Binocular vision supports the development of scene segmentation capabilities: Evidence from a deep learning model

Ross Goutcher Psychology Division, Faculty of Natural Sciences, University of Stirling, Stirling, UK

Christian Barrington Psychology Division, Faculty of Natural Sciences, University of Stirling, Stirling, UK; Computing Science and Mathematics Division, Faculty of Natural Sciences, University of Stirling, Stirling, UK

Paul B. Hibbard Department of Psychology, University of Essex, Colchester, UK

Bruce Graham Computing Science and Mathematics Division, Faculty of Natural Sciences, University of Stirling, Stirling, UK

The application of deep learning techniques has led to substantial progress in solving a number of critical problems in machine vision, including fundamental problems of scene segmentation and depth estimation. Here, we report a novel deep neural network model, capable of simultaneous scene segmentation and depth estimation from a pair of binocular images. By manipulating the arrangement of binocular image pairs, presenting the model with standard left-right image pairs, identical image pairs or swapped left-right images, we show that performance levels depend on the presence of appropriate binocular image arrangements. Segmentation and depth estimation performance are both impaired when images are swapped. Segmentation performance levels are maintained, however, for identical image pairs, despite the absence of binocular disparity information. Critically, these performance levels exceed those found for an equivalent, monocularly trained, segmentation model. These results provide evidence that binocular image differences support both the direct recovery of depth and segmentation information, and the enhanced learning of monocular segmentation signals. This finding suggests that binocular vision may play an important role in visual development. Better understanding of this role may hold implications for the study and treatment of developmentally acquired perceptual impairments.

Introduction

In both biological and artificial systems, visual processing supports the recovery of critical environmental properties, such as segmentation of figure from background, and the localization of objects in depth. In biological systems, numerous visual cues have been identified as supporting both figure-ground segmentation (e.g., Wagemans, Elder, Kubovy, Palmer, Peterson, Singh, & von der Heydt,2012) and the measurement of depth (e.g., Cutting & Vishton, 1995; Howard & Rogers, 2012; Welchman, 2016). These cues have typically formed the basis of efforts to model the processing of depth and object segmentation in the brain (e.g., Elder, Krupnik, & Johnston, 2003; Elder, 2018; Hildreth & Royden, 2011; Langer, Zheng, & Rezvankhah, 2016; Watt, Ledgeway, & Dakin, 2008). In machine vision research, problems of scene segmentation and depth estimation have been similarly addressed to support multiple applications including the guidance of autonomous vehicles (e.g., Smolyanskiy, Kamenev, & Birchfield, 2018), object tracking (Wang, Zhang, Bertinetto, Hu, & Torr, 2019) and enhanced scene understanding (e.g., Garcia-Garcia, Orts-Escolano, Oprea, Villena-Martinez, & Garcia-Rodriguez, 2017; Huang, Matzen, Kopf, Ahuja, & Huang, 2018; Jégou, Drozdzal, Vazquez, Romero, & Bengio, 2017; Song, Lichtenberg, & Xiao, 2015; Wang & Shen, 2018).

Citation: Goutcher, R., Barrington, C., Hibbard, P. B., & Graham, B. (2021). Binocular vision supports the development of scene segmentation capabilities: Evidence from a deep learning model. *Journal of Vision, 21*(7):13, 1–18, https://doi.org/10.1167/jov.21.7.13.

https://doi.org/10.1167/jov.21.7.13 Received March 17, 2021; published July 21, 2021 ISSN 1534-7362

Despite these overlapping concerns, substantial differences exist between the modeling approaches used in biological vision research, and those used in the development of machine vision systems. Despite notable early exceptions (e.g., Marr & Hildreth, 1980; Marr & Poggio, 1979), models of biological visual processing have often focused on accounting for the performance of human observers on an array of psychophysical tasks (e.g., Banks, Gepshtein, & Landy, 2004; Goutcher, 2016; Henricksen, Cumming, & Read, 2016; Hillis, Ernst, Banks, & Landy, 2002; Lovell, Bloj, & Harris, 2012). Many such models also constrain themselves to consider processing in biologically plausible terms, often focusing on the combination of physiologically inspired receptive field structures or other established response properties of neurons or neuronal populations in visual cortex (e.g., Banks et al., 2004; Chauhan, Masquelier, Montlibert, & Cottereau, 2018; Ecke, Papp, & Mallot, 2021; Goncalves & Welchman, 2017; Henricksen et al., 2016; Maiello, Chessa, Bex, & Scolari, 2020; May, Zhaoping, & Hibbard, 2012; Watt et al., 2008). Indeed, for a subset of physiologically inspired models, a primary aim is to account for the functioning or organization of particular sets of neurons, rather than to directly solve how such neurons contribute to specific visual or visuomotor behaviors (e.g., Bredfeldt, Read, & Cumming, 2009; DeAngelis, Ohzawa, & Freeman, 1991).

In contrast, machine vision research has largely focused on providing exactly these solutions. As such, the goal for machine vision models has been to maximize the precision and accuracy of performance, typically in estimation and categorization tasks, with real world, or close to real world, scenes. This typically means that such models disregard some of the peculiarities of biological visual processing that are often highlighted in the use of specific stimuli and experimental designs (e.g., Goutcher, Connelly & Hibbard, 2018; Goutcher & Wilcox, 2016; Goutcher & Wilcox, 2021; Harris, 2014; Kingdom, Yared, Hibbard, & May, 2020; Wardle, Palmisano, & Gillam, 2014; Wardle & Gillam, 2016). Yet these more targeted examinations of specific effects in visual processing (e.g., depth illusions, bias, etc.) can often reveal critical information about visual signals and visual processing that might otherwise be overlooked in the search for solutions to real world tasks. In this article, we show how a consideration of the visual cues for scene segmentation and depth measurement may inform the development of a deep neural network (DNN) model of these processes. This model was focused on examining the efficacy of binocular cues, which we consider in detail below.

Binocular signals for depth estimation and scene segmentation

Depth estimation and scene segmentation are two fundamental problems for visual processing. Depth estimation problems include the measurement of egocentric distance, as well as the measurement of relative depth and the estimation of three-dimensional (3D) object shape (cf., Parker, 2007; Howard & Rogers, 2012; Welchman, 2016). Scene segmentation problems can be similarly subdivided, delineating problems of figure-ground segmentation, object boundary classification, semantic segmentation and instance segmentation (Garcia-Garcia et al., 2017; Long, Shelhamer, & Darrell, 2015). In this article, we consider the problems of estimating egocentric distance, defined as the normalized absolute distance to the observer for each pixel in a scene, and semantic segmentation, defined as the production of a pixel-by-pixel map defining the identity and location of each object in a scene.

For depth estimation, numerous visual cues have been identified as informative of depth structure, including pictorial cues, such as shape-from-shading, texture gradients, and linear perspective, and dynamic cues such as motion parallax and structure-from-motion (cf., Howard & Rogers, 2012; Welchman, 2016). A similar array of cues has been proposed for scene segmentation, including the identification of edges defined by differences in luminance, contrast, color, texture and/or motion (e.g. Martin, Fowlkes, & Malik, 2004; Sundberg, Brox, Maire, Arbeláez, & Malik, 2011), as well as principles for the grouping and interpretation of such edges (Wagemans et al., 2012).

Binocular images provide a particularly important source of information for both depth estimation and scene segmentation. The role of binocular signals for depth estimation is relatively uncontroversial; small positional differences between left and right eye images, known as binocular disparities, are highly informative of the 3D structure of the distal scene. Use of this depth cue depends on the resolution of the problem of binocular correspondence, where matching points are found between left and right eye images. Numerous rules, constraints and heuristics have been proposed for correspondence resolution in biological vision (e.g., Goutcher & Hibbard, 2010; Goutcher & Hibbard, 2014; Marr & Poggio, 1979), many of which are implicitly implemented in biologically inspired algorithms that measure binocular disparity through cross-correlation, or cross-correlation-like processes (Banks et al, 2004; Qian & Zhu, 1997; Read & Cumming, 2007). Such models are often derived from the binocular energy model, where the energy at a given

disparity is measured as the sum of quadrature pairs of phase or position-shifted binocular simple cell receptive fields (DeAngelis et al, 1991; Fleet, Wagner & Heeger 1996). The use of such constraints is also evident in the measurement and optimization processes of many classical machine vision models of depth estimation (Hirschmulller, 2005; Scharstein & Szeliski, 2002).

In addition to their role in depth estimation, binocular images also contain important signals for scene segmentation. Depth differences at object boundaries may give rise to areas of binocular half-occlusion, where regions of an image are visible to one eye only (Harris & Wilcox, 2009; Nakayama & Shimojo, 1990; Tsirlin, Wilcox, & Allison, 2010). This absence of matching regions between left and right eyes results in changing patterns of disparity energy, where unmatched regions are likely to be associated with generally low levels of binocular correlation across a range of potential disparity values, and where the local image structure at these unmatched regions is more likely to match neighboring "background" image areas (Basgöze, White, Burge, & Cooper, 2020). Where binocular disparity information is available, disparity-defined boundaries are typically associated with large disparity gradients (Basgöze et al., 2020; Cammack & Harris, 2016; Goutcher et al., 2018; Goutcher & Wilcox, 2021). Note that, for both half-occlusion and disparity-defined boundaries, binocular segmentation signals are the result of image-based, rather than depth or distance-based, computations. This contrasts with many machine vision models of segmentation, where monocular images may be supplemented by an explicit depth channel, rather than make direct use of binocular imaging (Eitel, Springenberg, Spinello, Riedmiller, & Burgard, 2015; Silberman, Hoiem, Kohli, & Fergus, 2012; Song et al., 2015, although see some early work by, for example, Birchfield & Tomasi, 1999). The DNN model detailed in this article makes use of binocular image inputs to provide access to these image-based cues.

