95] and compare with the computation model presented in Section 5.5.

An example trial is shown in Fig. 3: the blue trace depicts the real saccadic scan-path whereas the yellow trace depicts the shortest distance between two fixations. The blue circles refer to fixations and their associated numerals represent

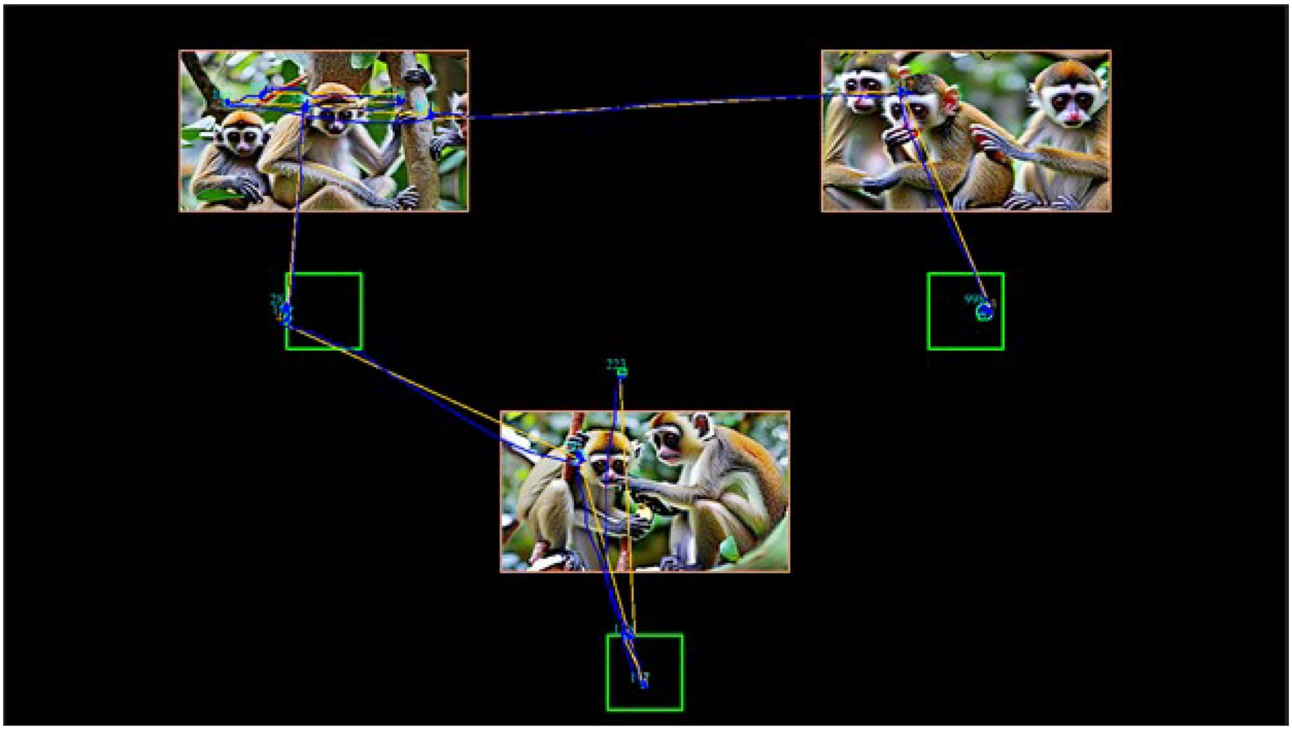

**Figure 3** Eye movement tracking for macaque monkeys.

duration of fixations (in ms). This image in Fig. 3 shows the memory test stage for 3-alternative forced choice recognition memory of a trial (the encoding stage is not shown here). These three test images were created using DreamStudio AI [96].

