Modeling Visual Impairments with Artificial Neural Networks: a Review

Lucia Schiatti [1,2,*], Monica Gori[2], Martin Schrimpf[3], Giulia Cappagli [2], Federica Morelli[4,5], Sabrina Signorini[4], Boris Katz[1] and Andrei Barbu[1]

[1] CSAIL & CBMM, Massachusetts Institute of Technology, USA

[2] UVIP, Istituto Italiano di Tecnologia, Italy

[3] NeuroX Institute, EPFL, Lausanne, Switzerland

[4] IRCCS Mondino Foundation, Italy

[5] DBBS, University of Pavia, Italy

[*] schiatti@mit.edu

Abstract

We present an approach to bridge the gap between the computational models of human vision and the clinical practice on visual impairments (VI). In a nutshell, we propose to connect advances in neuroscience and machine learning to study the impact of VI on key functional competencies and improve treatment strategies. We review related literature, with the goal of promoting the full exploitation of Artificial Neural Network (ANN) models in meeting the needs of visually impaired individuals and the operators working in the field of visual rehabilitation. We first summarize the existing types of visual issues, the key functional vision-related tasks, and the current methodologies used for the assessment of both. Second, we explore the ANNs best suitable to model visual issues and to predict their impact on functional vision-related tasks, at a behavioral (including performance and attention measures) and neural level. We provide guidelines to inform the future research about developing and deploying ANNs for clinical applications targeting individuals affected by VI.

1. Introduction

Vision is the principal sensory modality through which humans collect information about the world. A visual impairment can therefore have a huge impact on a person's life. The WHO [1] distinguishes among aspects of a vision lack related to changes at the organ level, i.e. anatomical, and at the person level, i.e. social and economical consequences of reduced abilities. Notably, impairments to an organ do not imply the person's complete loss of the ability to perform Activities of Daily Living (ADLs). Indeed, the majority of people with impaired vision are not blind, but in fact have residual vision [2]. The assessment of a person's ability to perform ADLs is often referred to, in clinical settings, as ”functional vision”, and it is at the heart of visual re-habilitation's interventions. The goal of visual rehabilitation is to manipulate environmental, medical, and human factors in order to minimize the negative effect of a disorder on functional vision, and ultimately to improve the person's participation in society and social life [3]. This is particularly true at an early age, when the neurological maturation is ongoing and functional abilities are developing, to prevent the onset of developmental delays and intellectual disabilities [4, 5, 6]. Visual training programs aim at teaching children how to make a functional use of their sense of sight. They include, for instance, training to improve efficient fixation, visual following of moving targets, and eye-hand coordination [7].

The majority of visual assistive technologies for people with VI, including applications of computer vision and deep learning, have been developed based on the ”sensory-substitution” paradigm, i.e. to replace the human functionality while performing a specific task (e.g. navigation, object detection) [8, 9]. However, this approach disregards the high potential of modern machine learning techniques to be deployed within visual rehabilitation settings, and it neglects neuroscientific findings [10, 11]. Indeed, for rehabilitative purposes, assistive technologies should rather enhance the person's capability to access information from the external environment exploiting his/her own resources, including residual vision [12].

Compared to previous works, which focused on a task-oriented classification of models for visual assistive technologies, this review offers an alternative perspective, rooted on neuroscience, to categorize existing ANNs. We provide an overview about how these models can be used to develop tools to understand visual impairments, and to predict their effect on vision-related behaviors, thus supporting visual assessment and rehabilitation. Indeed, current ANNs solve specific vision-related tasks, which are also relevant in the context of visual rehabilitation (e.g. reading, object detection, object recognition, face recognition), and their internal representations resemble - to a first extent - primate neural recordings [13, 14]. On the other hand, for clinical and assistive applications of machine learning and computer vision, it might be important to have models that are neurally aligned. Indeed, a human-model similarity plays in favor of ANNs' likelihood to predict the effect of perceptive, oculomotor and/or cognitive issues on how the person solves a specific functional task.

The review is organized as follows: in Section 2 we provide a categorization of VI based on how they affect visual functions (2.1), we identify the main vision-related (functional vision) skills involved in visual rehabilitation (2.2), and the main techniques used for the assessment of both visual functions and functional skills of subjects with VI (2.3). In Section 3 we review existing ANNs suitable to model behavioral outcomes (based on direct measures, 3.1, and gaze data, 3.2), and neural outcomes (3.3) of humans during vision-related functional tasks. In Section 4 we devise future research directions to use ANNs to model VI.

2. Characterization of visual impairments

Although it is a common notion that vision refers to the basic functioning of eyes, the visual process actually involves the integration among different structures within the visual system, including the eyes, the visual pathway, the visual cortex, and other brain or cortical areas. Vision occurs when all components of the system are intact and functioning [15, 16]. Modern visual assessment and rehabilitation paradigms rely on the distinction of two intercorrelated components of vision: visual functions, i.e. the functioning of anatomical organs (eyes and visual system), and functional vision, which refers to the ways in which a person "functions" in activities that are normally vision-dependant [17], such as reading, orientation and mobility, object recognition and social interaction. In the following paragraphs, we provide an overview of the different types of VI, their effect on the main functional skills, and existing (clinical and functional) assessment techniques (Fig. 1).

2.1. Visual functions

A description of the nature of VI from an anatomical point of view, as well as of the pathologies leading to visual issues, is out of the scope of this review. Here, we aim at describing the effect that VI can have on measurable outcomes related to functional abilities, because we view those as most readily tackled by current computational techniques. Starting from this idea, we identify the types of VI according to the distinction adopted in [18]: *peripheral*, *oculomotor*, and *cerebral* issues (Fig. 1).

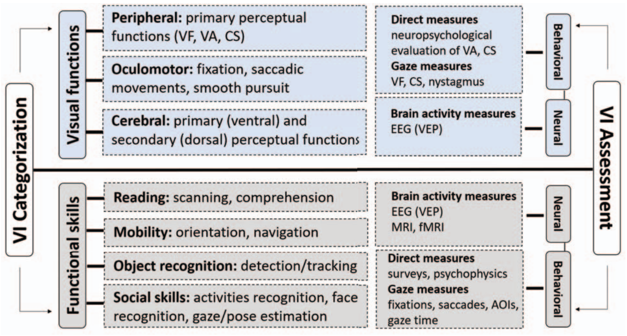
VI Categorization
Visual functions
Peripheral: primary perceptual functions (VF, VA, CS)
Oculomotor: fixation, saccadic movements, smooth pursuit
Cerebral: primary (ventral) and secondary (dorsal) perceptual functions
Functional skills
Reading: scanning, comprehension
Mobility: orientation, navigation
Object recognition: detection/tracking
Social skills: activities recognition, face recognition, gaze/pose estimation
Direct measures
neuropsychological evaluation of VA, CS
Gaze measures
VF, CS, nystagmus
Brain activity measures
EEG (VEP)
Brain activity measures
EEG (VEP)
MRI, fMRI
Direct measures
surveys, psychophysics
Gaze measures
fixations, saccades, AOIs, gaze time
Behavioral
Neural
Neural
Behavioral
VI Assessment

Figure 1. **Left-top**: types of VI, according to which visual functions are affected; **Left-bottom**: main functional skills affected by a vision's malfunctioning in ADLs; **Right**: assessment techniques of visual functions and functional skills in subjects with VI.

