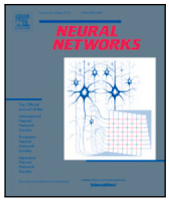
Neural Networks 181 (2025) 106758

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Exploring the categorical nature of colour perception: Insights from artificial networks

Arash Akbarinia

Department of Experimental Psychology, University of Giessen, Germany

ARTICLE INFO

Keywords:
Colour categories
Colour naming
Colour perception
Deep neural networks
Artificial psychophysics

ABSTRACT

The electromagnetic spectrum of light from a rainbow is a continuous signal, yet we perceive it vividly in several distinct colour categories. The origins and underlying mechanisms of this phenomenon remain partly unexplained. We investigate categorical colour perception in artificial neural networks (ANNs) using the odd-one-out paradigm. In the first experiment, we compared unimodal vision networks (e.g., ImageNet object recognition) to multimodal vision-language models (e.g., CLIP text-image matching). Our results show that vision networks predict a significant portion of human data (approximately 80%), while vision-language models account for the remaining unexplained data, even in non-linguistic experiments. These findings suggest that categorical colour perception is a language-independent representation, though it is partly shaped by linguistic colour terms during its development. In the second experiment, we explored how the visual task influences the colour categories of an ANN by examining twenty-four Taskonomy networks. Our results indicate that human-like colour categories are task-dependent, predominantly emerging in semantic and 3D tasks, with a notable absence in low-level tasks. To explain this difference, we analysed kernel responses before the winner-takes-all stage, observing that networks with mismatching colour categories may still align in underlying continuous representations. Our findings quantify the dual influence of visual signals and linguistic factors in categorical colour perception and demonstrate the task-dependent nature of this phenomenon, suggesting that categorical colour perception emerges to facilitate certain visual tasks.

1. Introduction

The electromagnetic spectrum of light from a rainbow reaches our eyes as a continuous signal, yet our conscious perception of the rainbow divides it into discrete categories, such as red, orange, yellow, green, blue, indigo, and violet—a phenomenon known as categorical perception (Harnad, 2003). The underlying mechanism of this phenomenon involves a multitude of perceptual and cognitive processes (Witzel, 2019). We have a good understanding of the early visual processes. In the initial stage of visual processing, cone photoreceptors absorb light at different wavelengths, creating a three-dimensional space often modelled by LMS colour space (i.e., Long-, Medium-, and Short-wavelength sensitive cones) (Gegenfurtner & Sharpe, 1999). In the subsequent stage, the output of cone photoreceptors is combined antagonistically, as modelled by the DKL colour space (i.e., Derrington-Krauskopf-Lennie) (Derrington, Krauskopf, & Lennie, 1984). At these neurophysiological levels, visual processing operates in the continuous domain, and it is estimated that the human eye can discern nearly two million colours (Linhares, Pinto, & Nascimento, 2008; Pointer & Attridge, 1998). In the later visual processes, the enormous complexity of the visual scene is simplified into cognitively manageable elements by segregating foreground objects from the background, and it is estimated that average subjects can recall about thirty colour categories (Derefeldt & Swartling, 1995). However, there remains a significant gap in our understanding of the intermediate operations in colour processing that connect a vast continuum of colours to a small set of discrete categories.

This gap has been a topic of discussion among scientists, particularly concerning the role of language in the formation of colour categories, resulting in two competing theories: universalism and relativism. Universalists argue that the mechanism underpinning categorical colour perception is an inherent aspect of physiological processes. This idea was initially suggested by Berlin and Kay (1969), who showed that eleven colour categories are remarkably similar across all cultures and languages. This view was hotly debated by relativists, associated with Benjamin Lee Whorf (Kay & Kempton, 1984), arguing that perception is shaped by the semantic categories of one's native language (Davidoff, 2001). They highlight challenges such as the difficulty children face in acquiring colour names (Roberson, Davidoff, Davies, & Shapiro, 2004), variations in colour terms across languages (Roberson, Davies,

E-mail address: arash.akbarinia@psychol.uni-giessen.de.

https://doi.org/10.1016/j.neunet.2024.106758
Received 3 June 2024; Received in revised form 26 August 2024; Accepted 23 September 2024
Available online 30 September 2024

& Davidoff, 2000), an instance of a patient with language impairments showing colour sorting problems (Roberson, Davidoff, & Braisby, 1999), and the absence of colour categories in some primates (Fagot, Goldstein, Davidoff, & Pickering, 2006). Nonetheless, scientific consensus leans towards moderate universality: universal patterns exist beyond superficial discrepancies across different cultures (Kay & Regier, 2006). This is supported by a large overlap of focal colours across diverse cultures (Regier, Kay, & Cook, 2005) and findings from non-verbal experiments (Indow, 1988), free categorisation tasks (Indow & Kanazawa, 1960), visual searches (Cropper, Kvansakul, & Little, 2013), and studies indicating categorical colour perception in pre-linguistic infants (Franklin, Pilling, & Davies, 2005; Skelton, Catchpole, Abbott, Bosten, & Franklin, 2017).

Despite the extensive body of literature on this topic, two key questions remain to be thoroughly addressed. First, we need to disentangle the bottom-up and top-down processes in categorical colour perception (Witzel & Gegenfurtner, 2018) and quantify the contributions of linguistic and perceptual processing to this phenomenon. Second, if the universalism theory is favoured, it remains unclear why our visual system adopts a categorical colour representation—whether this is due to the neural circuitry of our system or linked to the visual tasks we perform (de Vries, Akbarinia, Flachot, & Gegenfurtner, 2022). In this article, we aim to contribute to answering these questions by adopting a computational modelling approach, specifically leveraging the capabilities of artificial neural networks (ANNs), which possess sufficient complexity to emulate the ecological validity of human observers while remaining amenable to controlled experiments.

Computational modelling approaches to colour categorisation have a long-standing history in the literature. The visual mechanisms underlying colour categorisation have been modelled using continuous mathematics, such as fuzzy logic. Early studies fitted human data to variations of Gaussian functions (Mojsilovic, 2005). Subsequently, the k-means algorithm emerged as a natural choice for clustering psychophysical colour points (Seaborn, Hepplewhite, & Stonham, 2005). Other approaches that have been explored include the triple-sigmoidal parametric model (Benavente, Vanrell, & Baldrich, 2008) and probabilistic latent semantic indexing (van de Weijer, Schmid, Verbeek, & Larlus, 2009). More recently, three-dimensional ellipsoids have been proposed to partition the colour space into distinct colour categories (Parraga & Akbarinia, 2016). Currently, artificial neural networks are receiving considerable attention for investigating colour categories. Chaabouni, Kharitonov, Dupoux, and Baroni (2021) demonstrated that the accuracy-complexity trade-off in human colour terms emerges in two artificial agents playing a communication game. This finding aligns with efficient communication theory, which asserts that human colour categories closely approach the theoretically optimal limit (Gibson et al., 2017; Zaslavsky, Kemp, Regier, & Tishby, 2018), thereby reinforcing the pivotal role of language in shaping colour categories. In a contrasting approach, de Vries et al. (2022) illustrated that colour boundaries reported by human observers manifest in object recognition networks trained on natural images without any language component. This observation aligns with categorical perception theory, asserting that perceptual colour space is warped by stretching at category boundaries or by within-category compression (Bornstein & Korda, 1984; Witzel & Gegenfurtner, 2018).