## 5.5 Computational Model for Scene Category Recognition

The pre-processing stages of the model will utilise state-of-the-art deep neural networks for the target or object detection, which will be followed by the construction of the three unique vectors (spatial, size, pair-wise relationships) that will be learned for known scenes. The emphasis here is on the use of spiking neuron; much closer to the human visual system to take advantage of the minimal power consumption. The proposed model will involve four major activities:

1. Design of a model that goes beyond object detection and identification;
2. the introduction of a real-world and novel dataset that can be used to justify the model;
3. comparing human search performance with other deep learning approaches trained with large-scale images as well as our model for object and scene identification;
4. and finally making the architecture deployable on a low power neuromorphic hardware.

### 5.5.1 Design of a Model that goes Beyond Object Detection and Identification

This aspect of our model has four different tasks. The first will involve the use of a pre-trained convolutional neural network (deep learning) to identify all known objects in a scene. The input to the pre-trained CNN model will be series of images (scenes) representing normal day environment at home (Fig. 2), office and on the streets. The main aim here is to identify all known objects in the scene automatically, using the deep learning architecture. The second task is to group all common objects in the scene. For example, images of an office will normally consist of a desk, chairs, a keyboard, a monitor and other related objects. The collection of objects in the scene will then be used to form an object-set, containing all prominent objects commonly found in the defined environment. A threshold will be used to classify an object as part of the object-set for any given environment; thus, an object will have to appear in a specified number of scenes to be counted as part of that object-set. The third task will use the object-set to generate a vector that represents the spatial relationship between any two objects. The spatial relationship between any two objects will include their relative position, distance and orientation. Part of these measures will be acquired during the process of object-set generation. For every pair of objects in an object-set, a three-dimensional vector will be generated to represent the extremal as well as the average values. The fourth and final task will involve the generation of size relationship. Like the spatial relationship this will be generated for each pair of objects in each object-set, making use of their combined aspect ratio as well as scalar invariant features like Histogram of Oriented Gradients (HOG), Speeded-Up Robust features (SURF) and Binary Robust Independent Elementary Features (BRIEF).

### 5.5.2 Introduction of a Real-World and Novel Dataset that can be Used to Justify the Model

This aspect of the model will involve building a database of images that will be used to train, test and verify the model. The images will include indoor and outdoor scenes with typical objects found in that environment. To avoid the biases in most of the available datasets like IMAGENET and CAFFE, used for training deep CNN architectures, this work considers a collection of normal scene images to design a data-acquisition protocol for visual scene understanding in self-driving and surveillance systems. Synthetic images will also be generated to represent atypical scenes for training purposes. Like IMAGENET, available images on the internet with common themes to that of scenes being tested in this work will also be used.

### 5.5.3 Architectural Comparison

The last aspect of the model will involve the comparison of the model designed in this work with human search capabilities as well as off-the-shelf deep learning models (cafenet, VGG-16, GoogleNet, ResNet and Yolo2) trained on off-line large-scale images. The comparison is mainly to show how humans and deep learning architectures interpret scenes with varying objects in terms of size and position. These comparisons will also evaluate the global understanding of the scene to infer possible actions and identify any anomalies.

### 5.5.4 Neuromorphic Computing

As a specialised hardware designed mainly to mimic the structure and functionality of the human brain, the target implementation in this work is to perform task that involve processing information in ways similar to the human's brain neural networks function. The human brain is known to be highly efficient [97] at processing complex information, recognising visual patterns, and adapting to new situations. The traditional von Neumann architecture is not optimised for recognising or interpreting scenes and can be relatively power-hungry and slow when it comes to certain types of computations like pattern recognition and sensory processing.

Neuromorphic computing generally aims to address the limitations of the von Neumann architecture by designing hardware architectures inspired by the brain's structure and functionality. Neuromorphic architectures often involve large numbers of simple processing units (neurons) that are interconnected and can communicate with each other. The connections, similar to synapses in the brain, allow for the transmission of signals and the formation of networks that can adapt over time based on experience. Such an architecture, efficient in processing visual information is the target for the proposed scene understanding system.