Specifically, such a distinction relies on the (simplified) identification of three main functions of vision that are important to predict a functional outcome; in [18], these are described as "seeing", "looking" and "understanding", and can be isolated and assessed separately. According to this framework, **peripheral** issues are those related to the peripheral structures of the primary visual pathways (eye, optical nerve), and they directly affect the perceptual primary component of vision ("seeing"), including functions such as visual acuity, visual field, contrast sensitivity. **Oculomotor** issues concern the oculomotor system and its functions (fixations, saccades, smooth pursuit) and they affect the explorative component of vision ("looking"). Finally, **cerebral** issues refer to those generated by damages to the dorsal and ventral streams [19], which affect the visuo-cognitive abilities ("understanding"). Such categorization is not rigid, since visual issues, regardless of their origin, can manifest themselves at multiple levels of vision functioning. Cerebral issues can be associated with visual malfunctioning at primary perceptual (visual acuity and visual field) and oculomotor level (optokinetic nystagmus, i.e. instability of fixations) [20]. Generally, peripheral and oculomotor impairments, even when not arising from cortical damages, can affect the neuropsychological, social and cognitive development, as suggested by a body of works investigating, for instance, the (controversial) relationship between visual impairments and autistic-like behaviors [21].

Nevertheless, the distinction among peripheral, oculomotor and cerebral issues, based on the clinical assessment of visual functions, is useful to simulate these impairments within ANNs architectures, and test their differential effects on functional capabilities of individuals with VI.

2.2. Functional skills

Among functional skills that are affected by a vision deficit, visual rehabilitation focuses on different categories of competencies that are fundamental to enable a person's relationship with the surrounding environment and with other people, including: *reading*, *mobility*, *object recognition* and *social skills*.

Reading and mobility are two of the main skills addressed by visual rehabilitation training protocols [22], given their obvious importance in activities of daily life. Peripheral impairments, especially those related to central visual field defects, are those impairing **reading skills** the most. Rehabilitation includes interventions aimed at teaching how to use reading aids, and training for learning possible compensatory behaviors (e.g. eccentric fixation) [22].

Since visual information is fundamental for spatial processing, various studies have shown how the absence of vision impacts the development of **locomotor skills** [23, 24]. In [10], the authors present an extensive review about orientation (i.e. the ability to understand the spatial properties of the environment and its relationships with one's position) and mobility (i.e. the capability of efficiently and safely moving in an environment) in adults and children with and without visual disabilities. The authors discuss the link between the observed reduced orientation and mobility in children with VI and the related processes in which the visual modality is involved during development.

Many studies have documented the development of higher cognitive functions during childhood and adolescence, and they tried to correlate it with the processes of brain maturation [25]. It is known that visual abilities subtended by the primary visual cortex, such as simple shape discrimination, are already present at the age of 6 y.o., while higher visual abilities, such as visual **object recognition**, continue to develop later in childhood [26]. Given their nature, cerebral VI have been studied to investigate the link between brain damages and an impaired visual perception in such high-level tasks. Although some results pointed towards the hypothesis of specific visual perceptual impairments (object recognition) in children with cerebral VI [27], the research outcomes are inconclusive because of limited samples of examined subjects, the comorbidity with developmental delays, and a lack of standardized tests [28]. Oculomotor issues, on the other hand, may affect dynamic tasks such as object tracking and visual search, due to difficulties in holding fixations or performing saccadic movements.

The development of spatial cognition, locomotion skills, as well as higher perceptive visual function (including face recognition) are strictly related to the development of **social cognition** [11]. Because of an impaired capability of detecting social cues, i.e. body gestures and non verbal communication signals, children with VI may face difficulties in engaging in positive social interactions [29]. They demonstrate a low peer-related social competence [30], they do not display a full range of play behaviors, and they spend more time in solitary playing or interacting with adults than with their peers [31, 32]. In [33], the authors tested patients treated for bilateral congenital cataracts, and they found that early visual deprivation affects the development of face recognition. However, it seems unclear to which extent a lack of visual experience can affect the capability of producing emotional facial expressions [34].

2.3. Assessment techniques

The assessment of both visual functions (how the eye and the visual system function) and functional vision (how the person ”functions” on visual tasks) is the core of visual rehabilitation [3]. The relationship between these two allows to plan appropriate interventions and to verify the efficacy of a training protocol over time. As reported in Fig. 1, we take into account **behavioral** and **neural** measurements. Among behavioral measures, we distinguish among *”direct”* measures, i.e. task-related answers actively provided by the subject or performance measures, and *”gaze”* measures, i.e. indirect task-related measures based on eye movements. In the following sections, we review the current application of these techniques to the assessment of either visual functions (perceptual, oculomotor and visuo-cognitive) or functional vision-related skills (i.e. reading, mobility, object recognition and social interaction).

Visual Functions' Assessment: Visual and oculomotor functions are assessed by clinical tests, carried out by professionals. Such measures can be used both to assist in the diagnosis of the underlying disorder and to predict the functional consequences. Visual acuity (VA) is the most common metric for quantifying a subject's global visual functions. Indeed, even if a low VA can result from different disorders, it provides a good indicator about the impact of the person's ability to perform ADLs, e.g. reading [3]. Other visual functions commonly assessed include visual field (VF) and contrast sensitivity (CS) [35]. These metrics can be measured either with behavioral methods (e.g. reading chart for VA), where the subject needs to provide a verbal or behavioral answer to a stimulus, or neural methods (e.g. neural activity recordings) [36]. Eye-tracking techniques have been explored as an alternative to traditional visual functions assessment methods based on clinical evaluations, to provide a quantitative assessment both for primary perceptual functions (e.g. visual field, contrast sensitivity) and oculomotor functions (e.g. nystagmus, fixation, saccades, smooth pursuit) [37, 38]. Although the vast majority of studies assess visual functions through behavioral methods, neural methods such as pattern steady-state visual evoked potentials (VEP) are used in case of infants or cognitively impaired subjects. Since cerebral issues can cause defects in higher visual perceptual functions (i.e. there can be an impaired perceptual functioning in presence of normal acuity), neuro-imaging techniques are more appropriate to characterize visual functions in such cases [39].