In this study, we adopt a different approach from Chaabouni et al. (2021) and de Vries et al. (2022) by employing the linear probes technique (Alain & Bengio, 2017), trained on a generic colour discrimination task that is independent of the specific colour categorisation experiments we conducted. Our analysis focuses on two main experiments: (1) comparing multimodal vision-language and unimodal vision deep neural networks to dissect the contribution of each modality, and (2) examining the representations within an identical architecture (ResNet50) trained on different visual tasks to explore whether the system's functional role influences categorical colour representation. Our findings can be summarised in three key points:

1. **Role of language**: unimodal vision models, such as ImageNet object recognition networks, explain over eighty per cent of human data, with the remaining portion attributed to multimodal vision-language models, such as CLIP text-image matching networks. This suggests that categorical colour perception constitutes a language-independent representation, despite the significant influence that linguistic colour terms have on its developmental trajectory.
2. **Effect of visual task**: human-like colour categories predominantly emerge in models trained on semantic visual tasks, including image segmentation, object recognition, and scene classification. Networks optimised for 3D tasks exhibit moderately human-like colour categories, while those focused on 2D low-level tasks, such as autoencoding and denoising, fail to reproduce human-like colour categories.
3. **Internal representation**: The continuous representation of colour distribution remains rather similar across a network's layers, diverging into distinct colour categories only after the winner-take-all process partitions the internal space. This divergence typically occurs at a mid-level process, suggesting that categorical colour perception is shaped by mid-level visual representations, which likely facilitates certain visual tasks.

2. Methods

All research materials, including the source code for training/testing artificial neural networks and analysing the data, are openly accessible on our GitHub project page: https://arashakbarinia.github.io/projects/colourcats/.

2.1. Stimuli

To train and test the concept of colour categories in ANNs, we used images with a resolution of 224 × 224 pixels, consistent with the image resolution used during the pretrained stage of all examined networks. Each image stimulus consists of two uniformly coloured components: foreground and background (see Fig. 1). Foreground shapes were systematically selected from a set of 2904 geometrical shapes (refer to Appendix A for details). During training, the colours of the foreground shapes were selected from a uniform random distribution, while the background was always achromatic, also selected from a random uniform distribution (see below for details). At test time, foreground colours were systematically changed to different Munsell chips, and the background was always set to a mid-grey value ($R = G = B = 128$). Before inputting into a pretrained network, the images are normalised to the expected range of values for that network.

2.2. Comparison to human colour categories

We systematically investigated the categorical colour representation within artificial neural networks, utilising Munsell chips (see panel **d** in Fig. 1). This set gained prominence through its inclusion in the World Colour Survey (WCS) (Kay, Berlin, Maffi, Merrifield, & Cook, 2009) and is frequently employed in colour category literature (e.g., Chaabouni et al., 2021; Parraga & Akbarinia, 2016; Regier, Kay, & Khetarpal, 2007; Zaslavsky et al., 2018). In our experiments, we compared the networks to the human data from Berlin and Kay (1969) and Sturges and Whitfield (1995) on eight universal colour categories (pink, red, orange, brown, yellow, green, blue, and purple). Achromatic colour terms (white, grey, and black) were excluded because our stimuli background is achromatic during training and specifically grey at test time, resulting in grey Munsell chips not being visible in the image. The reported accuracies represent the average over the union of ground-truths provided by these two studies, which contains 209 Munsell chips (see panel **e** in Fig. 1).

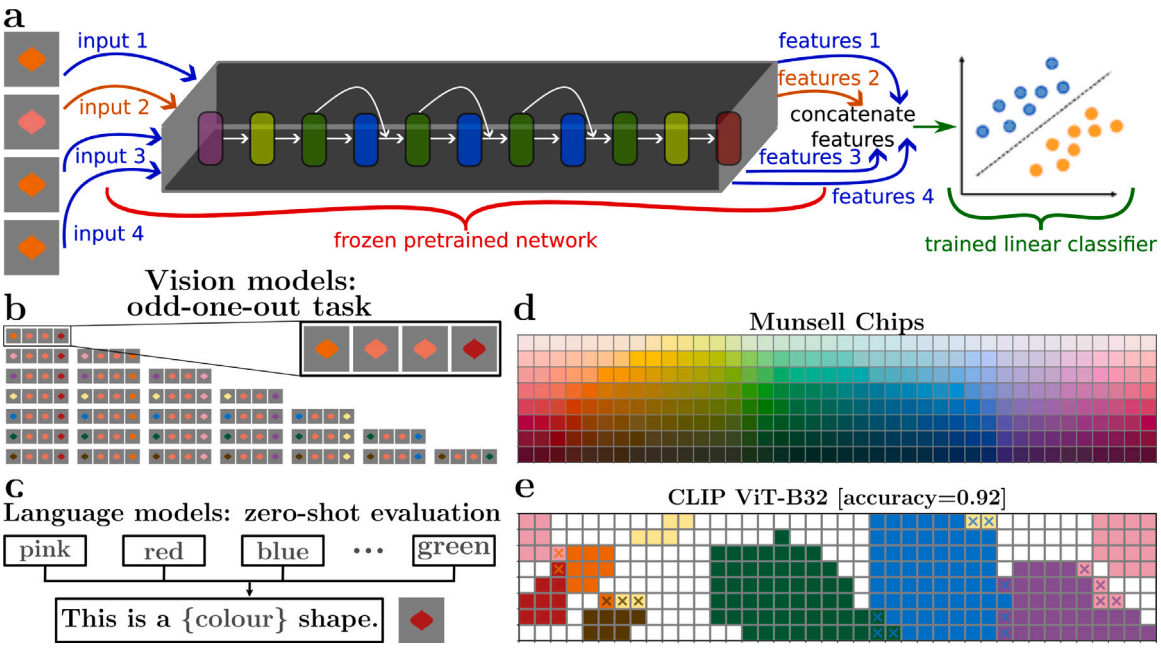


Fig. 1. The Psychophysical framework for assessing colour categories in artificial neural networks. Panel **a**: a linear classifier trained on features from a frozen pretrained network for a four-part odd-one-out task. Panel **b**: Vision layer assessment using conflicting odd images—test colour presented alongside two focal colours, with the category determined by the non-selected focal colour. This is systematically repeated for all pairs of focal colours to eliminate bias. Panel **c**: Language model colour category testing through zero-shot evaluation. The network is prompted with eight phrases, the category based on the term with the highest probability. Panel **d**: Displaying 320 Munsell chips as test colours. Panel **e**: Comparison of colour categories between one example network and human data (Berlin & Kay, 1969; Sturges & Whitfield, 1995). Filled cells represent network outputs, with mismatches indicated by a cross coloured based on human data. In the human data, white cells are not associated with any of the eight examined colour terms.