Deep learning as a tool for understanding the brain

Given the, often diverging, purposes of modeling endeavors in biological and machine vision research, one may wonder whether there are any benefits to utilizing machine vision approaches in an attempt to understand biological systems. Recently, this question has come under renewed focus with the rise of deep learning approaches in machine vision (cf., Lopez-Rubio, 2018; Majaj & Pelli, 2018; Richards et al., 2019). For many researchers in biological vision, deep learning networks provide an attractive and powerful way to conceive of the processes occurring in the mammalian visual system (Kriegeskorte, 2015; Rideaux & Welchman, 2020; Rideaux & Welchman, 2021; Srinath, Emonds, Wang, Lempel, Dunn-Weiss, Connor, & Nielsen, 2020). Like cells in the visual pathway, from retina to cortex, the filtering operations in DNNs make use of operations such as convolutions and max pooling, with some model architectures (e.g., "AlexNet"; Krizhevsky, Sutskever, & Hinton, 2017) exhibiting filter weights that bear similarity to the excitatory-inhibitory receptive field structures found in retinal ganglion cells, LGN and primary visual cortex. The activity of these forms of DNN has been used to draw inferences about the processing potential of areas further along the visuo-cortical pathways (e.g., Srinath et al., 2020). Yet many of the model architectures used in machine vision differ significantly from the processing pathways seen in biological visual systems. The development of DNN models also typically depends on supervised learning processes that differ markedly from the kinds of feedback available to active organisms (for detailed discussion, see Majaj & Pelli, 2018). Together, these differences suggest that, at the very least, substantial care must be taken when drawing comparisons between the activity in DNNs and the processing occurring in biological systems.

There is, however, another way to make use of DNN performance as a tool for understanding biological vision. Rather than consider DNNs as intrinsically informative of the processes occurring in biological visual systems, one may instead consider such networks as informative of the signals present in the input images. Thus, one may consider the capacity of DNNs to successfully perform a given task (e.g., segmentation and/or depth estimation) as indicative of the presence of task-relevant information within the input images and of its encoding by the network. By extension, one may therefore consider changes in performance in response to a principled stimulus manipulation as indicative of the efficacy of the manipulated stimulus information for the model's set task.

This approach is similar to existing model-based analysis methods, such as Bayesian-derived ideal observer-based measures of efficiency (Barlow, 1962; Pelli, 1990; Pelli, & Farell, 1999) and the application of support vector machines, for example in the classification of signal-relevant responses in neuroimaging data (LaConte, Strother, Cherkassky, Anderson, & Hu, 2005). It also extends image analysis-based approaches aimed at understanding the statistics of natural scenes (e.g., Adams, Elder, Graf, & Muryy, 2016; Burge & Geisler, 2014; Fowlkes, Martin, & Malik, 2007; Hunter & Hibbard, 2018) by determining whether useful visual information is readily recoverable. Using this approach, one may distinguish between signals that are *in principle* informative for a given task, and those that, given the complex, multi-object structures found in natural scenes, a sensory system can actually learn to use effectively. This can be done

by focusing on the input sensory information, and output task performance, without there necessarily being any direct relationship between the properties of the DNN hidden layers, and any particular features of the biological visual system (López-Rubio, 2018). Under this approach, network architectures can be considered as hypotheses on the importance of specific image properties for the task(s) under investigation. In this article, we examine the learning of depth estimation and segmentation signals from binocularly presented, rendered, multi-object scenes and examine the role played by binocular signals in this process. We show that manipulation of binocular image signals significantly impacts on model performance in both depth estimation and scene segmentation tasks, highlighting the importance of binocular viewing for these tasks. We further show that inputs from scene segmentation pathways in our network significantly enhance depth estimation performance.

Methods

To assess the importance of binocular cues for scene segmentation and depth estimation, we developed a DNN that took binocular images as inputs. This model was structured as an encoder-decoder network, an architecture that has previously proven useful for both segmentation and depth estimation tasks (Garcia-Garcia et al., 2017; Wang & Shen, 2018). Specifically, our architecture was based on the U-net network developed by Ronneberger, Fischer and Brox (2015). Following initial layers of feature extraction, left and right image input pathways converged on a common binocular stage. Subsequent processing stages were separated into parallel pathways for segmentation and depth estimation, producing a single set of depth estimates for distance to the left camera, and object identity segmentation maps for both left and right camera images. The model was trained on both tasks simultaneously.

A 3D rendered image training set

We constructed a training dataset of complex scenes, each containing multiple objects. Scenes were constructed using Blender (Blender Foundation, Amsterdam, Netherlands), and objects were drawn from an existing dataset of high-quality 3D renders of real objects (Solid Sight Dataset – Hibbard, Scarfe, Hornsey, & Hunter, 2016). Objects were scanned using a NextEngine 3D laser scanner (NextEngine Inc., Santa Monica, CA, USA), creating high-density 3D models of the object. Our complex scenes each contained 24 distinct objects from the dataset, arranged in pseudo-random positions to mimic objects distributed across a circular flat surface of radius 5.2 m. Scanned objects were an array of fruits, vegetables and toys, rendered and captured under a diffuse light source. Each scene was viewed within a hemispherical domed "sky" covering the full circular area of the scene. This was textured using image samples taken from the McGill Calibrated Colour Image Dataset (Olmos & Kingdom, 2004). Image samples were used to ensure that segmentation was not overly simplified by the presence of large blank areas, but instead were comprised of image information consistent with the statistics of natural scenes. Example images and ground truths are shown in Figure 1.

To train our networks we generated 40 such scenes, capturing pairs of 224 × 224 pixel RGB images from laterally separated binocular cameras. Images were captured for 50 frames following pseudo-random walks around the perimeter of the scene. The inter-camera distance was 6.5 cm, with parallel viewing geometry. Training images were accompanied by pixel-by-pixel ground-truth segmentation images for the left and right camera and equivalent depth maps, obtained as part of the image rendering process. Segmentation maps used One-Hot encoding (cf., Cerda, Varoquaux, & Kégl, 2018) to specify categories, matched to a colourmap for later visualization. Depth map values were normalized to fall between values of 0 and 1, with 0 being the value closest to the camera and 1 being the farthest value. Depth values were therefore on a relative scale, although, in practice, nearest and farthest distances were equivalent between scenes. In addition to this initial training set, we generated a further validation set of seven scenes, containing pairs of images for 50 frames to check trained model performance. All results reported below used a final test set of four novel scenes not previously presented to the model. The use of identical test images allowed for statistical analysis using related-samples approaches.

Model architecture and training regime

Our binocular encoder-decoder model is built on a TensorFlow 2.x and Keras backbone. In this model, features are encoded from the RGB image pair inputs and fed into a common binocular stage for the simultaneous learning of segmentation and depth from a shared feature pool. The binocular stage is a concatenation of prior monocular filter pathways, allowing differences in filter structure at equivalent image locations to affect subsequent processing. Information from this binocular concatenation stage is fed forward into the three output branches, left image segmentation, right image segmentation and depth prediction. As stated above, depth is calculated as normalized distance. This was calculated

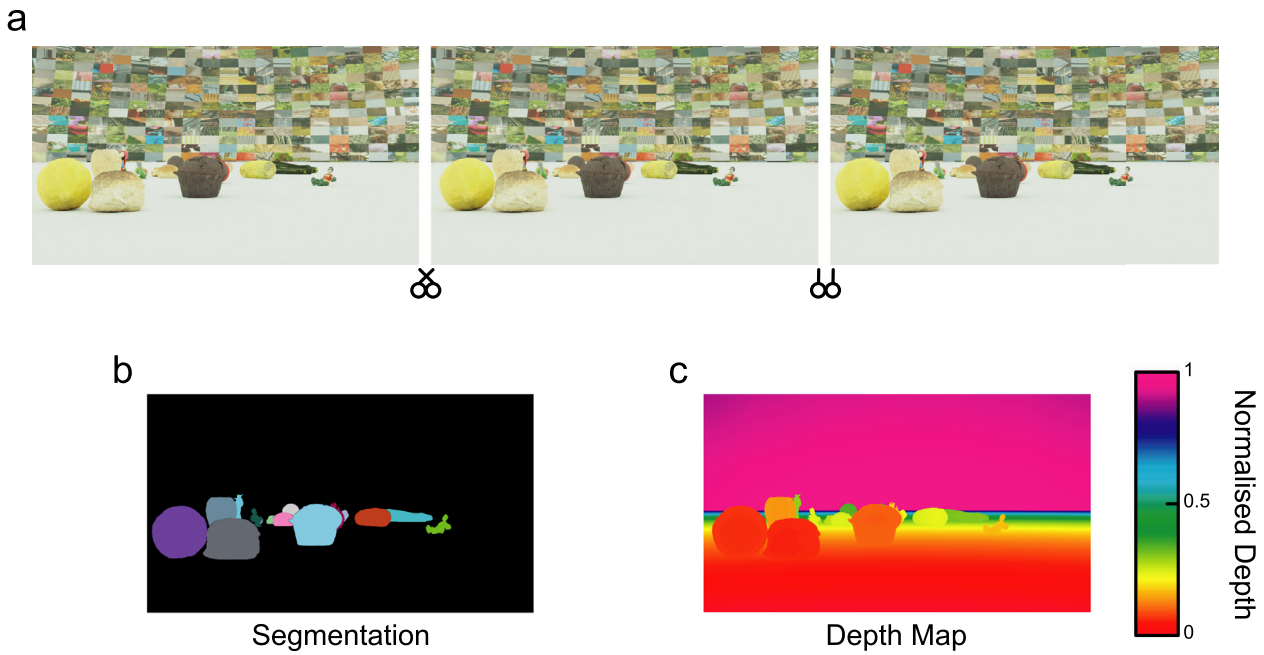

Figure 1. Example images from the dataset (a) arranged for crossed and uncrossed free fusion, together with (b) segmentation and (c) normalized depth ground truths. Segmentation image colors label each object category (see Figure 4b for details). Normalized depth values vary from 0 (closest, in red) to 1 (farthest, in violet).

relative to the left camera only. An illustration of the model architecture is shown in Figure 2, with full details provided in Supplementary Table S1. There were a total of 10,901,923 parameters in the network, with 10,889,955 trainable parameters and 11,968 nontrainable parameters. Code for the network is available at the following URL: https://github.com/StirlingChris/User_Version/tree/master.