Functional Vision Assessment: Functional vision is assessed by the ability to perform generic ADLs. The functional vision is typically assessed in two ways: 1) by developing systems for correlating visual functions' measurements to statistical estimates of functional abilities, or 2) by directly assessing such individual abilities. A recent example of the first approach is shown in [18], where the authors propose a novel protocol for visual functions' quantitative evaluation based on professional reporting (Visual Function Score). The system provides a global score of a subject's visual functioning, useful to monitor rehabilitation outcomes. The second approach is usually qualitative and takes the form of questionnaires in which the subject, the family and/or clinicians rate the ability to perform a series of ADLs [40, 41, 42]. Functional skills of subjects with VI have been also evaluated quantitatively, for instance, to assess the effectiveness of image enhancement techniques during reading, object detection/recognition, face/emotion recognition [43], or to determine the correlation between measures of VA and CS and the performance on ADLs, e.g. mobility speed and reading speed [44]. The assessment of various aspects of the social development on people with VI lacks of a unified definition of "social skills", and it has been mainly based on qualitative behavioral observations and questionnaires [45]. In contrast, quantitative behavioral approaches to eye analysis and gaze tracking have been widely used to investigate different aspects of social attention and social skills in children with Autism Spectrum Disorders [46]. Nevertheless, to our knowledge eye tracking has not yet been used to assess functional skills of people with VI, except one study investigating visual information processing in children [47]. Likewise, no eye tracking study investigated the onset of atypical behavioral patterns (i.e. the correlation between visual peripheral/oculomotor impairments and the emerging of autistic-like features [21]). Finally, neural measures (fMRI) were mainly used to investigate the relationship between visual functions (including object recognition, face recognition, visual memory, orientation, visual spatial perception, and motion perception) and cerebral visual impairments [28]. In a few cases, neural measures have been used to assess the functional outcome of visual training, e.g. in [48], VEP were used to assess the effectiveness of visual rehabilitation in improving obstacle detection skills of a child with severe VI.

3. ANNs to model visual impairments

Visual rehabilitation and computer vision share similar goals: improving human and model performance respectively, on a set of visual (functional) tasks. In this section, we review different ANN models, and we discuss the properties that make them good candidates to model VI, i.e. to predict behavioral (including direct and gaze measures) and neural human outcomes on visual tasks.

The conceptual framework underlying this approach is inspired by the work in [13, 49], and shown in Fig. 2. It is based on defining a common framework for the evaluation of both human and models' performance on key functional vision-related tasks, by identifying suitable and standardized experimental protocols and stimuli datasets. The goal is to generate large benchmark datasets of behavioral (including direct and eye-tracking measures) and neural measures from both typical subjects and subjects with VI, enabling a quantitative assessment of functional vision skills. This approach requires a close collaboration among researchers working in the fields of computer vision, computational neuroscience, people with VI, and clinicians. We define "Functional Vision Score, FVS" as the performance of a subject with VI compared to the baseline performance of typical subjects. The definition of common practices to label data from people with VI with metadata reporting clinical assessment of visual functions, together with the assessment of FVS on standardized tasks and techniques, are crucial aspects to model VI. Notably, efforts to model visual behaviors and neural mechanisms in neurotypical subjects are already underway [13], but models in humans with VI are missing. We believe that the directions proposed here would synergize well with existing efforts by adding behavioral and neural data from subjects with VI to guide and constrain model development, and conversely by making use of the leading models for typical vision.

3.1. Modeling direct behavioral performance

Current ANNs solve a range of visual tasks easily overlapping with those listed in 2.2. In the following sections, we will focus on lower level visual tasks, leaving aside reading and mobility. Indeed, reading and mobility applications of ANN models are mainly oriented at replacing the active role of the user with VI (e.g. [50] and [51]). At the same time, complex tasks such as navigation are composed by lower level visual tasks, i.e. object recognition and tracking, face recognition, pose estimation, and action recognition.

CNNs architectures: Until the recent advent of Transformers, CNNs stood out as the best deep models to solve visual tasks. Certain models in the 2010s (e.g. [52, 53, 54, 55, 56]) reached human-level performance in certain image classification tasks. Subsequent evolution led to the development of derived frameworks to solve object detection [57] and semantic segmentation [58] tasks, as well as image retrieval (which implies a visual search ability in humans) and pose estimation [59]. Particular CNNs have been considered for several years the state of the art for visual tasks. Their success was explained by their inductive biases, including translation and equivariance, inspired by the primate visual system [60, 61]. At the same time, CNN layers' activations have been used to explain neural measurements in the primate visual system [62, 63, 14].

Transformers architectures: Recently, another type of artificial network gained extreme popularity, which is not

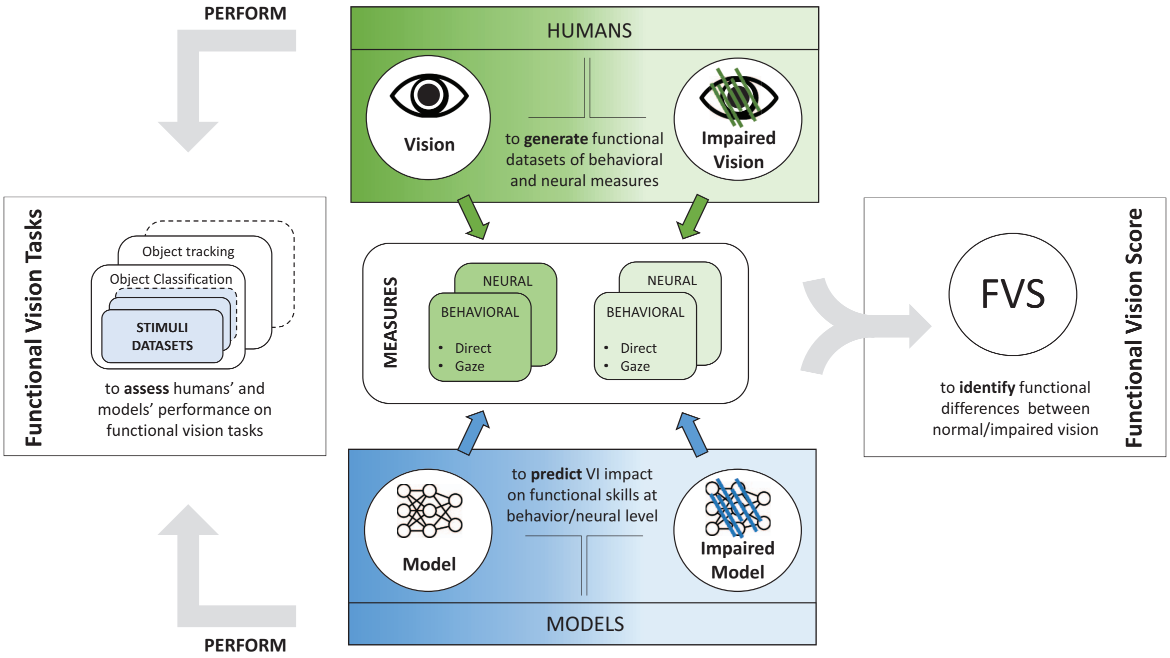
PERFORM
HUMANS
Vision
to generate functional datasets of behavioral and neural measures
Impaired Vision
Functional Vision Tasks
Object tracking
Object Classification
STIMULI DATASETS
to assess humans' and models' performance on functional vision tasks
MEASURES
NEURAL
BEHAVIORAL
Direct
Gaze
NEURAL
BEHAVIORAL
Direct
Gaze
FVS
to identify functional differences between normal/impaired vision
Functional Vision Score
to predict VI impact on functional skills at behavior/neural level
Model
Impaired Model
MODELS
PERFORM