2.3. Pretrained networks

We investigated twenty-eight pretrained artificial neural networks coming from three sources:

- **ImageNet**—An object recognition dataset containing 1.5 million images spanning over 1000 categories (Deng et al., 2009). Pretrained networks on ImageNet have been extensively compared to human vision, e.g., (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Khaligh-Razavi & Kriegeskorte, 2014; Storrs, Kietzmann, Walther, Mehrer, & Kriegeskorte, 2021; Yamins et al., 2014). Our investigation focused on two pretrained networks from this dataset, encompassing both contemporary architectural types: ResNet50 – a convolutional network (He, Zhang, Ren, & Sun, 2016) and ViT-B32 (Dosovitskiy et al., 2021) – a transformer network.
- **CLIP**—A multimodal framework designed to align images and captions (Radford et al., 2021). Contrastive Language-Image Pretraining (CLIP) networks comprise two encoders (text and image) that are simultaneously optimised to predict correct pairings of image-text batches. In our investigation, to align with the architectures of the chosen ImageNet pretrained networks, we specifically examined CLIP ResNet50 and CLIP ViT-B32, two pretrained networks released by OpenAI. CLIP features have been reported to account for some characteristics of the human visual system (Akbarinia, Morgenstern, & Gegenfurtner, 2023; Geirhos et al., 2021).
- **Taskonomy**—A multitask framework offering an excellent opportunity to compare visual features learnt from different visual tasks (Zamir et al., 2018). This dataset includes 24 pretrained networks, all designed with a common encoder architecture (ResNet50) and trained on an identical set of images (approximately 4 million images, predominantly depicting indoor scenes). Any differences observed between these 24 pretrained networks are attributable to the optimisation of their weights with respect to specific visual tasks, such as edge detection, depth estimation, and scene segmentation, etc. These task-specific ANNs have been previously used to study functions of the visual cortex (Akbarinia et al., 2023; Dwivedi, Bonner, Cichy, & Roig, 2021).

To summarise, we investigated twenty-eight pretrained networks: two with a ViT-B-32 architecture and twenty-six with a ResNet50 architecture. The weights of all aforementioned pretrained networks are sourced from their original PyTorch implementation and are kept frozen throughout the entirety of our experiments.

2.4. Language models: zero-shot evaluation

We conducted psychophysical experiments directly with the language layer of the CLIP ResNet50 and CLIP ViT-B-32 using zero-shot evaluation, eliminating the need for any intermediate steps (see panel **c** in Fig. 1):

- For each foreground shape, we applied a Munsell chip colour and presented the network with eight phrases corresponding to eight colour terms.
- We used the template "This is a {colour} shape"., where "{colour}" is one of the following: *red, orange, yellow, brown, green, blue, purple, pink*.
- The network outputted a probability for each phrase, indicating the likelihood of the phrase matching the image. The phrase with the highest probability was taken as the network's final output.
- For each Munsell chip, this procedure was repeated for all 2905 shapes, resulting in a total of 23,240 trials (8 × 2905).

2.5. Vision models: odd-one-out linear classifier

Directly querying a neural network trained on a visual task like object recognition about colour categories is not feasible, as the network is specialised for a different task. To address this, we employed the linear probing technique (Alain & Bengio, 2017) to evaluate the categorical representation of colours in unimodal vision networks.

This method enables the execution of psychophysical experiments with artificial neural networks using paradigms similar to those in human studies (Akbarinia et al., 2023). Linear probing also allows for the extraction of features at any layer, thereby facilitating the investigation of intermediate features.

We utilised the osculari Python package (Akbarinia, 2023) to implement a four-part odd-one-out colour discrimination task (see panel **a** in Fig. 1). During training, four images were individually input into a frozen pretrained network (i.e., with unaltered weights). The extracted features were then concatenated into a single vector and fed into a linear classifier. This classifier was trained to distinguish the odd image, which was identical to the other three in all aspects except for its foreground colour. To eliminate colour bias in the linear classifier, foreground colours were randomly selected from a uniform RGB distribution, while the background was uniformly chosen from achromatic colours (i.e., $R = G = B$). The training procedure for the linear classifier was consistent across all pretrained networks: a single GPU with a batch size of 32 for ten epochs of 15K samples (i.e., 150K iterations), using the stochastic gradient descent (SGD) optimiser with a learning rate of 0.1 and the categorical cross-entropy loss.

For each architecture, we assessed colour categories at six distinct layers, comprising five intermediate layers and the final layer. In ResNet50, the intermediate layers are defined as areas 0 to 4, while in ViT-B32, they correspond to blocks 1, 4, 7, 10, and 11. Although we endeavoured to align the intermediate layers across architectures by selecting layers at similar depths, it is important to note that an exact match is unattainable due to the inherent differences in their architectures.

It is well known that initial weights prior to training lead to substantial differences in the representations learnt by a neural network (Mehrer, Spoerer, Kriegeskorte, & Kietzmann, 2020). To address this issue and ensure our findings reflect the readout features from frozen pretrained networks rather than the trained linear classifier, we repeated the process of training the colour-discriminator linear classifier five times for each set of readout features from the pretrained networks. The resulting colour categories from these five instances exhibit remarkable consistency (refer to the almost imperceptible standard deviations in panel **a** of Fig. 2). This observation strongly implies that the colour categories assigned by artificial networks are predominantly shaped by features acquired during their pretraining phase, with minimal influence from the colour-discriminator linear classifier.

During testing, we assessed the categorical characteristics of pretrained networks by introducing conflicting odd images (see panel **b** in Fig. 1). In this scenario, the background colour is always mid-grey (i.e., $R = G = B = 128$). Two of the four images are identical, featuring the test colour in their foreground, while the other two images display the focal colour of two distinct categories in their foregrounds. The unselected focal colour indicates the colour category of the test colour from the perspective of the network. To mitigate bias associated with our categorical colour perception, this procedure is repeated for all twenty-eight pairs of colour categories ($\frac{8\times7}{2}$). This procedure was repeated for all 2904 shapes, and the positions of focal colours were swapped to ensure unbiased results. In total, 162,624 trials were conducted for each Munsell chip ($2904 \times 2 \times 28$).

3. Results

To investigate the role of language and visual signals in categorical colour perception, we examined two types of networks: unimodal vision and multimodal language-vision models.

3.1. Role of language

To study the role of language on categorical colour perception, we analysed four pretrained networks resulting from a combination of two tasks, multimodal text-image matching (Radford et al., 2021) and unimodal object recognition (Deng et al., 2009), and two architectures: Vision Transformer—ViT-B32 (Dosovitskiy et al., 2021) and Convolutional Network—ResNet50 (He et al., 2016). We examined the networks at six different layers to elucidate the role of low-, mid-, and high-level visual representation in explaining categorical colour perception. Fig. 2 illustrates the accuracy of predicting human data, measured by assigning the same colour category for each Munsell chip. Our findings reveal that unimodal vision models can explain up to 76% of human data. In contrast, multimodal language-vision models achieve higher accuracy, reaching up to 95% with their language component and notably 89% even without the language component when exclusively testing the vision layers. These results underscore the dual role that language plays in categorical colour perception: a significant portion of human data is explained independently of language, while language-vision models show a 16% improvement in explaining human data, even when tested exclusively with their vision modality (similar to nonverbal psychophysics). Interestingly, testing with the language module (similar to verbal psychophysics) increases accuracy by a moderate 5%, suggesting that language shapes the development of colour categories, but the resulting representation is language-independent.

To contextualise the accuracy of networks, we compared them to (1) the RGB baseline, and (2) state-of-the-art algorithms of colour categorisation (Parraga & Akbarinia, 2020).