## 6 Concluding Remark

In this paper, we have reviewed recent advances in deep learning architecture that have taken inspiration from human or primate learning and visual systems, and provided direction to future advancement on deep learning with inspiration from physiological experiments. Upon a review of areas that have benefited from deep learning, we specifically outline a physiologically inspired model for scene understanding that encodes three key components: object location, size and category. Human vision understanding can serve as a valuable source of inspiration for bio-inspired computer vision and for that matter bio-inspired AI. Through an effective emulation of the mechanisms and principles underlying human visual perception, bio-inspired computer vision can aim to achieve similar levels of performance and robustness as a human, when it comes to scene understanding. For example, the selective nature of human vision can be incorporated into bio-inspired AI to prioritise important features or regions in an image and effectively reduce computational cost. Similarly, human vision integrates contextual information like scene layout, object relationship and semantic context to make sense of visual scenes; attributes that can enhance scene understanding but hard to model into existing vision systems. The model proposed in this work goes beyond simple object detection and identification, it aims to introduce a novel real-world dataset, and ultimately provide a comparison between how humans and deep learning architectures interpret complex, naturalistic scenes.

**Author Contributions** The idea and direction of the work was defined by Appiah and Kwok, and the initial psychological experiments have been conducted by Kwok, Shi and Jin.

**Funding** The research results of this publication are sponsored by the Kunshan Municipal Government research funding Project Number: 23KKSGR017.

## Declarations

**Ethical Approval** The human and animal experiments included in this work have been approved by the Institutional Review Board (permission code: HR 295-2019) and the Institutional Animal Care and Use Committee (permission code M020150902-2018) at East China Normal University. All other datasets used in this work are publicly available or simulated.

**Conflict of Interest** The work was initiated and completed by the four named authors. There has not been any funding from a third party.

## References

1. Wiriyathammabhum, P., Summers-Stay, D., Fermüller, C., & Aloimonos, Y. (2017). Computer vision and natural language processing: Recent approaches in multimedia and robotics. *ACM Computing Surveys, 49*(4), 1–44. https://doi.org/10.1145/3009906
2. Coughlin, T. (2018). Digital storage in smartphones and wearables [the art of storage]. *IEEE Consumer Electronics Magazine, 7*(2), 108–120. https://doi.org/10.1109/MCE.2017.2773361
3. Hazelwood, K. (2018). Applied machine learning at facebook: A datacenter infrastructure perspective. *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 620–629. https://doi.org/10.1109/HPCA.2018.00059
4. Correll, N. (2018). Analysis and observations from the first amazon picking challenge. *IEEE Transactions on Automation Science and Engineering, 15*(1), 172–188. https://doi.org/10.1109/TASE.2016.2600527
5. Robert, J. A., & Breakspear, M. (2018). Synaptic assays: using biophysical models to infer neuronal dysfunction from non-invasive eeg. *Brain, 141*, 1583. https://doi.org/10.1093/brain/awy136
6. Martin, V., Séguier, R., Porcheron, A., & Morizot, F. (2018). Towards continuous health diagnosis from faces with deep learning. *Lecture Notes in Computer Science, 11121*, 1583. https://doi.org/10.1093/brain/awy136
7. Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine, 13*(3), 55–75. https://doi.org/10.1093/brain/awy136
8. Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., & Quillen, D. (2017). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research, 37*(4–5), 1–44. https://doi.org/10.1145/3009906
9. Michelsanti, D., Guichi, Y., Ene, A., Stef, R., Nasrollahi, K., & Moeslund, T. (2017). Fast fingerprint classification with deep neural network. *Visapp - International Conference on Computer Vision Theory and Applications*, 202–209. https://doi.org/10.5220/0006116502020209

10. Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1701–1708.
11. Xia, L., Luo, J., Sun, Y., & Yang, H. (2018). Deep extraction of cropland parcels from very high-resolution remotely sensed imagery. 2018 7th International Conference on Agrogeoinformatics (Agro-geoinformatics).
12. Maggiori, E., Tarabalka, Y., Charpiat, G., & Alliez, P. (2017). Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing, 55*(2), 645–657. https://doi.org/10.1109/TGRS.2016.2612821
13. Firdaus, Arkeman, Y., Buono, A., & Hermadi, I. (2017). Satellite image processing for precision agriculture and agroindustry using convolutional neural network and genetic algorithm. *IOP Conference Series: Earth and Environmental Science*, 54.
14. Nardari, G. V. (2018). Crop anomaly identification with color filters and convolutional neural networks. *2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE)*.
15. Thomas, G., Gade, R., Moeslund, T. B., Carr, P., & Hilton, A. (2017). Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding, 159*, 3–18.
16. Martinez, P., Ahmad, R., & Al-Hussein, M. (2019). A vision-based system for pre-inspection of steel frame manufacturing. *Automation in Construction, 97*, 151–163.
17. Xu, P., Dherbomez, G., Hery, E., Abidli, A., & Bonnifait, P. (2018). System architecture of a driverless electric car in the grand cooperative driving challenge. *IEEE Intelligent Transportation Systems Magazine, 10*(1), 47–59. https://doi.org/10.1109/MITS.2017.2776135
18. Gallardo, N., Gamez, N., Rad, P., & Jamshidi, M. (2017). Autonomous decision making for a driver-less car. 2017 12th System of Systems Engineering Conference (SoSE).
19. Zhang, Y., & Yuen, K.-V. (2022). In: *Cury, A., Ribeiro, D., Ubertini, F., Todd, M.D. (eds.) Applications of Deep Learning in Intelligent Construction*, pp. 227–245. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-81716-9_11
20. Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019). Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management. CIKM '19*, pp. 1441–1450. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3357384.3357895
21. Zhu, Q., & Luo, J. (2022). Generative pre-trained transformer for design concept generation: An exploration. *Proceedings of the Design Society, 2*, 1825–1834. https://doi.org/10.1017/pds.2022.185
22. You, C., Xiang, J., Su, K., Zhang, X., Dong, S., Onofrey, J., Staib, L., & Duncan, J. S. (2022). Incremental learning meets transfer learning: Application to multi-site prostate mri segmentation. In: *Albarqouni, S., Bakas, S., Bano, S., Cardoso, M. J., Khanal, B., Landman, B., Li, X., Qin, C., Rekik, I., Rieke, N., Roth, H., Sheet, D., Xu, D. (eds.) Distributed, Collaborative, and Federated Learning, and Affordable AI and Healthcare for Resource Diverse Global Health*, pp. 3–16. Springer, Cham.
23. Huang, Y. (2020). In: *Dong, H., Ding, Z., Zhang, S. (eds.) Deep Q-Networks*, pp. 135–160. Springer, Singapore. https://doi.org/10.1007/978-981-15-4095-0_4