The model consisted of five types of layers: convolutional, batch normalization, maximum pooling, up-sampling, SoftMax and skip connections/residual units. We used a convolutional filter size of 3 × 3 and a stride of one for all convolutional layers. Batch normalization is used to rescale the values of the results between zero and one to improve model efficiency and stability. We also utilized maximum pooling to summarize and reduce the dimensionality of the extracted features, decreasing the number of parameters in the overall model and the size of the input into the next layer. The up-sampling modules are used when decoding the features extracted and to increase the size of the input so that the output is the same size as the ground-truths the model is trained on. Skip connections have been shown to improve the passing on of information between layers and to help preserve spatial information (He, Zhang, Ren, & Sun, 2016; Jégou et al., 2017), which is of particular use for our model. We used skip connections to pass information between modules of matching size in the encoding and decoding ends of the network.

The output layer of the segmentation branch consists of SoftMax neurons, which output a probability that each pixel is a given class, producing a 224 × 224 RGB segmentation map as output. The argmax of all the SoftMax outputs is then taken as the most likely class for each pixel. The number of SoftMax neurons determines the maximum number of potential outputs. As such, we used 25 SoftMax neurons; one for each possible object category (24 object categories, plus background) in our dataset.

An additional feature of our model is that we introduced connections between these segmentation pathways and the depth branch. The depth branch of the network has features from both segmentation branches fed into it at 3 dimensionalities: 28 × 28, 56 × 56 and 112 × 112. These features are then fed forward to the linear activation output unit to produce a depth map of a 224 × 244 × 1 image as output. Note that there were, however, no direct connections from the network's depth branch to the segmentation branches. This means that any depth-based segmentation cues must be derived directly from underlying image properties and cannot be due to the explicit measurement of depth. The potential benefit of these connections instead lies in the enhancement of depth estimation processes. Several recent psychophysical findings (Cammack & Harris, 2016; Deas & Wilcox, 2014, 2015; Goutcher et al. 2018; Goutcher & Wilcox, 2021; Mamassian & Zannoli, 2020) point to the importance of segmentation boundaries in the quantitative perception of binocular depth.

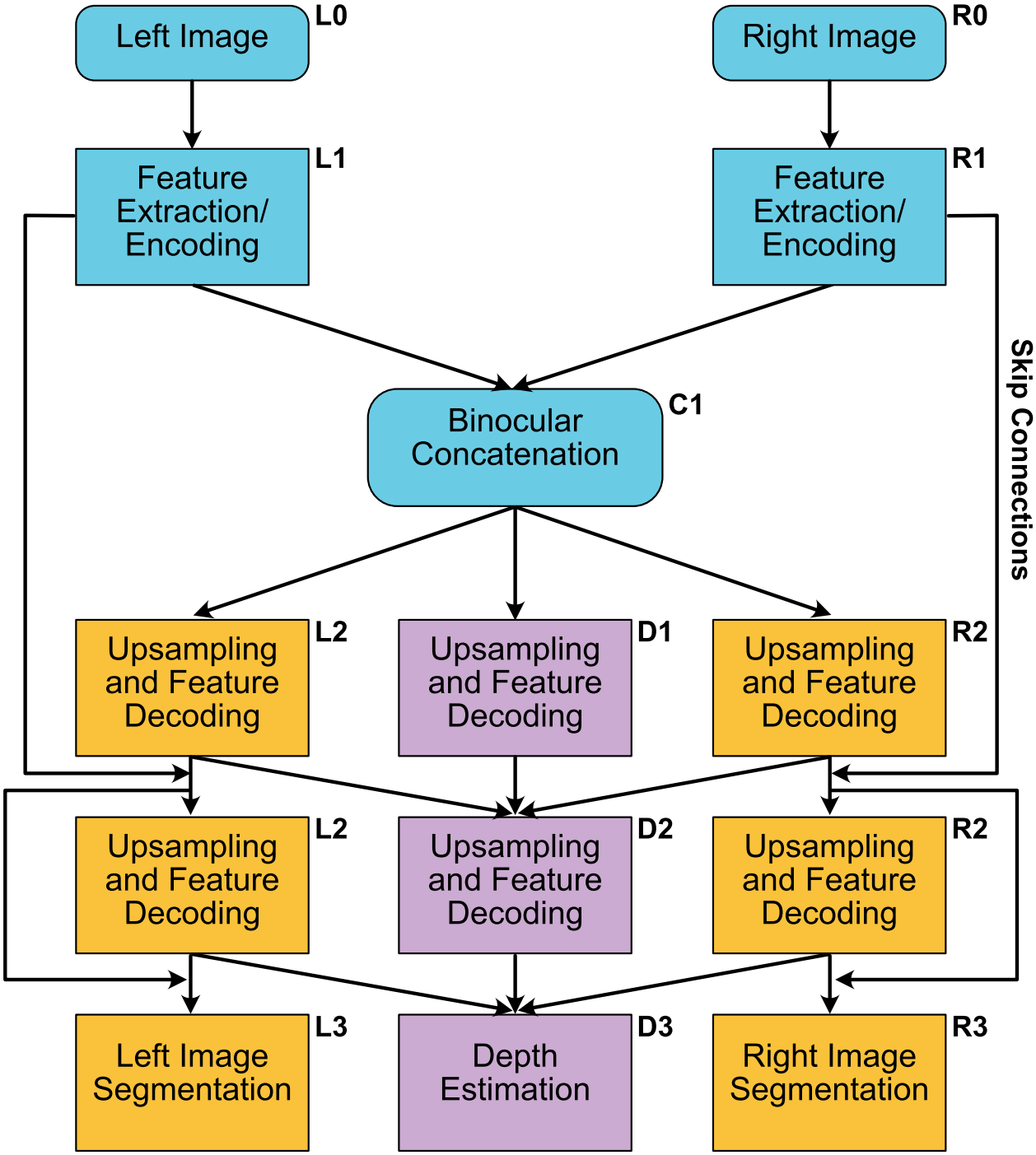

Figure 2. Illustration of the network architecture. Left and right input pathways feed into a common binocular concatenation stage, leading to distinct segmentation and depth estimation output pathways. Full details of the operations at each network layer are provided in Supplementary Table S1.

By explicitly manipulating the input of segmentation information into depth measurement processes, we tested whether our network was able to learn to make use of such signals.

The network was trained using the 40 binocular training scenes described in section 2.1, where each scene contained 50 binocular, 224 × 224 pixel, RGB frames. No data augmentation was used during our training processes, except for vertical flips which were tested but offered little for improving performance while also increasing training time. Standard data augmentation procedures were avoided, as these alter available disparity signals and would, therefore, affect the learning of disparity dependent

signals. No dropout was used on any part of the network.

The model was trained using the Adam optimizer with learning rate $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon =$ 1e -07, where α controls the step-size for weight changes on each iteration, β_1 and β_2 control the decay rate on moving averages of the first and second moment of the estimated gradient, and ε prevents division by zero (see Kingma & Ba, 2014, for further details). The ReLu activation function was used on all units except for the output neurons, which used SoftMax and a linear activation. We trained for a maximum of 150 epochs, using a batch size of 8 image pairs and a step size of 32. Image pairs were shuffled between epochs. Model performance was calculated on each iteration using a categorical cross-entry loss function for each segmentation pathway and by finding the per image root mean squared error (RMSE) for the depth estimation pathway. The total loss for the model was taken as the sum of these measures. An early-stopping condition was specified, where training was stopped and "best performing" weights saved if performance failed to improve for 15 consecutive epochs. Best performance was measured as minimized validation loss. Model weights at each layer of the network were held constant following the completion of these training regimes.

Binocular image manipulation

To examine the contribution of binocular image signals to model performance, we tested our DNN under three distinct viewing conditions. Following training and validation, the model was presented with test images under standard viewing conditions, identical to the image arrangements used in training, identical images or images swapped between left and right cameras. This type of swapped image presentation is typically referred to as *pseudoscopic viewing*, after the device developed by Charles Wheatstone (1852). The presentation of identical images removes all binocular disparity and monocular occlusion cues from the input images, while swapped presentation reverses these signals while leaving monocular depth and segmentation cues intact. All image manipulations were accompanied by appropriate adjustments to ground truth data, where these data were tied directly to the presented image(s). Thus, for pseudoscopic viewing, ground truth data was in direct opposition to binocular disparity-defined depth.

As a further examination of the contribution of binocular image signals, we also trained and tested adapted versions of our network. First, we ran a fully monocular version of our model, based on the U-net architecture (Ronneberger et al., 2015) from which our binocular network is derived. Our U-net architecture focused on segmentation only and used only single images as inputs. There were a total of 3,120,921 parameters in the U-net model, with 3,117,049 trainable parameters and 3,872 nontrainable parameters. The reduction in parameters compared with our binocular network is due to the absence of the depth estimation pathway and the reduction to only a single image input and single segmentation output pathway. The models, and number of equivalent parameters, are identical in all other respects. As a further comparison with the monocular U-net model, we also trained a segmentation only version of our binocular model. This version of our network was trained on binocular images, as with our standard approach, but with learning guided only by the segmentation loss functions.