1. Given that the input colour space to networks is RGB, we defined a categorical model that computes the Euclidean distance to focal colours, with the smallest distance assigning the category of a Munsell chip. This baseline achieved a high accuracy of 68% in explaining human data, equivalent to the accuracy achieved by ImageNet ResNet50. However, when we applied a threshold to the results for higher confidence, the RGB accuracy substantially dropped to a third, whereas the accuracy of the networks did not change considerably (compare the purple and blue curves in Fig. 2). These results indicate that the input colour space is not the primary determinant of categorical colour perception.
2. The highest reported accuracy in the state-of-the-art for colour categorisation is achieved by *Neural Isoresponsive Colour Ellipsoids* (Parraga & Akbarinia, 2016), with approximately 98% accuracy for the human data examined in this study. *Triple Sigmoid-Eliptic Sigmoid model* (Benavente et al., 2008) achieves the second highest score, around 92%. These two models, along with other state-of-the-art algorithms, are specifically fitted to human data, making their high accuracy unsurprising. In this study, we demonstrate that pretrained networks such as CLIP can achieve a similar level of accuracy without any fitting to human data. This finding suggests that pretrained networks are highly promising models for studying colour categorisation.

Undertaking a qualitative analysis, the right panel of Fig. 2 presents the outcomes of a network prediction on a rainbow image. The arches of the rainbow, sharing identical values in saturation and value, display a continuous increase in hue by one degree. Despite the absence of any physical discontinuity in the rainbow arches, we distinctly perceive them in different colour bands. How do artificial networks interpret this image? In this experiment, we evaluated networks utilising nine colour terms, including the teal/turquoise category, given its qualitative visibility in the rainbow image and widespread usage (Mylonas & MacDonald, 2016). Our observations reveal that the early layer differs significantly from our human colour perception, as it categorises bluish pixels as red and brown. In contrast, the intermediate representation closely mirrors how humans would categorise the rainbow image, except for the purple/pink split, almost entirely classified as purple. The

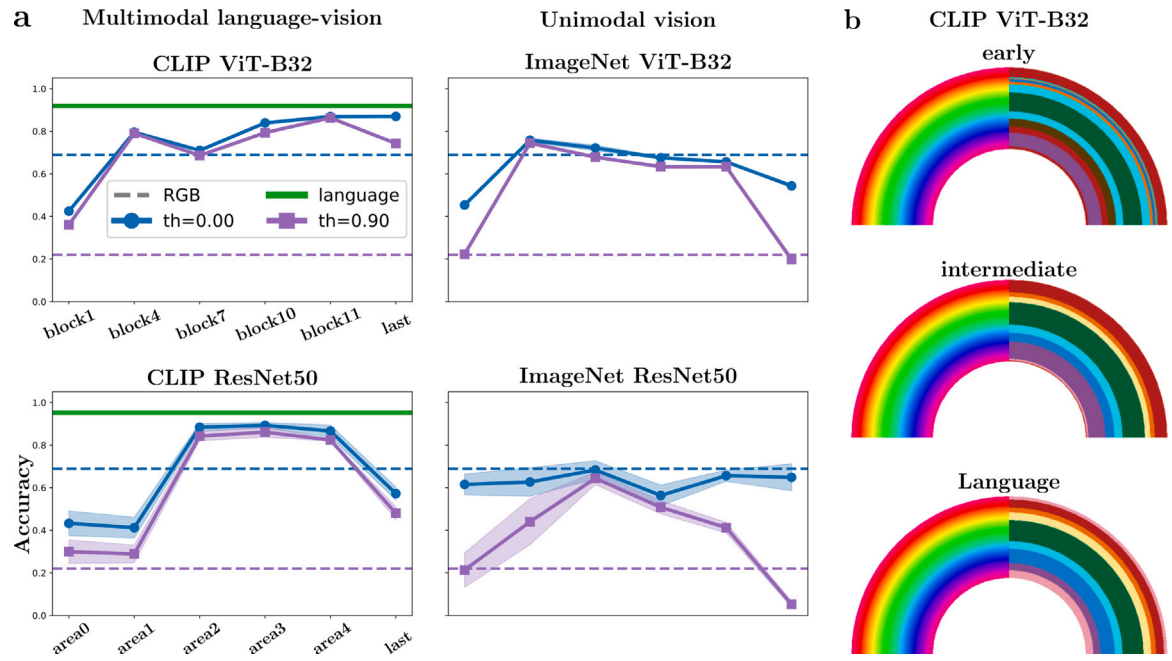


Fig. 2. The influence of language on colour categories. Panel **a**: Shows accuracy in matching human data across six layers of four networks. Blue curves include all results, while purple curves indicate outcomes thresholded at 90% confidence. Transparent regions depict one standard deviation among five instances of linear classifiers trained with the same pretrained network (see Section 2). The green horizontal line marks the accuracy when testing the network with the language module. Dashed horizontal lines represent colour categories based on Euclidean distance in RGB (networks' input colour space). Panel **b**: Displays a rainbow image with continuous hue arches on the left. On the right, colour categories are obtained from an example network at three different layers.

language layer resolves this discrepancy by adjusting the purple and pink categories, perfectly aligning with our perception of the rainbow image.

3.2. Effect of visual task

The Taskonomy dataset (Zamir et al., 2018) consists of twenty-four pretrained networks with an identical encoder architecture (ResNet50), trained on the same set of images for various visual tasks, spanning from low-level edge detection to mid-level depth estimation and high-level object classification. This dataset offers a unique opportunity to investigate the impact of a network's functional role (the visual task a network is optimised towards) on its categorical colour representation. Employing the same analysis as detailed earlier, we scrutinised the networks at six different layers.

A significant disparity is evident among networks in predicting human data—assigning the same colour category for each Munsell chip (see the left panel of Fig. 3). The networks are ranked based on their peak accuracy across six layers, highlighting a substantial gap between the best-performing network, achieving 82% accuracy, and the least-performing one, attaining 16% accuracy. On one end of the spectrum, networks optimised for high-level semantic tasks, like "Object Classification", consistently demonstrate human-like categorical representations. Conversely, networks performing 2D visual tasks, such as "2D Edge Detection", consistently fall short of achieving human-like colour categories. Their predictive capability essentially hovers around chance levels across all layers, markedly lower than the baseline (Euclidean distance in RGB, the network's input colour space). This implies that categorical colour representation is not a beneficial representation for networks trained on 2D visual tasks.

The taxonomy we adopted to classify these networks into four groups (2D, 3D, geometric, and semantic) relies on established criteria from prior literature, including methods such as representational similarity analysis (RSA) (Dwivedi & Roig, 2019) and feature transfer learning (Zamir et al., 2018). Remarkably, our analysis yields similar clusters: along the spectrum of explaining human data, 2D tasks are situated on the left, 3D tasks in the middle, and semantic tasks on the right. This distinction holds true even for equivalent perceptual tasks in different dimensions. For example, the network trained on "3D Edge Detection" achieves human-like colour categories, whereas its corresponding 2D networks do not (as observed in the right panel of Fig. 3). This pattern extends to other corresponding 2D/3D tasks, such as keypoint detection. Collectively, these findings suggest that the nature of the visual tasks a system is designed to perform strongly influences its representation of colour categories. It can be hypothesised that our categorical colour perception has evolved due to living in a three-dimensional space and tackling semantic tasks.

3.3. Internal representation

The comparison of networks/layers to human data has revealed a distinct division. Some networks/layers closely approximate human colour categories, while others fail to align with them. This raises the question of whether there is a fundamental difference in how these two groups of layers/networks represent colours. It is important to note that networks' colour categories are determined through a winner-take-all operation on an eight-class distribution. This is essentially a discrete procedure where one colour wins the category while the rest are silenced. However, before the discretisation stage, the underlying representation is a continuous distribution of the winning ratio among pairs of colour categories, which is a matrix of size 8×8 (refer to Fig. B.1 in the supplementary material). To compare the internal representations of colour categories in networks/layers, we calculated the average Spearman correlation coefficients on this eight-class confusion matrix for each Munsell chip.