Figure 2. Conceptual framework for developing ANN models of visual impairments (inspired by [13, 49]). The central idea is to identify visual tasks and stimuli datasets that can be used to functionally assess and benchmark both humans' and models' performance (left side). The collection of behavioral (including direct and gaze measures) and neural data from typical and visually impaired subjects (green arrows) will allow the construction of benchmarks for the definition of a "functional vision score (FVS)", quantifying the functional differences among individuals with and without VI in vision-related tasks (right). ANN models can be used simulate the effect of a specific visual impairment on behavioral and neural outcomes. The alignment to human data can then be tracked on behavioral and neural measures (blue arrows).

based on convolution and does not include architectural inductive biases towards local spatial structures. Transformers, mainly based on the mechanism of self-attention, i.e. learned allocation of attention, were first introduced in the context of natural language processing (NLP), achieving significant improvements in various NLP tasks [64, 65]. They were soon translated to vision applications [66] and are now considered a powerful alternative to CNNs and recurrent neural networks. Indeed, they achieved exceptional performance in different visual tasks such as image classification [66], object detection [67], semantic segmentation [68], and pose estimation [69]. Furthermore, thanks to their capability of modeling sequences, Transformers were applied in a number of video tasks, including action recognition [70]. Recent surveys of vision Transformers, categorized based on the visual task complexity can be found in [71] and [72]. Transformers architectures for action recognition are specifically reviewed by [73].

Model-Human comparison: ANNs performing visual tasks at a human level is a useful [62, 74] but not sufficient condition to use such models as a tool to investigate the mechanisms (cognitive and/or neural) through which humans perform the same visual tasks. Indeed, the model's mechanisms allowing to achieve a certain performance need to be investigated as well, to check for similarities or differences compared to humans. Previous research explored various approaches to investigate such human-models similarity, e.g. analyzing the model classification output on its match to human behavioral output, in behavioral comparisons. For example, [75] and [76] introduced metrics to quantify whether two decision-making systems yield the same outputs on the same inputs and make the same mistakes. Models performing better on image classification tasks are generally more consistent with human behavior [75, 14, 76, 77] (but see also [78]). The alignment of CNNs and Transformers with human attention and neural processes is further discussed in the next paragraphs.

3.2. Modeling gaze-based attention

Attention is a key aspect both for visual rehabilitation and as a tool to interpret decision mechanisms of ANNs. Some authors stressed that changes in behavior and visual attention, rather than the change of brain and visual functions, is the true goal of visual rehabilitation [79], and that ”looking” (i.e., acquired skill of paying attention to what is seen), not ”seeing” (i.e., light sensation and transportation to the brain) should be stimulated [80]. At the same time, a change in attention behavior, assessed by gaze data collection, may reflect atypical behavioral development (as it is the case for autism) [21]. Therefore, ANNs incorporating attention mechanisms, i.e. some strategy to weight the input based on task-related high-level features, are an interesting resource to investigate the model capability of capturing human-like strategies to solve complex visual tasks. This research approach is significant towards development of clinical applications of computer vision based on gaze measures, e.g. for the early detection of cognitive impairments, as well as in visual rehabilitation, to investigate the effect of primary perceptual impairments (e.g. peripheral VI) on cognitive development.

Bottom-up and Top-down attention: The computational modeling of gaze-based visual attention is typically addressed by bottom-up and top-down processing [81]. The first considers attention driven by the stimulus' low level visual features (color, contrast). This was the mainstream approach so far, and led to the development of saliency models. The advancement of ANNs and the availability of an increasing number of human gaze dataset benchmarks [82, 83, 84] led to the development of numerous models that are able to predict human fixations on images, mainly during free viewing tasks [85, 86, 87]. The second mechanism of attention is the top-down or endogenous one, i.e goal-directed attention. Since it is an internally induced process based on prior knowledge, it is more significant than bottom-up attention when investigating cognitive mechanisms underlying behavioral observations [88]. Human gaze data is the most common type of measurement to test attentional mechanisms implemented in models. Currently, most of available gaze datasets are not related to goal-oriented tasks, even though some recent research points towards this direction (e.g. [89] published a visual search dataset). Besides saliency models, gradient or attribution methods [90] are used to generate heatmaps indicating local relevancy given by a model to the input image.

Interpretability of attention: Evaluating the similarity of a saliency model to human behavior is still an open challenge. In [91], the authors propose some means of comparison applied to different evaluation metrics to improve interpretability of saliency scores. Different variants of attention have different interpretability properties. To test interpretability, there is a need to provide a definition of ”importance” (e.g. highest attention weight should identify the most influential representation in pushing towards the output class), define the threshold for the model to switch decision (especially in large output spaces, not only limited to few classes), and to evaluate more than one layer of attention [92]. While the majority of works on visual saliency focused on static images, some studies investigated how to understand and model visual attention over dynamic scenes (video saliency). For instance, [93] presents a CNN-LSTM architecture augmented with a supervised attention mechanism to learn temporal saliency representation across successive frames. However, artificial attention does not always align with human intuition. In [94], the authors tested the consistency between a set of representative ANNs with soft attention mechanisms, and human top-down attention based on gaze data, considering three tasks: saliency object segmentation, video action recognition, and fine-grained categorization (see [95] for an extensive survey on attention methods in deep learning, and [96] for a classification of attention mechanisms in deep learning based on data domain). They concluded that human attention can serve as a meaningful ground-truth for lower level tasks, when a higher artificial-human similarity leads to better performance, while this is non always the case for higher-level vision tasks. In other words, the comparison between models' attention and gaze data could be meaningful for visual sub-tasks, but less informative for general high-level tasks, i.e. gaze data represent the sub- visual goals of an overall complex visual task, but may not be informative about the overall strategy.