In addition to training the monocular U-net model and segmentation only version of our binocular model, we also trained and tested our binocular model with identical inputs only. This variation on our approach allowed for a direct comparison of the effects of binocular image presentation within the same model architecture. Further comparison with the results of the U-net model additionally allowed for an assessment of the benefits of binocular viewing even in the absence of binocular disparity signals, for example, through binocular summation (Baker, Lygo, Meese, & Georgeson, 2018; Blake & Fox, 1973; Blake, Sloane, & Fox, 1981).

Results

Figure 3 shows example segmentation and depth estimation outputs from the binocular image trained network, alongside ground truth images. Segmentation performance, including object identification, was highly accurate for binocular images, as was the estimation of depth at each image location. Simultaneous segmentation and depth estimates were also produced in near real-time, providing output images in a processing time of 136 ms (~8 frames on a 60Hz display). Timing estimates were obtained from a machine running the network with a GeForce GTX 1080 GPU.

Depth estimation performance

To quantify model performance for standard binocular image presentation, depth estimation errors were taken as the difference between estimated and ground truth depth values for each point in the image. Errors could vary between ±1, with negative errors indicating over-estimation of relative depth, and positive values indicating under-estimation. Depth errors were calculated on a per-pixel basis, for all images to provide measures of the distribution of depth errors across the test image set. Mean depth errors were 0.016,

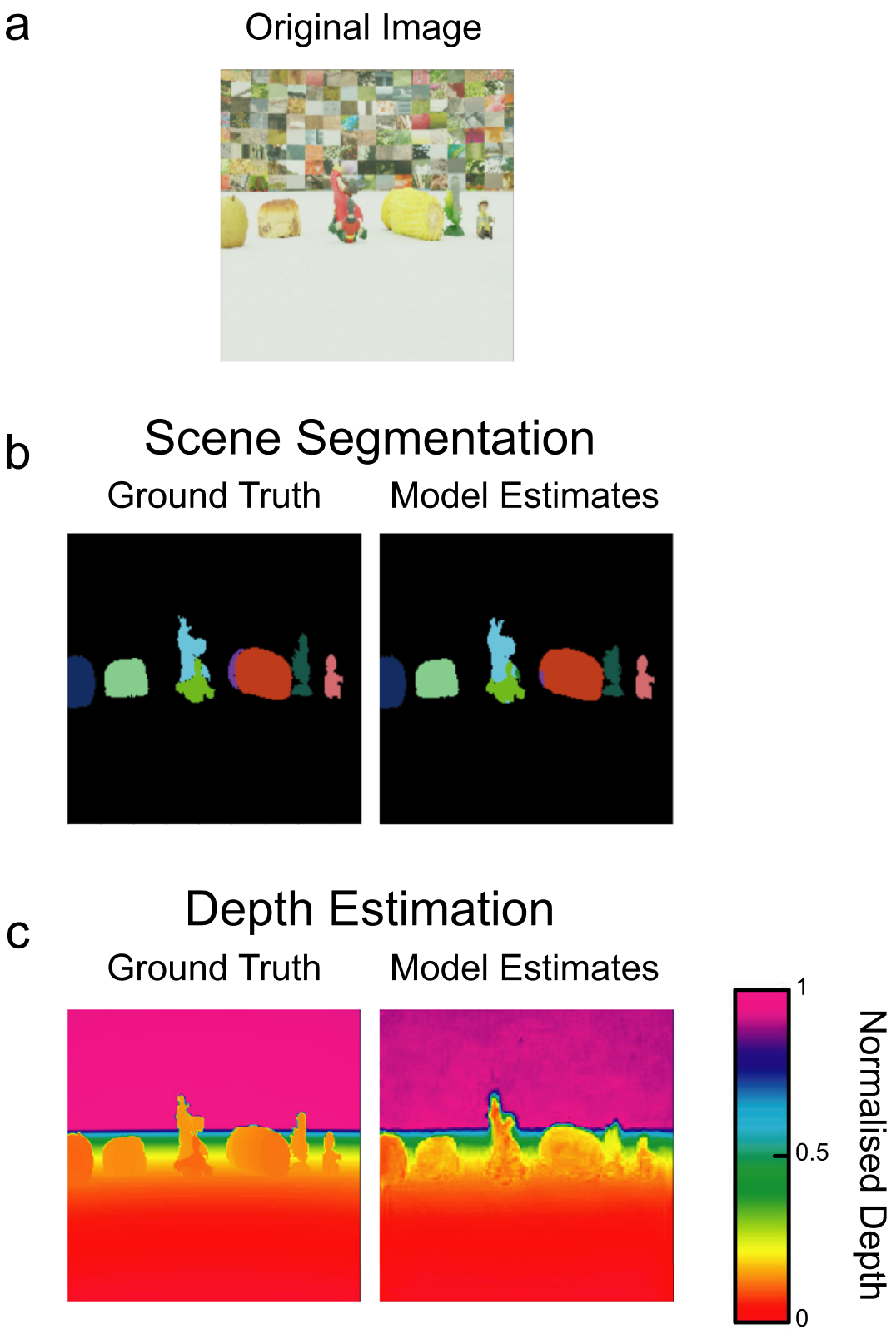

Figure 3. Model outputs for an example image (a), provided with comparison to ground truth images for (b) scene segmentation and (c) depth estimation. Outputs, images, and ground truths are shown for the left image only, because this was the basis for all depth calculations.

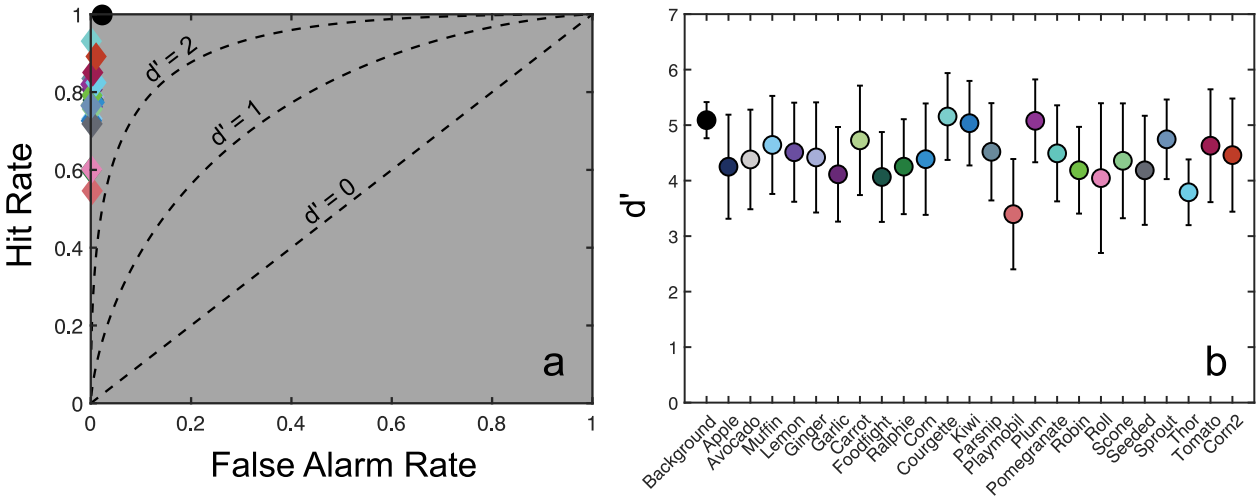

Figure 4. (a) Model segmentation performance for each object, plotted as the Hit rate against the Object False Alarm rate. Better segmentation performance is indicated by points lying closer to the top-left corner of this graph. Curves plotting differing levels of performance as d′ scores are shown for comparison. (b) Average d′ scores for each object in the dataset. *Error bars* show standard deviations, colors are matched to datapoints shown in (a).

with a standard deviation of 0.052, indicating a slight bias for positive (i.e., underestimation) errors. As a further summary of depth errors, we also calculated the average unsigned error (RMSE), which was 0.053 across all images, with a standard deviation of 0.013.

Scene segmentation performance

Pixel-by-pixel segmentation accuracy averaged 98.8% across all 100 test images (min 89.6%, max 99.8%). To further quantify the model's segmentation performance under standard binocular image presentation, we calculated the proportion of pixels on each image correctly identified as belonging to each object class, plus a "background" class (the Hit rate). In addition, we calculated the proportion of pixels mistakenly identified as belonging to each class (the False Alarm rate). There are multiple potential metrics for calculating these False Alarm rates. One may consider the proportion of pixels misidentified as belonging to the target object, either for all non-target pixels (including background pixels), or for non-target pixels that belong to another object (i.e., non-target, non-background pixels). We used this latter, more conservative, measure in all segmentation analyses. We refer to this as the Object False Alarm rate. These results are plotted in Figure 4a.

Segmentation hit rates were typically high (averaging 79% across objects), with consistently very low object false alarm rates, averaging 0.3%. False alarm rates were highest for the background pixels, indicating that target pixels were typically misidentified as belonging to the background. These values were summarized as d′ scores and are plotted for each object in Figure 4b. Average d′ values of 4.4 were found across objects and images. This value can be understood as quantifying the strength of the classification decision information available to the model, relative to the standard deviation of the noise in these decisions.

Quantifying the contribution of binocular signals

The use of different binocular viewing conditions allows for a more detailed understanding of the contributions of binocular signals to depth estimation and scene segmentation performance. We presented our model with correctly arranged left-right images, identical images or left-right swapped, pseudoscopic images. These latter manipulations have the effects of, respectively, removing or reversing binocular depth signals. Mean depth errors and segmentation performance for these models are shown in Figure 5. We also compared segmentation performance to a version of the monocular U-net model (Ronneberger et al., 2015), trained on our dataset. As noted in section 2.3, the segmentation pathway for our model used equivalent processing stages to the U-net model, except for the presence of the binocular convergence stage. As such, it offers a clear comparison for the benefits of binocular presentation for both learning and testing.