The left panel in Fig. 4 presents a pairwise comparison of all probed layers in CLIP and ImageNet networks. Notably, the continuous representation (upper triangle) exhibits better agreement across networks/layers compared to the discrete categories (lower triangle). The average correlation in categorical distributions (continuous) across all layers of CLIP/ImageNet ViT-B32/ResNet50 networks is 0.63 ± 0.12. In contrast, the percentage of matching colour categories (discrete)

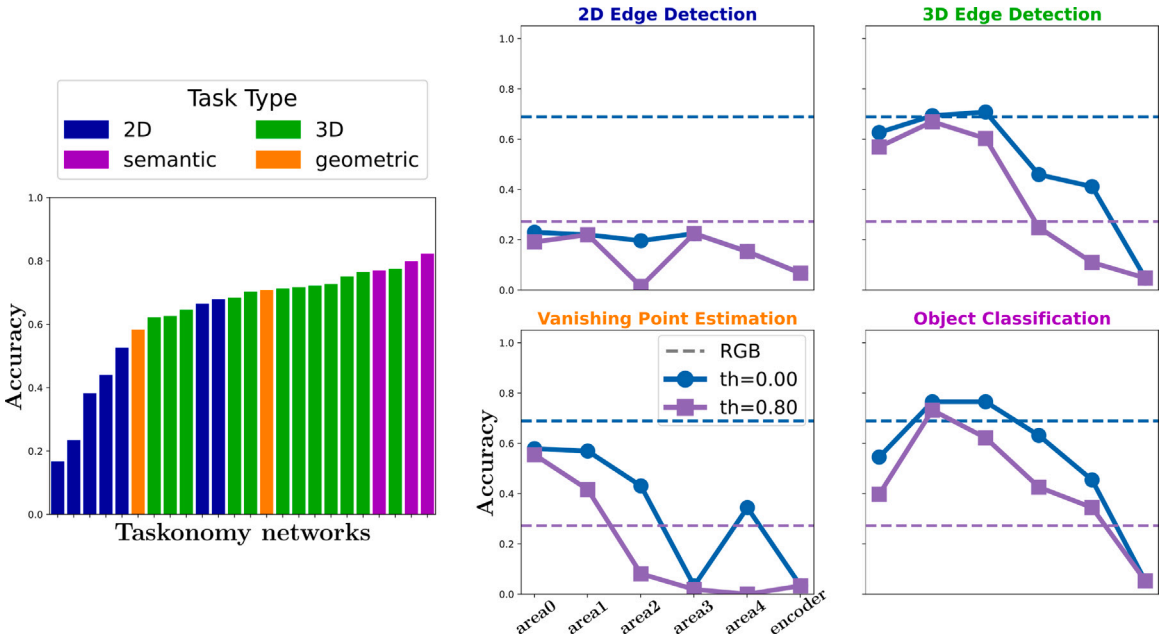


Fig. 3. Effect of Visual task on colour categories. **Left**: Ranks Taskonomy networks by their peak accuracy in explaining human colour categories. **Right**: shows accuracy in matching human data across six layers of four Taskonomy networks (see Fig. C.1 for all twenty-four networks). Blue curves include all results, while purple curves indicate outcomes thresholded at 80% confidence. Dashed horizontal lines depict colour categories based on Euclidean distance in RGB (networks' input colour space). Networks names are colour-coded by task types (Dwivedi et al., 2021): 2D, geometric, 3D, and semantic.

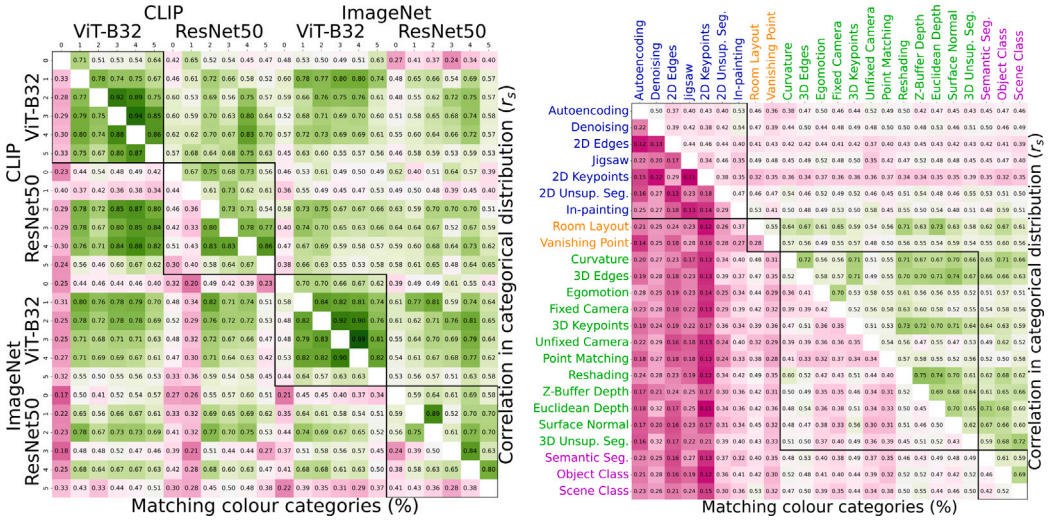


Fig. 4. Comparison of continuous and discrete representations. On the **left**, the upper triangular cells present Spearman correlations in the categorical distribution between pairs of layers, while the lower triangle indicates the percentage of matching colour categories. The dark-bordered squares represent layers within a single network. Cells are colour-coded: green indicates high correlation and high matching rate; purple indicates low correlation and low matching rate. On the **right**, the same format is applied to Taskonomy networks. The values in each cell are averages across the corresponding six layers in the networks. Network names are colour-coded based on task types (Dwivedi et al., 2021): 2D, geometric, 3D, and semantic. The dark-bordered squares delineate networks within a specific task type.

shows both a lower average and higher standard deviation (0.54 ± 0.18), indicating that the underlying continuous representations are considerably more similar than the discrete colour categories. This high correlation in the underlying continuous representation is particularly evident within the layers of a single network (depicted by dark-bordered squares; $r_s = 0.72 \pm 0.04$), and it is notably pronounced in ViT networks.

The right panel in Fig. 4 illustrates a parallel analysis conducted for the Taskonomy networks. The presented comparisons between networks are averaged over layerwise values (i.e., six layers). The first notable observation is the low ratio of matching colour categories across all 24 Taskonomy networks (purple cells in the lower triangle). This observation is not surprising, given the substantial variation in the ability of the Taskonomoy networks to explain human data (see Fig. 3 in the main text and Fig. C.1 in the supplementary material). A second noteworthy pattern is the moderate green cells in the upper triangle, indicating a decent correlation ($r_s = 0.65$) in the categorical distribution of most visual tasks, except the networks trained on 2D tasks (blue labels). This suggests that although the winner colour categories for these networks are notably different, the underlying representation is not significantly dissimilar.