Besides using human gaze as a ground truth to evaluate models' attention, some works explored the integration of gaze as a supervisory signal to guide neural attention mechanisms, e.g. for NLP [97] or object grasping [98]. Here, gaze data is used directly to make artificial attention more human-like. In the majority of cases, such design choice is motivated by the belief that a human-like attention also leads to improved performance and/or higher interpretability of network's decision mechanism. However, this is not always the case: e.g. [92] found that attention weights are noisy predictors of the importance of the input's intermediate representations in a text classification task.

Neurally inspired models of attention: Most of the computer vision approaches mentioned above bear no resemblance to the neurophysiological architecture of visual cortex. Rather than endeavouring the design of a human-like attention mechanism, the authors in [99] proposed a biologically inspired architecture to gain insight into the mechanisms that guide visual search. Their approach takes steps from neurophysiological knowledge about visual search, which is likely to happen in the form of a task-dependent modulation originating in the frontal cortex [100]. The proposed model provides an approximation to the mechanisms integrating bottom-up and top-down signals during search in natural scenes. Another difference between attention/search mechanisms in artificial models and humans is that typically ANNs process images with space-invariant resolution, contrary to the human visual system, where acuity drops rapidly from the fovea to the peripheral regions of the retina. Attempts of implementing human-like models to predict the fixations' scanpath, i.e. foveation mechanism, recently emerged [101].

Transformers' attention: In contrast to CNNs, Transformers embed a self-attention mechanism within their backbone architecture. Attention was indeed the architectural feature that boosted the performance of these models on many NLP and visual tasks. It is therefore of much interest to interpret their decisions, and this is currently an open problem. A common practice is to consider the learnt self-attention values to visualize a Transformer's relevancy score, either for a single layer, or averaging them on multiple layers. However, this often results in a not meaningful visualization since the attention originating in each layer gets inter-mixed in subsequent layers in a complex manner [72]. In [90], the authors showed how simplistic assumptions (e.g. attention roll-out and attention flow methods) miss to consider different roles of different layers (e.g. deeper layers are more semantic) and they proposed a method to maintain the total relevancy across layers, which also includes the property of class-based separation by design (i.e. different visualizations for different classes). Several papers addressed the following question: is attention in Transformers comparable to the human attention? For instance, [102] compared human attention (based on gaze data) and neural attention for CNN, LSTM and Transformer networks, on a reading comprehension task. While finding the best performance for the Transformer architecture, the authors determined that it was not correlated with higher similarity to human attention (while this correlation held for CNN and LSTM). On the other hand, [103] found that large language models are predictive of human eye fixations during task-specific reading, in a comparable way as classical cognitive models of human attention. Such correlation between Transformers learnt self-attention and overt human attention (assessed with gaze data) during reading tasks is supported also by [104]. Compared to the language domain, few studies investigated the correlation between neural and human attention in visual Transformers. Recently, [105] disputed the similarity between Transformers' and human visual attention, arguing that, from a computational point of view, the purely feed-forward attention's architecture in Transformers (not affected by higher-level factors) performs similarity grouping of visual features, only capturing bottom-up signals. Human visual attention is, in turn, known to be modulated by bottom-up and top-down mechanisms that in early stages of visual processing, allow to organize the perceptual visual input to figures and ground. They conclude that the quest for a computational model that implements human-like visual attention mechanisms has not come to an end with current Transformers.

3.3. Modeling neural processes

Modeling neural processes and neural differences in presence of perceptual or cognitive impairments is crucial for gaining a deeper understanding of each specific issue, and to develop assistive and rehabilitation technologies rooted in neuroscientific findings. With regards to VI, this is especially valuable for subjects with cerebral issues, where perceptual defects originate from brain damage.

Animal models: Animal models, specifically non-human primate models, have been the main source of data for modeling visual cortex. The convolution operation of CNNs, which boosted the performance of computer vision to and above human level in tasks such as object categorization, has a neurobiological basis, and draws its inspiration from the visual processing mechanism of the primate early visual cortex [60, 106]. This was enabled by the fact that non-human primates have a similar developmental profile, including the development of visual functions, as well as a similar visual system organization and level of vision, compared to humans [107, 75]. Animal models of the visual cortex have also been used to identify the causal mechanisms underlying some types of cerebral visual impairments, e.g. amblyiopia, a sensory developmental disorder impacting the structure and function of the visual pathways beginning at the level of the visual cortex. In [108], the author reviews how data generated from macaque models provided useful insights about the neural mechanisms underlying amblyopia, and how such findings are consistent with critical periods and treatment strategies in children with this type of cerebral visual issue.

Brain-like networks: Given the success of models such as CNNs, which are known to have many brain-like properties, it is natural to question whether models performing well in visual tasks, such as recent Vision Transformers, also show human-like properties. Indeed, it is of great interest to investigate the potential of brain-like features to scale up the ANNs performance. To date, few works addressed these questions. In particular, [14] introduced a new large-scale composite of neural and behavioral benchmarks, called Brain-Score, for quantifying the functional fidelity, i.e. how similar an ANN is to the brain's representations in the primate ventral stream as well as to human behavioral measures. They demonstrated that better neural and behavioral alignment correlates with higher model performance on ImageNet, a popular computer vision benchmark. Subsequent work attempted to make the model architecture more similar to the brain's neuroanatomy, e.g. by including recurrent connections [74, 109]. In [13], the authors proposed to extend this approach to the development of integrative benchmarking platforms putting together large-scale brain and behavioral data in the form of accessible benchmarks, and computational models that aim to explain these data. The goal is to push forwards the development of models explaining intelligence in various domains beyond visual intelligence, e.g. language and motor control. In [49], the authors applied this approach to an higher-level cognitive task, i.e. human language processing. They found that particular Transformer models such as GPT2-xl are predictive of neural responses across different recording modalities and datasets (fMRI and ECoG), and that models' fits to behavioral responses are correlated with both neural fits and accuracy on the task of predicting the next word. In a similar vein, in [110], different models, including CNNs and ViTs, were evaluated in their capability to predict neural activities of the human visual cortex, considering as a metric the alignment among model layers and visual regions. Similarly, [111] evaluated the brain-like properties of different types of models, including CNNs, Transformers, and their hybrids. The evaluation focused on the ability of the networks to explain brain activity on the human visual cortex (based on two neural fMRI datasets), and the hierarchical correspondence of ANNs and visual regions. They found that both CNNs and ViTs show hierarchical correspondences to the ventral stream, but neither one is an optimal paradigm to model the visual pathway (even if CNNs perform better on the entry-level and mid-level visual cortex, while ViTs perform better on the higher visual cortex). Critics of the above-mentioned approaches were emboldened by [112], arguing that most behavioral and brain benchmarks for testing models' alignment to human data do not account for the findings and hypotheses from psychological research. Including psychological findings seems in the spirit of Brain-Score [13] which aims to integrate behavioral and neural datasets for models' evaluation.