Depth estimation errors (Figure 5a) increased from average RMSE values of 0.053 across all 100 test images for the standard model, to 0.072 and 0.110 for identical image and swapped image conditions, respectively. These differences were significant on

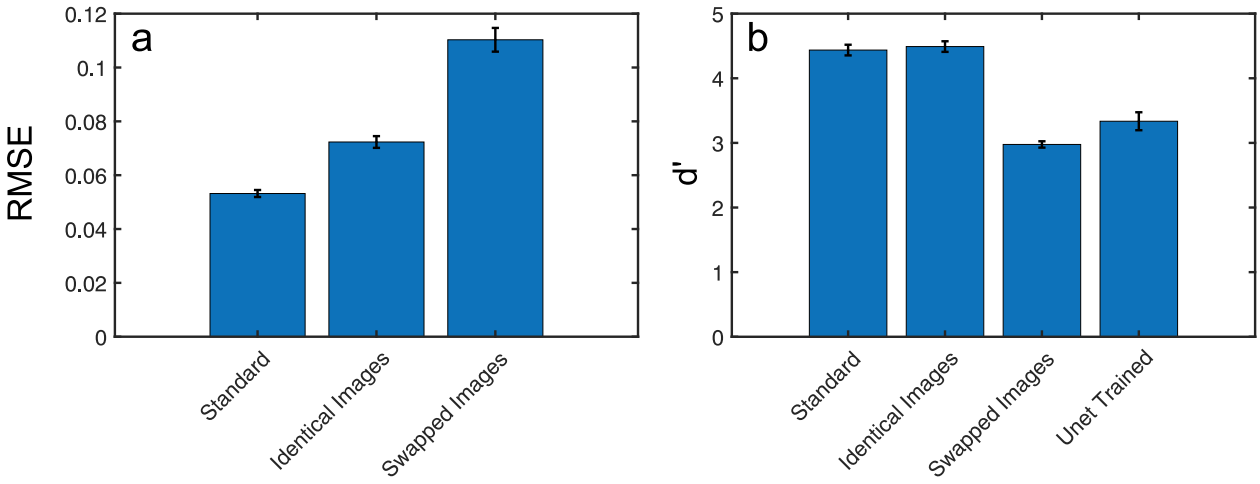

Figure 5. (a) Depth errors (RMSE) for binocular presentation, identical images, and swapped images. Errors increase for non-standard binocular presentation. (b) Segmentation performance plotted as d′ scores across all objects for the same presentation arrangements as in (a), plus the monocular U-net segmentation model. Reductions in d′ scores show poorer segmentation performance for swapped and U-net trained models, compared to standard binocular presentation. *Error bars* show standard errors on the mean.

Bonferroni corrected, two-tailed, related-samples t-tests ($t_{99} = 14.97$, $p < 0.001$, Cohen's $d = 1.5$; $t_{99} = 16.78$, $p < 0.001$, Cohen's $d = 2.0$; $t_{99} = 11.44$, $p < 0.001$, Cohen's $d = 1.4$, for standard-same, standard-swapped, and same-swapped comparisons, respectively). These results indicate that our model is able to use monocular, pictorial depth cues, while also showing sensitivity to eye-of-origin dependent binocular image differences.

The segmentation pathway shows a similar pattern of results (Figure 5b): d′ values were 4.44 for the standard model, 4.49 for identical images and 2.98 for swapped images. Analysis with related-sample *t*-tests, using mean d′ values for each object showed significant differences between standard and swapped presentation ($t_{24} = 17.05$, $p < 0.001$, Cohen's $d = 3.4$) and between standard and identical image presentation ($t_{24} = 3.35$, $p = 0.003$, Cohen's $d = 0.7$). These results indicate that swapped binocular image presentation led to an impairment in segmentation performance while also, surprisingly, showing a slight preference for monocular presentation. This was not the case, however, for testing with the monocularly trained U-net model, where d′ values averaged only 3.34. U-net performance contrasts markedly with the d' values obtained when our binocular model was presented with identical images. Differences were again significant on a two-tailed, related-samples t-test ($t_{24} = 19.41$, $p < 0.001$, Cohen's $d = 3.9$). This result suggests that binocular presentation during training is beneficial for the learning of monocular segmentation signals.

In addition to comparing our model's performance against an equivalent, purely monocular, model, we also conducted a further comparison, where our binocular model was trained on identical images and tested either against identical or against standard binocular images. Depth estimation performance for this identical image trained model is shown, compared to our standard model performance in Figure 6a. Depth estimation errors increased for the identical image trained model, compared to our standard model, with RMSE values rising from 0.053 to 0.059 under identical image testing and 0.063 under binocular image testing. These differences were significant on related-samples *t*-tests ($t_{99} = 3.14$, $p = 0.002$, Cohen's $d = 0.3$; $t_{99} = 4.77$, $p < 0.001$, Cohen's $d = 0.5$). The difference in depth estimation performance for the identical image trained network was also significant for identical, compared with binocular image testing ($t_{99} = 3.66$, $p < 0.001$, Cohen's $d = 0.4$). Note that depth estimation errors were still much lower than for identical image and pseudoscopic viewing in the standard model (see Figure 5a), indicating a dependence on pictorial depth cues in the identical image trained model and further demonstrating the impact of binocular depth cues in our standard network.

Segmentation performance for the identical image trained network is shown under both identical image and binocular viewing conditions in Figure 6b. When tested with identical images, this network showed impaired segmentation performance compared to our standard binocular image trained network, with mean object d' values declining from 4.44 to 4.22 ($t_{24} = 5.38$, $p < 0.001$, Cohen's $d = 1.1$), although these performance levels were significantly greater than found with the monocular U-net ($t_{24} = 11.42$, $p < 0.001$, Cohen's $d = 2.3$). Segmentation performance for our standard model was also better than for the identical image trained model, tested under binocular viewing conditions with mean d' values for the latter falling to 4.14 ($t_{24} = 5.52$,

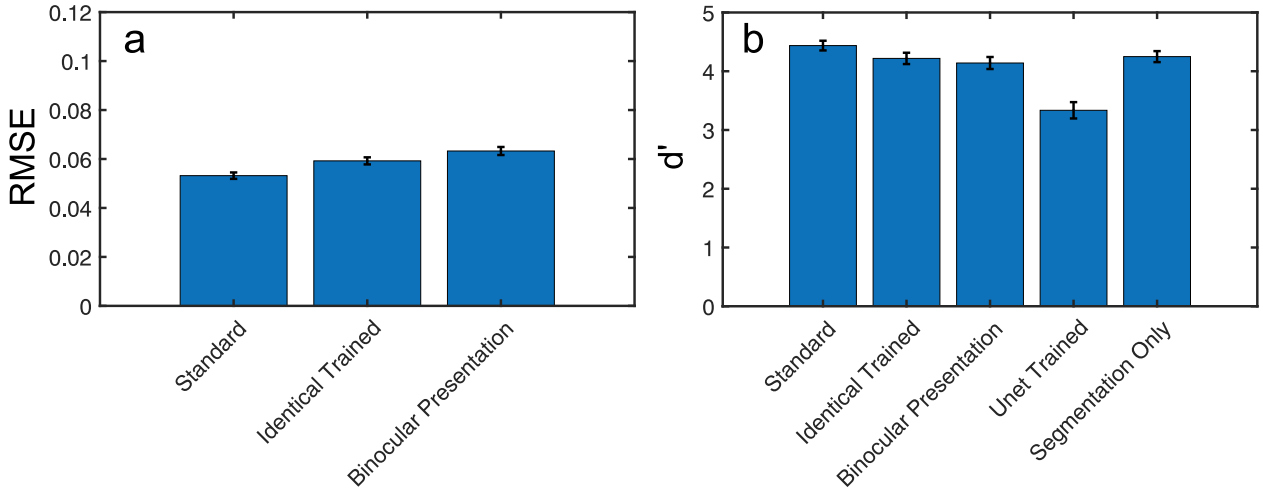

Figure 6. (a) Depth errors (RMSE) for the standard model, plus a version of the model trained and tested on identical image pairs, and one trained on identical image pairs and tested on standard binocular pairs. (b) Segmentation performance plotted as d′ scores for the same sets of models as in (a), plus the monocular U-net model and a version of the network trained on the segmentation task only. *Error bars* show standard errors on the mean.

$p < 0.001$, Cohen's $d = 1.1$). Unlike depth estimation, there was no significant difference in d′ values for the identical image trained model under these two viewing conditions ($t_{24} = 1.80$, $p = 0.085$, Cohen's $d = 0.4$).

The improved performance of our standard model compared to the identical image trained model provides further support for the benefits of binocular image viewing in both depth estimation and scene segmentation. Yet, the improved segmentation scores for this identical image trained model, relative to the monocular U-net model shows that binocular viewing is not the only driver of performance in our network. While one possibility is that these improvements were due to factors such as binocular summation (Baker et al., 2018; Blake & Fox, 1973; Blake et al., 1981), this is not consistent with the lack of impairment in performance when the identical image trained network was tested with standard binocular image pairs. To further examine this issue, we investigated a remaining difference between the monocular U-net model and our binocular network, the presence of a simultaneous depth estimation pathway. We compared the performance of our standard binocular network to a network trained on the segmentation task only (see Figure 6b). Although this binocular trained segmentation network performed substantially better than the monocular U-net, with average d′ values of 4.25, compared to 3.34 ($t_{24} = 10.54$, $p < 0.001$, Cohen's $d = 2.1$ on a related samples t-test), this was still lower than average d′ values found for our standard network ($t_{24} = 5.49$, $p < 0.001$, Cohen's $d = 1.1$). These results suggest that simultaneous training on depth estimation and segmentation is itself beneficial for segmentation performance.