We further examined the continuous representations of the networks in relation to human consistency data from colour naming experiments involving British and German adults (Witzel, Flack, Sanchez-Walker, & Franklin, 2021). The language layers in CLIP networks demonstrated a high correlation coefficient ($r_s = 0.65$), closely aligning

with the correlation observed between British and German speakers ($r_s = 0.67$). The vision layers in multimodal language-vision networks (CLIP) exhibited a similar correlation coefficient (maximum $r_s = 0.63$), whereas unimodal vision networks (ImageNet) showed a more moderate correlation ($r_s = 0.35$). These findings indicate that pretrained networks exhibit a considerable degree of similarity to human categorical colour perception, even at the level of continuous representation.

4. Discussion

Communication plays an integral role in categorical colour perception, evident in our frequent use of colour names, even during inner speech. Recent studies, supporting the universalists' standpoint, propose that efficient communication underlies the formation of colour categories (Chaabouni et al., 2021; Twomey, Roberts, Brainard, & Plotkin, 2021; Zaslavsky et al., 2018). This concept is also seen in the animal kingdom, where colour categories are intertwined with nonverbal communication needs like sexual mating (Caves et al., 2018). Similarly, the role of communication cannot be fully discarded in nonverbal human experiments indicating the emergence of colour categories independent of language, such as those involving stroke patients (Siuda-Krzywicka et al., 2019) and prelinguistic infants (Skelton et al., 2017). The interaction between language and vision is inseparable, due to the nature of our brain throughout its development. In the realm of artificial agents, the inherent co-development of language and vision can be excluded, enabling the testing of models devoid of language and communication components. This advantage has been leveraged in artificial neural networks for object recognition, revealing that human-like colour categories emerge to a considerable extent based solely on their utility for a particular vision task (de Vries et al., 2022). Our results advance this understanding by quantifying the contributions of each essential component—visual signals and linguistic factors. Notably, we find that a significant portion (about 80%) of human colour categories emerge in unimodal vision models. Nevertheless, a small yet important portion (about 20%) remains unexplainable purely on the basis of visual signals, which is clarified by the inclusion of multimodal language-vision models, underscoring the intricate interplay of these components in the development of categorical colour perception.

The utility of colour naming in communication is evident, as it is unfeasible to reference every discriminable tristimulus value with a unique colour name (Lindsey & Brown, 2021). Hence, using distinct colour names for a broader range of hues proves efficient. However, the direct relevance of colour categories to a visual system is less apparent in the absence of communication or language interactions. To explore this, we examined Taskonomy networks, encompassing twenty-four distinct functional roles (i.e., visual tasks defining the optimisation loss) using an identical neural circuitry (i.e., ResNet50 encoder architecture) and training environment (i.e., exposed to the same set of images). The results resonate with the idea that the primary function of colour is to provide information relevant to behavioural tasks in the natural environment (Conway, 2018) by revealing the task-dependent nature of colour categories (Koida & Komatsu, 2007; Webster & Kay, 2012) in a dualistic manner. While human-like categorical colour representation does not emerge in networks trained on 2D tasks, it is not scarce in other functional roles. This challenges the proposition of a unique connection between object recognition and colour categorisation (de Vries et al., 2022). Indeed, our findings suggest that, besides semantic tasks, 3D tasks such as shade parametrisation, depth estimation, and 3D edge detection yield human-like colour categories. The exact benefits of colour categories for specific functional roles, as opposed to others, remain to be investigated. However, categorical colour representation might be associated with foreground-background segmentation (Gibson et al., 2017), a fundamental task continuously performed by infants in their daily lives, potentially explaining the early development of categorical colour perception in prelinguistic infants (Maule, Skelton, & Franklin, 2023).

The first and second stages of colour processing, involving cone activation to different wavelengths of light and the antagonistic combination into colour opponency, are well-established (Gegenfurtner & Kiper, 2003) and reported to manifest in artificial neural networks (Akbarinia & Gil-Rodríguez, 2021; Rafegas & Vanrell, 2018). While these low-level mechanisms account for colour discrimination thresholds, they prove insufficient in explaining colour categories (Roberson, Hanley, & Pak, 2009; Witzel & Gegenfurtner, 2013) and their robustness across illuminant changes (Morimoto, Yamauchi, & Uchikawa, 2023). Our experiments affirm this limitation; irrespective of the network's architecture, modality, or training dataset, the initial layer does not exhibit any categorical effect. It has been postulated that, given the inadequacy of low-level mechanisms in elucidating colour categories, higher-level cognitive processes influenced by linguistic terms mediate categorical colour perception (Roberson et al., 2009). Our results challenge this notion by demonstrating that beneath different colour categories, a similar continuous colour representation may exist. This observation is independent of language modulation and consistently emerges in unimodal vision models. The involvement of high-level visual processes in categorical colour encoding remains uncertain (Bird, Berens, Horner, & Franklin, 2014; Conway et al., 2010; Witzel & Gegenfurtner, 2018). However, our findings do not support this perspective in artificial networks, as the peak accuracy in matching human colour categories is never observed in the final layer. Conceptually, this aligns with the idea that high-level concepts should not strongly associate their representation with colour categories (e.g., recognising an apple based on its shape rather than its colour), and low-level processes should favour generic features in a continuous colour representation (e.g., detecting edges based on fine details of pixel values rather than coarser colour categories).

The connection between continuous colour perception and discrete colour categories remains a major challenge in the field of colour science (Siuda-Krzywicka, Boros, Bartolomeo & Witzel, 2019; Witzel, 2019). Our analysis of twenty-eight artificial neural networks indicates a greater degree of similarity in colour representation at the continuous level compared to discrete categories. This phenomenon may also be present in humans, where individuals across different languages and cultures, despite diverging slightly in their colour terms due to the cognitive influence of language, show a higher degree of alignment in their colour perception. For instance, in several languages, including Russian (Paramei, 2005), Italian (Paggetti, Menegaz, & Paramei, 2016), and Japanese (Kuriki et al., 2017), there are two terms covering the blue category of English; nevertheless, in colour discrimination, their patterns of results are not markedly different (Martinovic, Paramei, & MacInnes, 2020; Roberson et al., 2009). To better disentangle the perceptual and cognitive components of colour categorisation, we posit that a meticulous analysis of intermediate layers in artificial networks can offer valuable insights into this intricate topic. In our experiments, Taskonomy networks (ResNet50 architecture) consistently show categorical colour representation emerging early in area 1, with peak accuracy sustained at mid-level representation (usually areas 1–2), followed by a rapid decline in deeper layers. Similar patterns are observed in ImageNet and CLIP networks (across both ResNet50 and ViT-B32 architectures). However, language models experience a more moderate drop in deeper layers, likely attributed to language modulation interacting directly with the final visual layer. These findings suggest that categorical colour representation is a mid-level feature in artificial neural networks, loosely aligning with the observation in rhesus monkeys that mechanisms for encoding colour categorically should occur earlier than visual area V4 (Walsh, Kulikowski, Butler, & Carden, 1992). The investigation into why mid-level mechanisms favour a categorical colour representation remains a subject for future exploration, yet insights from artificial neural networks propose that they may hold the key to advancing our understanding of categorical colour perception (Peirce, 2015).