Impaired models: Neurally plausible models could help to disentangle the reasons behind failures in solving a specific visual task. In [113], the authors presented a method to impair a Transformer language model by deliberately modifying parameters in specific layers of the model self-attention, to generate text with characteristics associated with Alzheimer's disease. By pairing such a degraded model with its unimpaired counterpart, they discriminate between language produced by cognitively healthy and impaired individuals, relaxing the need of large training datasets (which are notoriously harder to build and/or retrieve for impaired categories of subjects than for healthy ones). We believe that such an approach would also be useful in the context of visual impairments. Furthermore, neurally aligned ANNs are likely to be more readily usable to model VI. Specifically, starting from a model that is aligned with data from neurotypical subjects, particular VI could be induced in the ANN that aim to replicate the same behavioral change we observe between people with and without VI. Such impaired models could be used as in-silico testbeds to unravel representational and behavioral changes in individuals with VI. Besides understanding human differences in visual tasks in presence or absence of VI, the comparison between unimpaired/impaired ANNs could also help to better interpret the inner working of the models.

4. Limitations and future directions

There are several limitations towards using ANNs as efficient models of VI. While current ANNs are now considered adequate models of several visual behaviors and the neural mechanisms underlying them [13], (i) they have yet to capture complex functional tasks that are crucial for the development and daily living of humans, e.g. social interactions. One crucial shortcoming is (ii) a lack of standardized tasks, large-scale stimuli and extensive data benchmarks for the assessment of models in tasks related to human social skills [114]. Furthermore, (iii) neural systems associated with cognitive and behavioral processes involved in social situations are far from being understood. The computational modeling of such processes will enable detecting, and aiding various social impairments, even beyond VI [115].

Using ANNs to understand VI is hampered by (iv) the difficulty or retrieving large-scale data from groups of impaired subjects. To tackle this, we encourage the definition of common experimental paradigms and integrated public data benchmarks to improve the availability and accessibility of data from subjects with VI. At the same time, modeling these data with ANN ”impairments” could connect peripheral or neural deficits with behavioral outcomes, and provide a rapid in-silico testbed for treatment strategies.

Computer vision and ANN techniques can be used at several stages of diagnosis, treatment, rehabilitation, and assistance of VI people. Here, we focused on applications for visual assessment and rehabilitation, and we highlighted the need for a closer connection between computer vision scientists and clinical practice. Indeed, a deeper collaboration with operators working in the field of rehabilitation is a crucial step for closing the gap between the two domains for the development of effective tools.

5. Conclusions

This review provides computer vision scientists with a high-level background to approach the field of technologies for visually impaired people. It also provides cognitive neuroscientists and researchers investigating impaired vision with the vocabulary to interact with machine learning practitioners. We demonstrate the complexity of VI and their consequences on key functional abilities, and provide useful guidelines to inform choices about how to develop and apply ANNs to effectively support clinical practices in the field of visual rehabilitation.

Acknowledgments

This research has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 896415, and it was supported by the Center for Brains, Minds and Machines, NSF STC award CCF-1231216.

References

[1] World Health Organization et al. International statistical classification of diseases and related health problems. 2009.

[2] A Mariotti and D Pascolini. Global estimates of visual impairment. *Br J Ophthalmol*, 96(5):614–8, 2012.

[3] August Colenbrander. Measuring vision and vision loss. *Duane's clinical ophthalmology*, 5:1–39, 2001.

[4] Jane N Erin and Beth Paul. Functional vision assessment and instruction of children and youths in academic programs. *Foundations of low vision: Clinical and functional perspectives*, pages 185–220, 1996.

[5] NJ Anastasiow. Implications of the neurobiological model for early intervention. *Handbook of early childhood intervention*, pages 196–216, 1990.

[6] S Carlson, L Hyvärinen, and ANTTI Raninen. Persistent behavioural blindness after early visual deprivation and active visual rehabilitation: a case report. *British Journal of Ophthalmology*, 70(8):607–611, 1986.

[7] Duane Lundervold, Lewis M Lewin, and Larry K Irvin. Rehabilitation of visual impairments: A critical review. *Clinical Psychology Review*, 7(2):169–185, 1987.

[8] Marco Leo, G Medioni, M Trivedi, Takeo Kanade, and Giovanni Maria Farinella. Computer vision for assistive technologies. *Computer Vision and Image Understanding*, 154:1–15, 2017.

[9] Marco Leo, Antonino Furnari, Gerard G Medioni, Mohan Trivedi, and Giovanni M Farinella. Deep learning for assistive computer vision. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[10] Luigi F Cuturi, Elena Aggius-Vella, Claudio Campus, Alberto Parmiggiani, and Monica Gori. From science to technology: Orientation and mobility in blind children and adults. *Neuroscience & Biobehavioral Reviews*, 71:240–251, 2016.

[11] Monica Gori, Giulia Cappagli, Alessia Tonelli, Gabriel Baud-Bovy, and Sara Finocchietti. Devices for visually impaired people: High technological devices with low user acceptance and no adaptability for children. *Neuroscience & Biobehavioral Reviews*, 69:79–88, 2016.

[12] Andrei Barbu, Dalitso Banda, and Boris Katz. Deep video-to-video transformations for accessibility with an application to photosensitivity. *Pattern Recognition Letters*, 137:99–107, 2020.

[13] Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020.

[14] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

[15] Pamela S Roberts, John-Ross Rizzo, Kimberly Hreha, Jeffrey Wertheimer, Jennifer Kaldenberg, Dawn Hironaka, Richard Riggs, and August Colenbrander. A conceptual model for vision rehabilitation. *Journal of rehabilitation research and development*, 53(6):693, 2016.

[16] Susan Silveira. Exploring the dualism of vision–visual function and functional vision. *Vision rehabilitation international*, 10(1):1–10, 2019.

[17] August Colenbrander. Aspects of vision loss–visual functions and functional vision. *Visual impairment research*, 5(3):115–136, 2003.

[18] Sabrina Signorini, Antonella Luparia, Giulia Cappagli, Eleonora Perotto, Mauro Antonini, Federica Morelli, Giorgia Aprile, Elena Ballante, Silvia Figini, Renato Borgatti, et al. Visual function score: A new clinical tool to assess visual function and detect visual disorders in children. *Frontiers in Pediatrics*, 10:868974, 2022.

[19] William V Good, James E Jan, Susan K Burden, Ann Skoczenski, and Rowan Candy. Recent advances in cortical visual impairment. *Developmental medicine and child neurology*, 43(1):56–60, 2001.