Segmentation and depth estimation

The preceding comparisons of depth estimation and segmentation performance show clear benefits of binocular presentation in both the training and testing of our network. The architecture of our network, with its lateral connections providing inputs from the segmentation pathways into the depth estimation pathway, also allowed for a further comparison, examining the potential role of segmentation information *for* depth estimation. To examine this relationship, we trained and tested our network under an additional condition where these lateral connections were removed. This left the segmentation pathways unchanged but led to significant impairment of depth estimation performance (Figure 7a). RMSE values across images rose from 0.053 with lateral connections, to 0.078 without ($t_{99} = 10.73$, $p < 0.001$, Cohen's $d = 1.1$). Notably, depth estimates from the model without lateral connections varied more smoothly over space, with an absence of sharp depth edges between objects (see Figure 7b).

Discussion

In this article, we have reported the results of a new DNN model for simultaneous depth estimation and scene segmentation, using binocular image inputs. This model shows selectivity for binocular eye-of-origin signals, a necessary prerequisite for the encoding of both binocular disparities and monocular

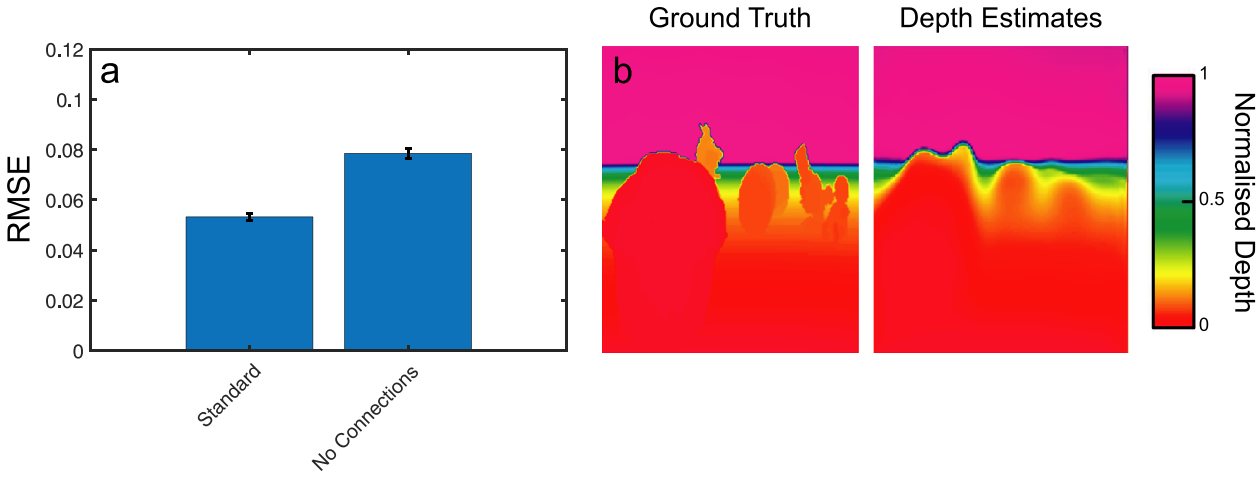

Figure 7. (a) Depth errors (RMSE) for the standard binocular model, as well as a version of the model that removed lateral connections between the network's segmentation pathways and the depth estimation branch. Error bars show standard errors on the mean. (b) An example depth estimation output from the "no lateral connections" model, with accompanying ground truth. Depth estimates were visibly smoothed over space compared with those from the standard binocular model (see Figure 3c).

occlusions. Here, we consider what may be learned about the function of biological visual systems from the performance of this model.

Binocular image training enhances monocular segmentation

There has been significant debate among researchers in biological vision as to the adaptive significance of binocular vision for real world tasks (cf. Heesy, 2009). Although benefits of binocular viewing have been demonstrated in laboratory situations for a number of tasks, including color constancy (Yang & Shevell, 2002), object tracking and visual search (Dunser & Mancero, 2009; Nakayama & Silverman, 1986; Vishwanathan & Mingolla, 2002), and the perception of shape, depth, distance and heading direction (Brenner & van Damme, 1999; McCann, Hayhoe, & Geisler, 2018; Macuga, Loomis, Beall, & Kelly, 2006; Scarfe & Hibbard, 2006; van den Berg & Brenner, 1994), a number of factors suggest that its benefits may be more limited in natural environments. One commonly raised issue is that deficits in binocular vision are relatively commonplace. Coutant and Westheimer (1993) found that 2.7% of a student population sample had stereoacuity thresholds of greater than 2 arcmin, with 20% unable to meet a stricter stereoacuity criterion of 30 arcsec. Other studies have shown dominance for non-binocular visual cues under multi-cue presentation conditions (Allison & Howard, 2000a; Allison & Howard, 2000b; Hill & Johnston, 2007; van Ee, van Dam, & Erkelens, 2002), including in scene recognition tasks (Valsecchi, Caziot, Backus, & Gegenfurtner, 2013).

Here, we provide evidence for the importance of binocular signals for depth estimation and scene segmentation tasks, while also supporting the idea that, for segmentation at least, we typically depend upon the information from monocular signals. For depth estimation, binocular image presentation resulted in lower average errors than either monocular presentation or pseudoscopic presentation. These results clearly show the benefits of adding correctly arranged binocular signals for estimating depth. In addition, however, they also show that the reversal of these binocular signals, although effective in impairing performance, does not entirely override the information provided by monocular depth cues. In this respect our model behaves in a manner consistent with human observers, where pseudoscopic viewing typically only fully reverses depth in abstract stimuli (van den Enden & Spekreijse, 1989). It should be noted, however, that pseudoscopic perception is not particularly well understood with regards to how binocular and pictorial depth cues interact to determine perceived three-dimensional structure. Our model may therefore prove useful in allowing for the generation of predictions on perceived depth and shape under these viewing conditions.

This benefit for binocular presentation extends to scene segmentation tasks. Model segmentation performance for standard binocular image presentation surpassed that found for pseudoscopically presented images. This effect of presentation order shows that the binocular signals encoded by our network are eye-of-origin dependent, a prerequisite for any subsequent measurement of binocular disparity or monocular occlusion. Yet our results are not simply an indication that binocular information is, by default,

critical or beneficial for this task. Two main results complicate this issue. First, identical image presentation offers a small, but significant, benefit over binocular presentation for our model. Second, this benefit for images without binocular disparity information does not extend to a purely monocular version of the segmentation model or to a model trained on identical image pairs.

To understand these issues, one must consider the nature of both monocular and binocular object boundary signals. For binocular cues, although monocular occlusions are effective for signaling object boundaries, they also give rise to the problem of ascribing both depth and object identity to monocular regions. These difficulties could underpin the slightly poorer performance of our model under binocular, compared with identical image, presentation.

A consideration of monocular boundary cues can help to explain the differences in performance between our model, under identical image conditions, and the equivalent, purely monocular, U-net segmentation model. The primary difficulty faced by monocular segmentation processes, is the requirement to differentiate between pattern edges and object edges. As noted, binocular presentation offers a powerful means to resolve this problem through the addition of binocular disparity-defined edges and monocular occlusions. The benefits of these binocular signals are, potentially, twofold. First, they can directly contribute to boundary detection, as under standard binocular viewing conditions. They may, however, also offer systems a means to learn how to distinguish between monocular signals associated with boundary edges and pattern edges. In this way, binocular signals may hold adaptive value not just as segmentation signals in and of themselves, but as a means to support the development of such abilities using other image cues. This finding offers potential insight into developmentally acquired perceptual impairments, such as amblyopia, where disruptions in the development of binocular vision lead to perceptual deficits in both amblyopic and fellow eyes (Parker, Smith, & Krug, 2016; Simmers & Bex, 2004; Simmers, Ledgeway, Hess, & McGraw, 2003). Relatedly, impairments in scene segmentation abilities have also been reported in human patients recovering from early blindness, with signals from motion-defined boundaries seemingly critical for the development of these abilities (Ostrovsky, Meyers, Ganesh, Mathur, & Sinha, 2009).

Learning depth estimation supports enhanced segmentation

By manipulating the arrangement of binocular image pairs in the training and testing of our network, we have shown that the binocular viewing of scenes enhances segmentation performance and supports the further development of monocular segmentation capabilities. Binocular viewing offers these enhancements whether compared to a purely monocular model, or to a model with the same binocular architecture, trained on identical image pairs. Although our monocular comparison model shows poorer segmentation performance than the identical image trained model, further comparison with a version of our binocular model trained only on the segmentation task suggests that this may be due to a beneficial effect of our dual task training. This suggests that training on both depth estimation and scene segmentation leads to depth-related differences, through the feedback of depth errors into the initial left-right and binocular encoding layers, with these differences acting as beneficial cues for segmentation. This raises the intriguing possibility that information relevant for the scene segmentation task is learned by the network only by virtue of its initial relevance for another task.

Scene segmentation in the brain

Although any attempts to compare DNN architectures and model weights to human and nonhuman animal physiology should be made with extreme caution (Lopez-Rubio, 2018; Majaj & Pelli, 2018), it is worth considering where similarities do exist and how this may inform our knowledge of biological visual systems. Critically, we do not view any similarities between the structure of our network and mammalian physiology as essential for our findings to hold relevance for our understanding of biological visual systems. Although the learning that occurred within our network need not necessarily translate to the information learned by biological systems, the benefits of binocular viewing for segmentation and depth estimation in our network may still be considered as hypotheses on the likely pattern of sensitivities to be found in such biological systems. The model reported here offers the possibility for two comparisons to biological vision. First, the architecture of our network maps onto several known aspects of the physiology of mammalian visual systems. As in such biological systems, our network begins with monocular encoding stages, followed by a point of binocular convergence. After this point the network proceeds along specialized pathways for segmentation and depth estimation.