5. Conclusion

In this article, we demonstrate that artificial neural networks are a suitable tool for investigating colour perception through hypothesis-driven experiments (Bowers et al., 2023). We examined the roles of vision and language in categorical colour perception by quantifying the contribution of each modality. Our results indicate that a significant portion of human data can be explained by vision models that lack any language component. Furthermore, we show that not all vision models achieve human-like colour categories, and such representations are highly dependent on the network's training task. High-level semantic tasks, such as object recognition, are the primary drivers of categorical colour representation, reinforcing the view that the primary function of colour is to provide information relevant to behavioural tasks in the natural environment (Conway, 2018). The implications of our findings for future research are twofold:

- **Theoretical**: This study systematically investigated two aspects of categorical colour representation: the role of language and visual tasks. An important aspect that remains unexplored is the network's training dataset. The networks we examined were pre-trained on datasets of natural images, which closely align with the statistics of the natural environment. By systematically manipulating the global and local statistics of the pretrained dataset, one can investigate the influence of the environment on categorical colour representation. This approach would provide a clearer understanding of the utility of categorical colour representation to a visual system (Geisler, 2008).
- **Applied**: Current state-of-the-art colour categorisation algorithms operate at the pixel level, meaning that an arbitrary RGB value will always correspond to the same colour category, regardless of its context and surroundings. Our findings demonstrate that pretrained artificial neural networks inherently achieve state-of-the-art accuracy. Given these networks' sophisticated representation of complex scenes, they are ideally suited for developing a dynamic colour categorisation model that can account for local and global context, which is known to significantly impact human colour perception (Gegenfurtner & Kiper, 2003).

CRediT authorship contribution statement

Arash Akbarinia: Writing – review & editing, Writing – original draft, Visualization, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Arash Akbarinia reports financial support was provided by Justus Liebig University Giessen. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft SFB/TRR 135 (grant number 222641018) TP S. Portions of this work were presented in abstract form at the European Conference on Visual Perception in 2022 and 2023 (Akbarinia, 2022). We extend our gratitude to Christoph Witzel for providing the human colour naming consistency data.

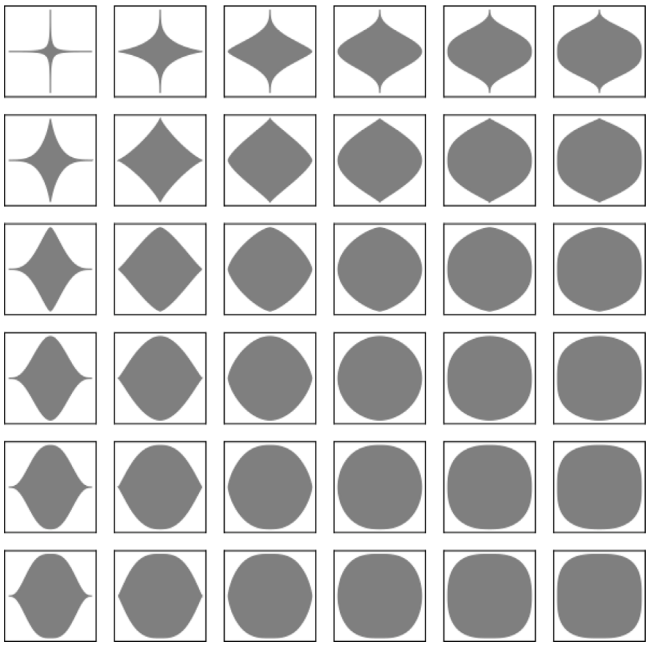


Fig. A.1. Example of thirty-six superellipse shapes obtained by keeping $a = b = 0.5$ and systematically varying the m and n values in Eq. (1).

Appendix A. Stimuli shapes

To create the test shapes in our study, we employed the superellipse, defined in the Cartesian coordinate system as the set of all points (x, y) satisfying the equation

$$\left|\frac{x}{a}\right|^m + \left|\frac{y}{b}\right|^n = 1, \tag{1}$$

where a, b, m and n are positive numbers. Fig. A.1 depicts thirty-six examples of these superellipse shapes. The selection of a systematic geometrical shape serves the purpose of exploring the interaction between object shape and colour perception, although this aspect falls outside the scope of the current article.

Appendix B. Raw experimental data

The exhaustive examinations conducted to evaluate the categorical representation of colours in vision layers through linear probing yield an 8×8 multi-class confusion matrix, as illustrated in Fig. B.1. Several noteworthy aspects of this matrix warrant attention:

- Higher values indicate a robust category effect, while values close to 0.5 (chance level) suggest an absence of categorical representation.
- The summation of winning ratios for a specific pair of colours may not necessarily equate to 1.0. For example, in Fig. B.1, the sum of winning ratios for the orange-red colour categories is 0.99. The remaining percentage pertains to scenarios where the test colour has been selected as the odd image. This can be construed as noise in the linear classifier. Overall, the magnitude of this noise is minimal, accounting for only 0.02 across all layers.
- The relationship between colour categories is not entirely transitive. In Fig. B.1, although orange prevails over red 78% of the time, when compared to brown and purple categories, red obtains a marginally higher winning ratio (1% more, i.e., 100 versus 99). Whether this discrepancy is attributable to noise in the linear classifier or signifies the non-transitive nature of colour categories remains unclear. Nevertheless, similar to the aforementioned point, the impact is exceedingly marginal.

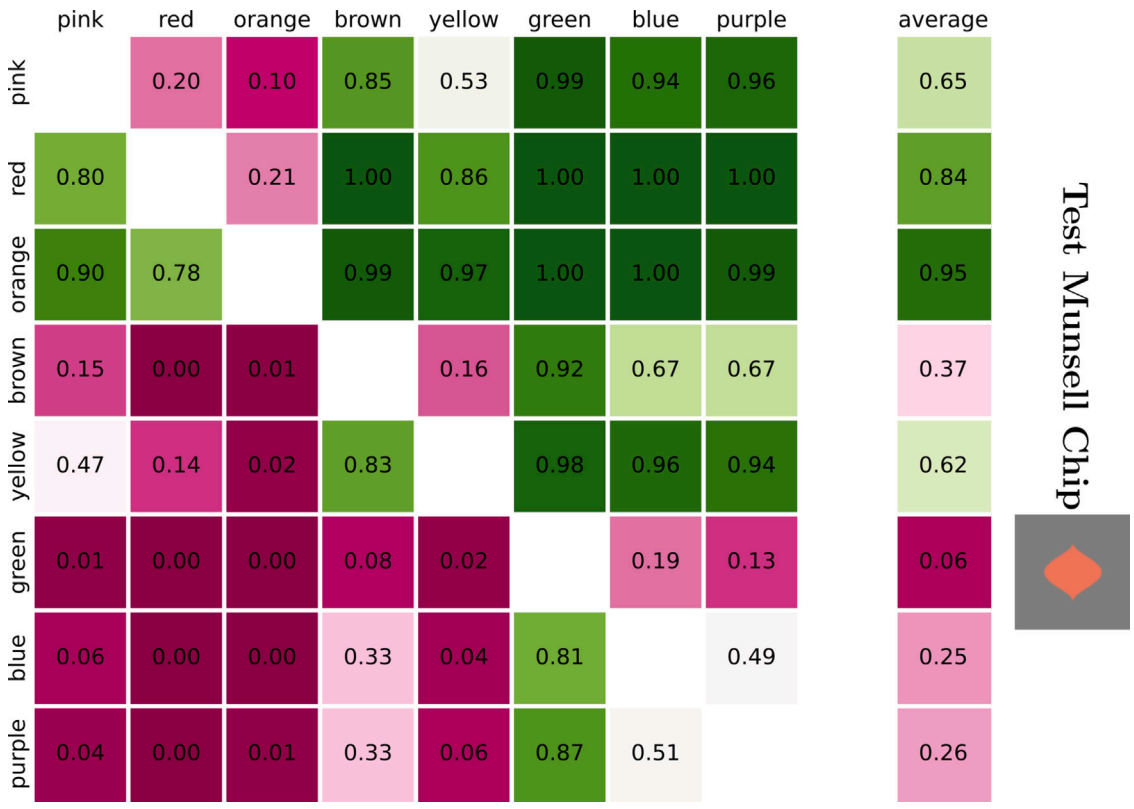


Fig. B.1. Distribution of winning colour categories (Derived from Block-10 of CLIP Vit-B32). Each cell denotes the percentage of a category selected as the colour for the illustrated test-Munsell chip. The values in the upper and lower triangles may not necessarily add up to 1; the remaining percentage (typically minimal) indicates instances where neither category is chosen. The numbers reflect the average across 5810 tests. Cells are colour-coded, with green representing 1 and purple representing 0.