[20] Elisa Fazzi, Sabrina G Signorini, Roberta La Piana, Chiara Bertone, Walter Misefari, Jessica Galli, Umberto Balottin, and Paolo Emilio Bianchi. Neuro-ophthalmological disorders in cerebral palsy: ophthalmological, oculomotor, and visual aspects. *Developmental Medicine & Child Neurology*, 54(8):730–736, 2012.

[21] Anna Molinaro, Serena Micheletti, Andrea Rossi, Filippo Gitti, Jessica Galli, Lotfi B Merabet, and Elisa Maria Fazzi. Autistic-like features in visually impaired children: a review of literature and directions for future research. *Brain sciences*, 10(8):507, 2020.

[22] Susanne Trauzettel-Klosinski. Current methods of visual rehabilitation. *Deutsches Ärzteblatt International*, 108(51-52):871, 2011.

[23] Takashi Nakamura. Quantitative analysis of gait in the visually impaired. *Disability and Rehabilitation*, 19(5):194–197, 1997.

[24] John J Rieser, David A Guth, and Everett W Hill. Sensitivity to perspective structure while walking without vision. *Perception*, 15(2):173–188, 1986.

[25] Kolb and B Fantie. Development of the child's brain and behavior./: Handbook of clinical child neuropsychology. eds. cr reynolds, e. fletcher-janzen, 1997.

[26] Stefania M Bova, Elisa Fazzi, Alessia Giovenzana, Cristina Montomoli, Sabrina G Signorini, Marina Zoppello, and Giovanni Lanzi. The development of visual object recognition in school-age children. *Developmental neuropsychology*, 31(1):79–102, 2007.

[27] Peter Stiers, Paul De Cock, and Erik Vandenbussche. Impaired visual perceptual performance on an object recognition task in children with cerebral visual impairment. *Neuropediatrics*, 29(02):80–88, 1998.

[28] FH Boot, JJM Pel, J Van der Steen, and HM Evenhuis. Cerebral visual impairment: which perceptive visual dysfunctions can be expected in children with brain damage? a systematic review. *Research in developmental disabilities*, 31(6):1149–1159, 2010.

[29] Giulia Cappagli, Sara Finocchietti, Gabriel Baud-Bovy, Leonardo Badino, Alessandro D'Ausilio, Elena Cocchi, and Monica Gori. Assessing social competence in visually impaired people and proposing an interventional program in visually impaired children. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):929–935, 2018.

[30] Michael J Guralnick, Robert T Connor, Mary A Hammond, John M Gottman, and Kelly Kinnish. The peer relations of preschool children with communication disorders. *Child development*, 67(2):471–489, 1996.

[31] Marie Celeste. Play behaviors and social interactions of a child who is blind: In theory and practice. *Journal of visual impairment & Blindness*, 100(2):75–90, 2006.

[32] Sharon Sacks and Karen E Wolffe. *Teaching social skills to students with visual impairments: From theory to practice*. American Foundation for the Blind, 2006.

[33] Lisa Putzar, Kirsten Hötting, and Brigitte Röder. Early visual deprivation affects the development of face recognition and of audio-visual speech perception. *Restorative neurology and neuroscience*, 28(2):251–257, 2010.

[34] Dannyelle Valente, Anne Theurel, and Edouard Gentaz. The role of visual experience in the production of emotional facial expressions by blind people: A review. *Psychonomic bulletin & review*, 25(2):483–497, 2018.

[35] Gary S Rubin, Sheila K West, Beatriz Munoz, Karen Bandeen-Roche, Scott Zeger, Oliver Schein, and Linda P Fried. A comprehensive assessment of visual impairment in a population of older americans. the see study. salisbury eye evaluation project. *Investigative ophthalmology & visual science*, 38(3):557–568, 1997.

[36] Susan J Leat, Naveen K Yadav, and Elizabeth L Irving. Development of visual acuity and contrast sensitivity in children. *Journal of optometry*, 2(1):19–26, 2009.

[37] Marlou JG Kooiker, Johan JM Pel, Hélène JM Verbunt, Gerard C de Wit, Maria M van Genderen, and Johannes van der Steen. Quantification of visual function assessment using remote eye tracking in children: validity and applicability. *Acta ophthalmologica*, 94(6):599–608, 2016.

[38] Deepesh Kumar, Anirban Dutta, Abhijit Das, and Uttama Lahiri. Smarteye: Developing a novel eye tracking system for quantitative assessment of oculomotor abnormalities. *IEEE Transactions on neural systems and rehabilitation engineering*, 24(10):1051–1059, 2016.

[39] William V Good, Chuan Hou, and Anthony M Norcia. Spatial contrast sensitivity vision loss in children with cortical visual impairment. *Investigative Ophthalmology & Visual Science*, 53(12):7730–7734, 2012.

[40] Vijaya K Gothwal, Jan E Lovie-Kitchin, and Rishita Nutheti. The development of the lv prasad-functional vision questionnaire: a measure of functional vision performance of visually impaired children. *Investigative ophthalmology & visual science*, 44(9):4131–4139, 2003.

[41] Jyoti Khadka, Barbara Ryan, Tom H Margrain, J Margaret Woodhouse, et al. Development of the 25-item cardiff visual ability questionnaire for children (cvaqc). *British Journal of Ophthalmology*, 94(6):730–735, 2010.

[42] Elisa Fazzi and Serena Micheletti. Questionnaires as screening tools for children with cerebral visual impairment. *Developmental Medicine & Child Neurology*, 62(8):891–891, 2020.

[43] Howard Moshtael, Tariq Aslam, Ian Underwood, and Baljean Dhillon. High tech aids low vision: a review of image processing for the visually impaired. *Translational vision science & technology*, 4(4):6–6, 2015.

[44] Sheila K West, Gary S Rubin, Aimee T Broman, Beatriz Munoz, Karen Bandeen-Roche, Kathleen Turano, SEE Project Team, et al. How does visual impairment affect performance on tasks of everyday life?: The see project. *Archives of Ophthalmology*, 120(6):774–780, 2002.

[45] Valérie Caron, Alessio Barras, Ruth MA van Nispen, and Nicolas Ruffieux. Teaching social skills to children and adolescents with visual impairments: A systematic review. *Journal of Visual Impairment & Blindness*, 117(2):128–147, 2023.

[46] Quentin Guillon, Nouchine Hadjikhani, Sophie Baduel, and Bernadette Rogé. Visual social attention in autism spectrum disorder: Insights from eye tracking studies. *Neuroscience & Biobehavioral Reviews*, 42:279–297, 2014.

[47] Marlou JG Kooiker, Johan JM Pel, Sanny P van der Steen-Kant, and Johannes van der Steen. A method to quantify visual information processing in children using eye tracking. *JoVE (Journal of Visualized Experiments)*, (113):e54031, 2016.