24. Marchesini, E., & Farinelli, A. (2022). Enhancing deep reinforcement learning approaches for multi-robot navigation via single-robot evolutionary policy search. In: 2022 International Conference on Robotics and Automation (ICRA), pp. 5525–5531. https://doi.org/10.1109/ICRA46639.2022.9812341
25. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117.
26. Li, Z., & Hoiem, D. (2018). Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(12), 2935–2947. https://doi.org/10.1109/TPAMI.2017.2773081
27. Li, P., Chen, Z., Yang, L. T., Zhang, Q., & Deen, M. J. (2018). Deep convolutional computation model for feature learning on big data in internet of things. *IEEE Transactions on Industrial Informatics, 14*(2), 1–44. https://doi.org/10.1109/TII.2017.2739340
28. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural network. *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 1, 1097–1105.
29. Bartunov, S., Santoro, A., Richards, B., Marris, L., Hinton, G. E., & Lillicrap, T. (2018). Assessing the scalability of biologically-motivated deep learning algorithms and architectures. *Proceedings of Advances in Neural Information Processing Systems*, 31, 9390–9400.
30. Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*.
31. Yann, L., Yoshua, B., & Geoffrey, H. (2015). Deep learning. *International Journal of science: Nature*, 521. https://doi.org/10.1038/nature14539
32. Dundar, A., Jin, J., Martini, B., & Culurciello, E. (2017). Embedded streaming deep neural networks accelerator with applications. *IEEE Transactions on Neural Networks and Learning Systems, 28*(7), 1572–1583. https://doi.org/10.1109/TNNLS.2016.2545298
33. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition, 77*, 354–377.
34. Shawahna, A., Sait, S., & El-Maleh, A. (2019). Fpga-based accelerators of deep learning networks for learning and classification: A review. *IEEE Access*, 7823–7859. https://doi.org/10.1109/ACCESS.2018.2890150
35. YouTubeStatistics. (2019). https://youtube.com/yt/press/statistics.html
36. FacebookStatistics. (2019). https://zephoria.com/top-15-valuable-facebook-statistics/
37. Crawford, C. H., Henning, P., Kistler, M., & Wright, C. (2008). Accelerating computing with the cell broadband engine processor. In *Proceedings of the 5th conference on Computing frontiers (CF '08)*, 3–12. https://doi.org/10.1145/1366230.1366234
38. Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 49*(4), 427–436.
39. Wiriyathammabhum, P., Summers-Stay, D., FermÜller, C., & Aloimonos, Y. (2017). Computer vision and natural language processing: Recent approaches in multimedia and robotics. *ACM Computing Surveys, 49*(4), 1–44. https://doi.org/10.1145/3009906
40. Daeyeol, L., Hyojung, S., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience, 35*(1), 287–308. https://doi.org/10.1146/annurev-neuro-062111-150512
41. Silver. (2017). Mastering the game of go without human knowledge. *Nature, 550*(254). https://doi.org/10.1038/nature24270
42. Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60–88.

43. Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
44. Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D., Bernstein, M., & Fei-Fei, L. (2015). Image retrieval using scene graphs. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3668–3678.
45. Song, S., Yu, F., Zeng, A., Chang, A. X., Savva, M., & Funkhouser, T. (2017). Semantic scene completion from a single depth image. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*
46. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV).*
47. Yang, S., Wang, W., Liu, C., & Deng, W. (2019). Scene understanding in deep learning-based end-to-end controllers for autonomous vehicles. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49*(1), 53–63. https://doi.org/10.1109/TSMC.2018.2868372
48. Kohli, P., & Chadha, A. (2018). Enabling pedestrian safety using computer vision techniques: A case study of the 2018 uber inc. self-driving car crash.
49. Zhou, W., Zyner, A., Worrall, S., & Nebot, E. (2019). Adapting semantic segmentation models for changes in illumination and camera perspective. *IEEE Robotics and Automation Letters, 4*(2), 461–468. https://doi.org/10.1109/LRA.2019.2891027
50. Vincent, K. (2018). Ethical implications: The ACM/IEEE-CS software engineering code applied to tesla's
51. Norton, A., Ober, W., Baraniecki, L., Shane, D., Skinner, A., & Yanco, H. (2018). Perspectives on human-robot team performance from an evaluation of the darpa robotics challenge. *Springer Tracts in Advanced Robotics, 121.*
52. Barbara, C., & Bernward, J. (2018). Robotization of work as presented in popular culture, media and social sciences. *Research Report - Gothenburg Research Institute*.
53. Droniou, A., Ivaldi, S., & Sigaud, O. (2015). Deep unsupervised network for multimodal perception, representation and classification. *Robotics and Autonomous Systems, 71*, 83–98.
54. Chen, X., Ghadirzadeh, A., Folkesson, J., Björkman, M., & Jensfelt, P. (2018). Deep reinforcement learning to acquire navigation skills for wheel-legged robots in complex environments. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).
55. Robicquet, A., Alahi, A., Sadeghian, A., Anenberg, B., Doherty, J., Wu, E., & Savarese, S. (2016). Forecasting social navigation in crowded complex scenes.
56. Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage, 152*, 184–194.
57. Peelen, M. V., & Downing, P. E. (2017). Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia, 105*, 177–183.
58. Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., Raju, B., Shahrzad, H., Navruzyan, A., Duffy, N., & Hodjat, B. (2017). Computer vision and natural language processing: Recent approaches in multimedia and robotics. *ACM Computing Surveys, 49*(4), 1–44. https://doi.org/10.1145/3009906
59. Zheng, L., Yang, Y., & Tian, Q. (2018). Sift meets cnn: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(5), 1224–1244. https://doi.org/10.1109/TPAMI.2017.2709749
60. Zhao, Y., Zhang, Y., Cheng, R., Wei, D., & Li, G. (2015). An enhanced histogram of oriented gradients for pedestrian detection. *IEEE Intelligent Transportation Systems Magazine, 7*(3), 29–38. https://doi.org/10.1109/MITS.2015.2427366
61. Chen, C., Jafari, R., & Kehtarnavaz, N. (2015). Action recognition from depth sequences using depth motion maps-based local binary patterns. *2015 IEEE Winter Conference on Applications of Computer Vision.*
62. Ye, Q., Hu, Y., Ku, Y., Appiah, K., & Kwoki, S. C. (2018). Locally distributed abstraction of temporal distance in human parietal cortex. bioRxiv The Preprint Server for Biology. https://doi.org/10.1101/249904
63. Wiriyathammabhum, P., Summers-Stay, D., Fermüller, C., & Aloimonos, Y. (2017). Computer vision and natural language processing: Recent approaches in multimedia and robotics. *ACM Computing Surveys, 49*(4), 1–44. https://doi.org/10.1145/3009906
64. Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex[j]. *The Journal of Physiology, 160*(1), 106–154.
65. Yann, L. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation, 1*(4), 541–551.
66. Christian, S., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2017). Going deeper with convolutions. *ACM Computing Surveys, 49*(4), 1–44. https://doi.org/10.1145/3009906
67. Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Reviews, 94*(2), 115–147. https://doi.org/10.1037/0033-295X.94.2.115
68. Turner, M. H., Giraldo, L. G. S., Schwartz, O., & Rieke, F. (2019). Stimulus- and goal-oriented frameworks for understanding natural vision. *Nature Neuroscience, 22*(1), 15–24. https://doi.org/10.1038/s41593-018-0284-0
69. Aravindh, M., & Vedaldi, A. (2015). Understanding deep image representations by inverting them. https://doi.org/10.1145/3009906
70. Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks[c]. *European Conference on Computer Vision.*
71. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database[c]. Advances in neural information processing systems, 487–495.
72. Van Essen, D. C., Anderson, C. H., & Felleman, D. J. (1992). Information processing in the primate visual system: an integrated systems perspective[j]. *Science, 255*(5043), 419–423.
73. Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks[c]. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249–256.
74. Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system[j]. *Nature, 381*(6582), 520.
75. Zhang, S., Gong, Y., Wang, J., & Zheng, N. (2016). A biologically inspired deep CNN model. *17th Pacific-Rim Conference on Advances in Multimedia Information Processing*, 9916, 540–549. https://doi.org/10.1007/978-3-319-48890-5_53
76. Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience, 39*(33), 7255–7269. https://doi.org/10.1523/JNEUROSCI.0388-18.2018
77. Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology: General, 145*(1), 82–94. https://doi.org/10.1037/xge0000129
78. Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *Elife Journal*, 7. https://doi.org/10.7554/eLife.32962

79. Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., & Rust, N. C. (2019). A neural correlate of image memorability. bioRXiv. https://doi.org/10.1101/535468
80. Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience, 35*(39), 13402–13418. https://doi.org/10.1523/JNEUROSCI.5181-14.2015
81. Ullman, S. (2019). Using neuroscience to develop artificial intelligence. *Science, 363*, 692–693. https://doi.org/10.1126/science.aau6595
82. Redmon, J., Divvala, S. K., Girshick, R. B., & Farhad, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 779–788.
83. Hao, S., Zhou, Y., & Guo, Y. (2020). A brief survey on semantic segmentation with deep learning. *Neurocomputing, 406*, 302–321. https://doi.org/10.1016/j.neucom.2019.11.118
84. Muhammad, K., Hussain, T., Ullah, H., Ser, J. D., Rezaei, M., Kumar, N., Hijji, M., Bellavista, P., & de Albuquerque, V. H. C. (2022). Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. *IEEE Transactions on Intelligent Transportation Systems, 23*(12), 22694–22715. https://doi.org/10.1109/TITS.2022.3207665
85. Gaurav, R., Tripp, B., & Narayan, A. (2022). Spiking approximations of the maxpooling operation in deep snns. In: *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892504
86. Davies, M., Srinivasa, N., Lin, T.-H., Chinya, G., Cao, Y., Choday, S. H., Dimou, G., Joshi, P., Imam, N., Jain, S., Liao, Y., Lin, C.-K., Lines, A., Liu, R., Mathaikutty, D., McCoy, S., Paul, A., Tse, J., Venkataramanan, G., … Wang, H. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro, 38*(1), 82–99. https://doi.org/10.1109/MM.2018.112130359
87. Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology, 27*, 2827–2832.
88. Izadinia, H., Sadeghi, F., & Farhadi, A. (2014). Incorporating scene context and object layout into appearance modeling. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 232–239
89. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 48*(6), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009
90. Segraves, M. A., Kuo, E., Caddigan, S., Berthiaume, E. A., & Kording, K. P. (2017). Predicting rhesus monkey eye movements during natural-image search. *Journal of Vision, 17*(12). https://doi.org/10.1167/17.3.12
91. Shepherd, S. V., Steckenfinger, S. A., Hasson, U., & Ghazanfar, A. A. (2010). Human-monkey gaze correlations reveal convergent and divergent patterns of movie viewing. *Current Biology, 20*(7), 649–656. https://doi.org/10.1016/j.cub.2010.02.032
92. Berg, D. J., Boehnke, S. E., Marino, R. A., Munoz, D. P., & Itti, L. (2009). Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision, 9*(19). https://doi.org/10.1167/9.5.19
93. RESEARCH, S. EyeLink [online]. https://www.sr-research.com/products/eyelink-1000-plus/
94. Urgolites, Z. J., Smith, C. N., & Squire, L. R. (2018). Eye movements support the link between conscious memory and medial temporal lobe function. *Proceedings of the National Academy of Sciences*, *115*(29), 7599–7604. https://www.pnas.org/doi/pdf/10.1073/pnas.1803791115. https://doi.org/10.1073/pnas.1803791115
95. Wang, L., Zhou, X., Yang, J., Zeng, F., Zuo, S., Kusunoki, M., Wang, H., Zhou, Y.-D., Chen, A., & Kwok, S. C. (2022). Mixed selectivity coding of content-temporal detail by dorsomedial posterior parietal neurons. *BioRxiv: The Preprint Server for Biology*. https://doi.org/10.1101/2022.07.16.500237
96. DreamStudio: Early Access to SDXL. https://beta.dreamstudio.ai/generate. Accessed 27 Jun 2023.
97. Majumdar, S., Tan, H., Qin, Q. H., & van Dijken, S. (2019). Energy-efficient organic ferroelectric tunnel junction memristors for neuromorphic computing. *Advanced Electronic Materials, 5*(3), 1800795. https://doi.org/10.1002/aelm.201800795