This specialization is similar, in some respects, to the divergence of dorsal and ventral pathways in humans (cf., de Hann & Cowey, 2011). Although binocular processing occurs in multiple areas of the human brain, it does so in very different ways with, for example, ventral areas V4, TE and IT showing selectivity for both surface segmentation and descriptors of surface shape (Verhoef, Vogels, & Janssen, 2016). That our

model is able to simultaneously learn and perform both segmentation and depth estimation tasks shows the suitability of our common encoding pathway as a generalized stage of visual processing. As noted in section 4.2, there may be additional benefits to these common early processing stages, where selectivity for signals may develop in response to specific task requirements but may also then prove useful for a range of other tasks.

A second point of comparison between our network and human vision comes in the use of connections between our specialized segmentation and depth estimation output pathways. These connections enhance depth estimation performance, with RMSE values for depth estimation increasing significantly in their absence. Depth estimates without lateral connections were also notably poorer across object boundaries, with a visible smoothing of depth between objects. Not only do these lateral connections map onto those found between functionally specialized areas in the mammalian brain (Cloutman, 2013; de Haan & Cowey, 2011; Schenk & McIntosh, 2010), their measured effects also provide a qualitative match to the psychophysical behavior of human observers in a number of depth perception tasks (Cammack & Harris, 2016; Deas & Wilcox, 2014; Deas & Wilcox, 2015; Goutcher et al., 2018; Goutcher & Wilcox, 2021; Mamassian & Zannoli, 2020; Wardle & Gillam, 2016). This psychophysical work has demonstrated the importance of segmentation boundaries on perceived depth, with Mamassian and Zannoli (2020) suggesting that object boundaries are used to delimit depth averaging processes. The benefits of lateral connections between segmentation and depth estimation pathways in our network support this suggestion.

Conclusions

The novel DNN reported in this article shows high levels of accuracy in scene segmentation and depth estimation tasks and is capable of learning and performing these dual tasks simultaneously. By manipulating the arrangement of binocular images, we have shown that this model encodes eye-of-origin information, necessary for the encoding of both binocular disparities and monocular half-occlusions. Results from these manipulations show that binocular signals are informative for both tasks and may play a role in supporting the learning of monocular segmentation signals. They further show that the simultaneous learning of depth estimation and segmentation tasks may in itself be beneficial for the development of sensitivity to depth-related segmentation cues. These findings provide a further example to illustrate the usefulness of deep learning approaches as a tool for understanding biological vision, providing evidence on where and how stimulus information can be used and recovered by suitable network structures.

Keywords: deep learning, binocular vision, segmentation, depth perception

Acknowledgments

Supported by DASA/DSTL Grant No. ACC6012885 to RG and BBSRC Grant No. BB/K018973/1 to PBH.

Commercial relationships: none.
Corresponding author: Ross Goutcher.
Email: ross.goutcher@stir.ac.uk.
Address: Psychology Division, Faculty of Natural Sciences, University of Stirling, Cottrell Building, Stirling, FK9 4LA, UK.

References

Adams, W. J., Elder, J. H., Graf, E. W., Leyland, J., Lugtigheid, A. J., & Muryy, A. (2016). The Southampton-York natural scenes (syns) dataset: Statistics of surface attitude. *Scientific Reports, 6*(1), 1–17.

Allison, R. S., & Howard, I. P. (2000a). Temporal dependencies in resolving monocular and binocular cue conflict in slant perception. *Vision Research, 40*(14), 1869–1885.

Allison, R. S., & Howard, I. P. (2000b). Stereopsis with persisting and dynamic textures. *Vision Research, 40*(28), 3823–3827.

Baker, D. H., Lygo, F. A., Meese, T. S., & Georgeson, M. A. (2018). Binocular summation revisited: Beyond$\sqrt{2}$. *Psychological Bulletin, 144*(11), 1186.

Banks, M. S., Gepshtein, S., & Landy, M. S. (2004). Why is spatial stereo resolution so low? *Journal of Neuroscience, 24*(9), 2077–2089.

Barlow, H. B. (1962). A method of determining the overall quantum efficiency of visual discriminations. *The Journal of Physiology, 160*(1), 155–168.

Basgöze, Z., White, D. N., Burge, J., & Cooper, E. A. (2020). Natural statistics of depth edges modulate perceptual stability. *Journal of Vision, 20*(8):10, 1–21, https://doi.org/10.1167/jov.20.8.10.

Birchfield, S., & Tomasi, C. (1999). Depth Discontinuities by Pixel-to-Pixel Stereo. *International Journal of Computer Vision, 35*(3), 269–293.

Blake, R., & Fox, R. (1973). The psychophysical inquiry into binocular summation. *Perception & Psychophysics, 14*(1), 161–185.

Blake, R., Sloane, M., & Fox, R. (1981). Further developments in binocular summation. *Perception & Psychophysics, 30*(3), 266–276.

Bredfeldt, C. E., Read, J. C. A., & Cumming, B. G. (2009). A quantitative explanation of responses to disparity-defined edges in macaque V2. *Journal of Neurophysiology, 101*(2), 701–713.

Brenner, E., & van Damme, W. J. (1999). Perceived distance, shape and size. *Vision Research, 39*(5), 975–986.

Burge, J., & Geisler, W. S. (2014). Optimal disparity estimation in natural stereo images. *Journal of Vision, 14*(2), 1–1.

Cammack, P., & Harris, J. M. (2016). Depth perception in disparity-defined objects: finding the balance between averaging and segregation. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1697), 20150258.

Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning, 107*(8), 1477–1494.

Chauhan, T., Masquelier, T., Montlibert, A., & Cottereau, B. R. (2018). Emergence of Binocular Disparity Selectivity Through Hebbian Learning. *Journal of Neuroscience, 38*(44), 9563–9578.

Cloutman, L. L. (2013). Interaction between dorsal and ventral processing streams: Where, when and how? *Brain & Language, 127*, 251–263.

Coutant, B. E., & Westheimer, G. (1993). Population distribution of stereoscopic ability. *Ophthalmic and Physiological Optics, 13*(1), 3–7.

Cutting, J. E., & Vishton, P. M. (1995). Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In W. Epstein, & S. Rogers (Eds.) *Perception of Space and Motion* (pp. 69–117). Cambridge, MA: Academic Press.

Deas, L. M., & Wilcox, L. M. (2014). Gestalt grouping via closure degrades suprathreshold depth percepts. *Journal of Vision, 14*(9), 14–14.

Deas, L. M., & Wilcox, L. M. (2015). Perceptual grouping via binocular disparity: The impact of stereoscopic good continuation. *Journal of Vision, 15*(11), 11–11.

DeAngelis, G. C., Ohzawa, I., & Freeman, R. D. (1991). Depth is encoded in the visual cortex by a specialized receptive field structure. *Nature, 352*(6331), 156–159.

de Haan, E. H., & Cowey, A. (2011). On the usefulness of 'what' and 'where' pathways in vision. *Trends in Cognitive Sciences, 15*(10), 460–466.

Dünser, A., & Mancero, G. (2009). The use of depth in change detection and multiple object tracking. *The Ergonomics Open Journal, 2*(1).

Ecke, G. A., Papp, H. M., & Mallot, H. A. (2021). Exploitation of image statistics with sparse coding in the case of stereo vision. *Neural Networks, 135*, 158–176.

Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., & Burgard, W. (2015, September). Multimodal deep learning for robust RGB-D object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 681–687). IEEE.

Elder, J. H., Krupnik, A., & Johnston, L. A. (2003). Contour grouping with prior models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(6), 661–674.

Elder, J. H. (2018). Shape from contour: Computation and representation. *Annual Review of Vision Science, 4*, 423–450.

Fleet, D. J., Wagner, H., & Heeger, D. J. (1996). Neural encoding of binocular disparity: energy models, position shifts and phase shifts. *Vision Research, 36*, 1839–1857.

Fowlkes, C. C., Martin, D. R., & Malik, J. (2007). Local figure–ground cues are valid for natural images. *Journal of Vision, 7*(8), 2–2.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., & Garcia-Rodriguez, J. (2017). A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*.

Goncalves, N. R., & Welchman, A. E. (2017). "What not" detectors help the brain see in depth. *Current Biology, 27*(10), 1403–1412.

Goutcher, R., & Hibbard, P. B. (2010). Evidence for relative disparity matching in the perception of an ambiguous stereogram. *Journal of Vision, 10*(12), 35–35.

Goutcher, R. (2016). Motion direction influences surface segmentation in stereo transparency. *Journal of Vision, 16*(15), 17–17.

Goutcher, R., Connolly, E., & Hibbard, P. B. (2018). Surface continuity and discontinuity bias the perception of stereoscopic depth. *Journal of Vision, 18*(12), 13–13.

Goutcher, R., & Hibbard, P. B. (2014). Mechanisms for similarity matching in disparity measurement. *Frontiers in Psychology, 4*, 1014.

Goutcher, R., & Wilcox, L. M. (2016). Representation and measurement of stereoscopic volumes. *Journal of Vision, 16*(11), 16–16.

Goutcher, R., & Wilcox, L. M. (2021). Surface slant impairs disparity discontinuity discrimination. *Vision Research, 180*, 37–50.

Harris, J. M. (2014). Volume perception: Disparity extraction and depth representation in complex three-dimensional environments. *Journal of Vision, 14*(12), 11–11.

Harris, J. M., & Wilcox, L. M. (2009). The role of monocularly visible regions in depth and surface perception. *Vision Research, 49*(22), 2666–2685.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity Mapping in Deep Residual Networks. In *European Conference on Computer Vision* (pp. 630–645). Cham: Springer.

Heesy, C. P. (2009). Seeing in stereo: the ecology and evolution of primate binocular vision and stereopsis. *Evolutionary Anthropology: Issues, News, and Reviews, 18*(1), 21–35.

Henriksen, S., Read, J. C. A., & Cumming, B. G. (2016). Neurons in striate cortex signal disparity in half-matched random-dot stereograms. *Journal of Neuroscience, 36*(34), 8967–8976.