[48] Li-Ting Tsai, Ling-Fu Meng, Wei-Chi Wu, Yuh Jang, and Yu-Chin Su. Effects of visual rehabilitation on a child with severe visual impairment. *The American Journal of Occupational Therapy*, 67(4):437–447, 2013.

[49] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

[50] Jothi Ganesan, Ahmad Taher Azar, Shrooq Alsenan, Nashwa Ahmad Kamal, Basit Qureshi, and Aboul Ella Hassanien. Deep learning reader for visually impaired. *Electronics*, 11(20):3335, 2022.

[51] Feng Hu, Hao Tang, Aleksandr Tsema, and Zhigang Zhu. Computer vision for sight: Computer vision techniques to assist visually impaired people to navigate in an indoor environment. In *Computer Vision for Assistive Healthcare*, pages 1–49. Elsevier, 2018.

[52] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[53] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[54] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[55] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[56] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[57] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2147–2154, 2014.

[58] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[59] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S Lew. Deep learning for visual understanding: A review. *Neurocomputing*, 187:27–48, 2016.

[60] Kunihiko Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, 1(2):119–130, 1988.

[61] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[62] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.

[63] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[65] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[66] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[67] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[68] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[69] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021.

[70] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 244–253, 2019.

[71] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

[72] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

[73] Anwaar Ulhaq, Naveed Akhtar, Ganna Pogrebna, and Ajmal Mian. Vision transformers for action recognition: A survey. *arXiv preprint arXiv:2209.05700*, 2022.

[74] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32, 2019.

[75] Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. 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*, 38(33):7255–7269, 2018.

[76] Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33:13890–13902, 2020.

[77] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021.

[78] Thomas Fel, Ivan F Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems*, 35:9432–9446, 2022.

[79] Lori Goetz and Kathleen Gee. Teaching visual attention in functional contexts: Acquisition and generalization of complex visual motor skills. *Journal of Visual Impairment & Blindness*, 81(3):115–117, 1987.

[80] Patricia M Sonksen, Aviva Petrie, and Kristina J Drew. Promotion of visual development of severely visually impaired babies: evaluation of a developmentally based programme. *Developmental Medicine & Child Neurology*, 33(4):320–335, 1991.

[81] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[82] Matthias Kümmerer, Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit/tübingen saliency benchmark. https://saliency.tuebingen.ai/.

[83] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 798–814. Springer International Publishing.

[84] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*, 2015. arXiv preprint arXiv:1505.03581.

[85] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

[86] Ali Borji, Dicky N Sihite, and Laurent Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *Image Processing, IEEE Transactions on*, 22(1):55–69, 2013.

[87] Matthias Kümmerer, Matthias Bethge, and Thomas SA Wallis. Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5):7–7, 2022.

[88] Fumi Katsuki and Christos Constantinidis. Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist*, 20(5):509–521, 2014.

[89] Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11(1):8776, 2021.

[90] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

[91] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605*, 2016.

[92] Sofia Serrano and Noah A Smith. Is attention interpretable? *arXiv preprint arXiv:1906.03731*, 2019.

[93] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):220–237, 2019.

[94] Qiuxia Lai, Salman Khan, Yongwei Nie, Hanqiu Sun, Jianbing Shen, and Ling Shao. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 23:2086–2099, 2020.

[95] Mohammed Hassanin, Saeed Anwar, Ibrahim Radwan, Fahad S Khan, and Ajmal Mian. Visual attention methods in deep learning: An in-depth survey. *arXiv preprint arXiv:2204.07756*, 2022.

[96] Yuting Yang, Licheng Jiao, Xu Liu, Fang Liu, Shuyuan Yang, Zhixi Feng, and Xu Tang. Transformers meet visual learning understanding: A comprehensive review. *arXiv preprint arXiv:2203.12944*, 2022.

[97] Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. Improving natural language processing tasks with human gaze-guided neural attention. *Advances in Neural Information Processing Systems*, 33:6327–6341, 2020.

[98] Ivan Gonzalez-Diaz, Jenny Benois-Pineau, Jean-Philippe Domenger, Daniel Cattaert, and Aymar de Rugy. Perceptually-guided deep neural networks for ego-action prediction: Object grasping. *Pattern Recognition*, 88:223–235, 2019.

[99] Mengmi Zhang, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman. Finding any waldo with zero-shot invariant and efficient visual search. *Nature communications*, 9(1):3730, 2018.

[100] Narcisse P Bichot, Matthew T Heard, Ellen M DeGennaro, and Robert Desimone. A source for feature-based attention in the prefrontal cortex. *Neuron*, 88(4):832–844, 2015.

[101] Hristofor Lukanov, Peter König, and Gordon Pipa. Biologically inspired deep learning model for efficient foveal-peripheral vision. *Frontiers in Computational Neuroscience*, 15:746204, 2021.

[102] Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. Interpreting attention models with human visual attention in machine reading comprehension. *arXiv preprint arXiv:2010.06396*, 2020.

[103] Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of*

the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4295–4309, 2022.

[104] Emmanuele Chersoni, Nora Hollenstein, Cassandra L Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. Proceedings of the workshop on cognitive modeling and computational linguistics. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 2021.

[105] Paria Mehrani and John K Tsotsos. Self-attention in vision transformers performs perceptual grouping, not attention. *arXiv preprint arXiv:2303.01542*, 2023.

[106] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[107] Nikolaus Kriegeskorte, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008.

[108] Lynne Kiorpes. Understanding the development of amblyopia using macaque monkey models. *Proceedings of the National Academy of Sciences*, 116(52):26217–26223, 2019.

[109] Aran Nayebi, Daniel Bear, Jonas Kubilius, Kohitij Kar, Surya Ganguli, David Sussillo, James J DiCarlo, and Daniel L Yamins. Task-driven convolutional recurrent models of the visual system. *Advances in neural information processing systems*, 31, 2018.

[110] Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *bioRxiv*, pages 2022–03, 2022.

[111] Qiongyi Zhou, Changde Du, and Huiguang He. Exploring the brain-like properties of deep neural networks: a neural encoding perspective. *Machine Intelligence Research*, 19(5):439–455, 2022.

[112] Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico Adolfi, John E Hummel, Rachel F Heaton, et al. Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, pages 1–74, 2022.

[113] Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. Gpt-d: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. *arXiv preprint arXiv:2203.13397*, 2022.

[114] Ravi Tejwani, Yen-Ling Kuo, Tianmin Shu, Boris Katz, and Andrei Barbu. Social interactions as recursive mdps. In *Conference on Robot Learning*, pages 949–958. PMLR, 2022.

[115] Corneliu Florea, Laura Florea, and Constantin Vertan. Computer vision for cognition: An eye focused perspective. In *Computer Vision for Assistive Healthcare*, pages 51–74. Elsevier, 2018.