Hibbard, P., Scarfe, P., Hornsey, R., & Hunter, D. (2016). A 3D database of everyday objects for vision research. *Journal of Vision, 16*(12), 289–289

Hildreth, E. C., & Royden, C. S. (2011). Integrating multiple cues to depth order at object boundaries. *Attention, Perception, & Psychophysics, 73*(7), 2218–2235.

Hill, H., & Johnston, A. (2007). The hollow-face illusion: Object-specific knowledge, general assumptions or properties of the stimulus?. *Perception, 36*(2), 199–223.

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. *Science, 298*(5598), 1627–1630.

Hirschmuller, H. (2005). Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. *2*, pp. 807–814). IEEE.

Howard, I. P., & Rogers, B. J. (2012). *Perceiving in Depth.* Oxford, UK: OUP.

Huang, P. H., Matzen, K., Kopf, J., Ahuja, N., & Huang, J. B. (2018). Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2821–2830).

Hunter, D. W., & Hibbard, P. B. (2018). The effect of image position on the Independent Components of natural binocular images. *Scientific Reports, 8*(1), 1–15.

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., & Bengio, Y. (2017). The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 11–19).

Kingdom, F. A., Yared, K. C., Hibbard, P. B., & May, K. A. (2020). Stereoscopic depth adaptation from binocularly correlated versus anti-correlated noise: Test of an efficient coding theory of stereopsis. *Vision Research, 166*, 60–71.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science, 1*, 417–446.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM, 60*(6), 84–90.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., & Hu, X. (2005). Support vector machines for temporal classification of block design fMRI data. *NeuroImage, 26*(2), 317–329.

Langer, M. S., Zheng, H., & Rezvankhah, S. (2016). Depth discrimination from occlusions in 3D clutter. *Journal of Vision, 16*(11), 11–11.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3431–3440).

López-Rubio, E. (2018). Computational functionalism for the deep learning era. *Minds and Machines, 28*(4), 667–688.

Lovell, P. G., Bloj, M., & Harris, J. M. (2012). Optimal integration of shading and binocular disparity for depth perception. *Journal of Vision, 12*(1), 1–1.

McCann, B. C., Hayhoe, M. M., & Geisler, W. S. (2018). Contributions of monocular and binocular cues to distance discrimination in natural scenes. *Journal of Vision, 18*(4), 12–12.

Macuga, K. L., Loomis, J. M., Beall, A. C., & Kelly, J. W. (2006). Perception of heading without retinal optic flow. *Perception & Psychophysics, 68*(5), 872–878.

Maiello, G., Chessa, M., Bex, P. J., & Scolari, F. (2020). Near-optimal combination of disparity across log-polar scaled visual field. *PLoS Computational Biology, 16*(4): e1007699.

Majaj, N. J., & Pelli, D. G. (2018). Deep learning—Using machine learning to study biological vision. *Journal of Vision, 18*(13), 2–2.

Mamassian, P., & Zannoli, M. (2020). Sensory loss due to object formation. *Vision Research, 174*, 22–40.

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences, 207*(1167), 187–217.

Marr, D., & Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London. Series B. Biological Sciences, 204*(1156), 301–328.

Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(5), 530–549.

May, K. A., Zhaoping, L., & Hibbard, P. B. (2012). Perceived direction of motion determined by adaptation to static binocular images. *Current Biology, 22*(1), 28–32.

Nakayama, K., & Shimojo, S. (1990). Da Vinci stereopsis: Depth and subjective occluding contours from unpaired image points. *Vision Research, 30*(11), 1811–1825.

Nakayama, K., & Silverman, G. H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature, 320*(6059), 264–265.

Olmos, A., & Kingdom, F. A. A. (2004). A biologically inspired algorithm for the recovery of shading and reflectance images, *Perception, 33*, 1463–1473.

Ostrovsky, Y., Meyers, E., Ganesh, S., Mathur, U., & Sinha, P. (2009). Visual parsing after recovery from blindness. *Psychological Science, 20*(12), 1484–1491.

Parker, A. J. (2007). Binocular depth perception and the cerebral cortex. *Nature Reviews Neuroscience, 8*(5), 379–391.

Parker, A. J., Smith, J. E., & Krug, K. (2016). Neural architectures for stereo vision. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1697), 20150261.

Pelli, D. G. (1990). The quantum efficiency of vision. In C. Blakemore (Eds). *Vision: Coding and Efficiency* (pp. 3–24), Cambridge, UK: Cambridge University Press.

Pelli, D. G., & Farell, B. (1999). Why use noise?. *Journal of the Optical Society of America A, 16*(3), 647–653.

Qian, N., & Zhu, Y. (1997). Physiological computation of binocular disparity. *Vision Research, 37*(13), 1811–1827.

Read, J. C. A., & Cumming, B. G. (2007). Sensors for impossible stimuli may solve the stereo correspondence problem. *Nature Neuroscience, 10*(10), 1322–1328.

Richards, B. A., Lillicrap, T. P., & Beaudoin, P. et al. (2019) A deep learning framework for neuroscience. *Nature Neuroscience, 22*, 1761–1770.

Rideaux, R., & Welchman, A. E. (2020). But still it moves: static image statistics underlie how we see motion. *Journal of Neuroscience, 40*(12), 2538–2552.

Rideaux, R., & Welchman, A. E. (2021). Exploring and explaining properties of motion processing in biological brains using a neural network. *Journal of Vision, 21*(2), 11–11.

Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241). Cham: Springer.

Scarfe, P., & Hibbard, P. B. (2006). Disparity-defined objects moving in depth do not elicit three-dimensional shape constancy. *Vision Research, 46*(10), 1599–1610.

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision, 47*(1), 7–42.

Schenk, T., & McIntosh, R. D. (2010). Do we have independent visual streams for perception and action? *Cognitive Neuroscience, 1*(1), 52–62.

Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012, October). Indoor segmentation and support inference from rgbd images. In *European conference on Computer Vision* (pp. 746–760). Berlin: Springer.

Simmers, A. J., & Bex, P. J. (2004). The representation of global spatial structure in amblyopia. *Vision Research, 44*(5), 523–533.

Simmers, A. J., Ledgeway, T., Hess, R. F., & McGraw, P. V. (2003). Deficits to global motion processing in human amblyopia. *Vision Research, 43*(6), 729–738.

Smolyanskiy, N., Kamenev, A., & Birchfield, S. (2018). On the importance of stereo for accurate depth estimation: An efficient semi-supervised deep neural network approach. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018*, 1007–1015.

Song, S., Lichtenberg, S. P., & Xiao, J. (2015). Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 567–576).

Srinath, R., Emonds, A., Wang, Q., Lempel, A. A., Dunn-Weiss, E., Connor, C. E., … Nielsen, K. (2020). Early emergence of solid shape coding in natural and deep network vision. *Current Biology*. *31*, 1–15.

Sundberg, P., Brox, T., Maire, M., Arbeláez, P., & Malik, J. (2011, June). Occlusion boundary detection and figure/ground assignment from optical flow. In *CVPR 2011* (pp. 2233–2240). IEEE.

Tsirlin, I., Wilcox, L. M., & Allison, R. S. (2010). Monocular occlusions determine the perceived shape and depth of occluding surfaces. *Journal of Vision, 10*(6), 11–11.

Valsecchi, M., Caziot, B., Backus, B. T., & Gegenfurtner, K. R. (2013). The role of binocular disparity in rapid scene and pattern recognition. *i-Perception, 4*(2), 122–136.

Van den Berg, A. V., & Brenner, E. (1994). Why two eyes are better than one for judgements of heading. *Nature, 371*(6499), 700–702.

Van Den Enden, A., & Spekreijse, H. (1989). Binocular depth reversals despite familiarity cues. *Science, 244*(4907), 959–961.

Van Ee, R., Van Dam, L. C., & Erkelens, C. J. (2002). Bi-stability in perceived slant when binocular disparity and monocular perspective specify different slants. *Journal of Vision, 2*(9), 2–2.

Verhoef, B. E., Vogels, R., & Janssen, P. (2016). Binocular depth processing in the ventral visual pathway. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1697), 20150259.

Viswanathan, L., & Mingolla, E. (2002). Dynamics of attention in depth: Evidence from multi-element tracking. *Perception, 31*(12), 1415–1437.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., … von der Heydt, R. (2012). A Century of Gestalt Psychology in Visual Perception I. Perceptual Grouping and Figure-Ground Organization. *Psychological Bulletin, 138*(6), 1172–1217

Wang, K., & Shen, S. (2018, September). Mvdepthnet: Real-time multiview depth estimation neural network. In *2018 International conference on 3d vision (3DV)* (pp. 248–257). IEEE.

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., & Torr, P. H. (2019). Fast online object tracking and segmentation: A unifying approach. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019*, 1328–1338.

Wardle, S. G., Palmisano, S., & Gillam, B. J. (2014). Monocular and binocular edges enhance the perception of stereoscopic slant. *Vision Research, 100*, 113–123.

Wardle, S. G., & Gillam, B. J. (2016). Gradients of relative disparity underlie the perceived slant of stereoscopic surfaces. *Journal of Vision, 16*(5), 16–16.

Watt, R., Ledgeway, T., & Dakin, S. C. (2008). Families of models for Gabor paths demonstrate the importance of spatial adjacency. *Journal of Vision, 8*(7), 23–23.

Welchman, A. E. (2016). The human brain in depth: how we see in 3D. *Annual Review of Vision Science, 2*, 345–376.

Wheatstone, C. (1852). I. The Bakerian Lecture – Contributions to the physiology of vision: On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions of the Royal Society of London*, (142), 1–17.

Yang, J. N., & Shevell, S. K. (2002). Stereo disparity improves color constancy. *Vision Research, 42*(16), 1979–1989.