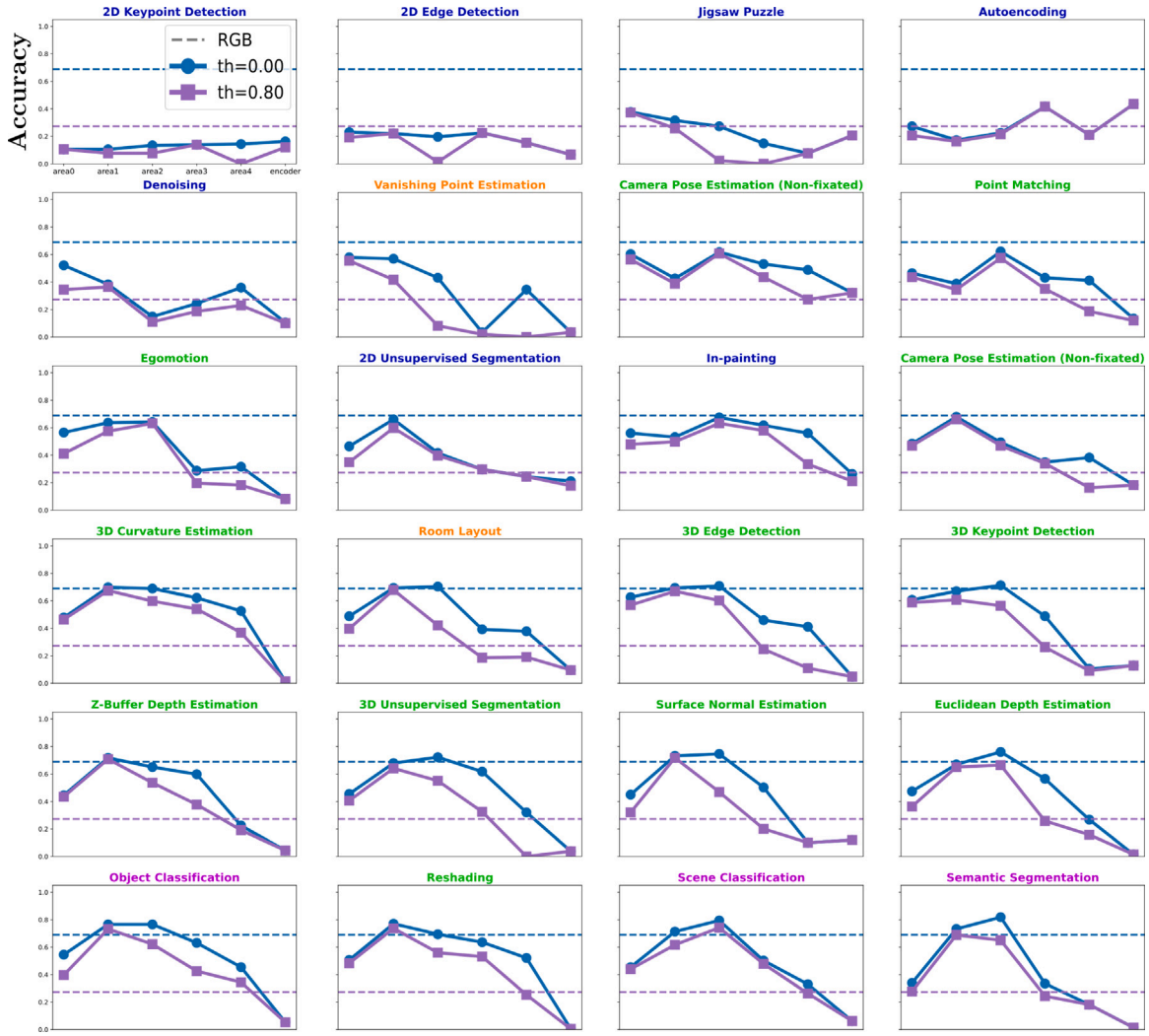


Fig. C.1. Results of Taskonomy networks. Accuracy matching human data in six layers of 24 Taskonomy networks. Blue curves include all results, while purple curves indicate outcomes thresholded at 80% confidence. Dashed horizontal lines depict colour categories based on Euclidean distance in RGB (networks' input colour space). Networks names are colour-coded by task types (Dwivedi et al., 2021): 2D, geometric, 3D, and semantic.

Appendix C. Taskonomy results

Fig. C.1 illustrates the accuracy in matching with human colour categories for all twenty-four Taskonomy networks across six layers. The networks are arranged in ascending order based on their peak accuracy in explaining human data. Notably, in the top two rows, all networks grouped under the 2D task type (Dwivedi et al., 2021) demonstrate inferior performance compared to the RGB baseline. This observation implies that categorical colour representation is inconsequential to their functional role.

References

Akbarinia, A. (2022). Interaction between colour and form in vision transformers. *Perception*, *51*, 186. http://dx.doi.org/10.1177/03010066221141167.

Akbarinia, A. (2023). Osculari: a Python package to explore artificial neural networks with psychophysical experiments. http://dx.doi.org/10.5281/zenodo.10420544.

Akbarinia, A., & Gil-Rodríguez, R. (2021). Color conversion in deep autoencoders. *Journal of Perceptual Imaging*, *29*(1), 89. http://dx.doi.org/10.2352/J.Percept.Imaging.2021.4.2.020401.

Akbarinia, A., Morgenstern, Y., & Gegenfurtner, K. R. (2023). Contrast sensitivity function in deep networks. *Neural Networks*, *164*, 228–244.

Alain, G., & Bengio, Y. (2017). Understanding intermediate layers using linear classifier probes. In *International conference on learning representations*.

Benavente, R., Vanrell, M., & Baldrich, R. (2008). Parametric fuzzy sets for automatic color naming. *Journal of the Optical Society of America A*, *25*(10), 2582–2593.

Berlin, B., & Kay, P. (1969). Basic color terms: Their universality and evolution. Univ of California Press.

Bird, C. M., Berens, S. C., Horner, A. J., & Franklin, A. (2014). Categorical encoding of color in the brain. *Proceedings of the National Academy of Sciences*, *111*(12), 4590–4595.

Bornstein, M. H., & Korda, N. O. (1984). Discrimination and matching within and between hues measured by reaction times: Some implications for categorical perception and levels of information processing. *Psychological Research*, *46*(3), 207–222.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., et al. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, Article e385.

Caves, E. M., Green, P. A., Zipple, M. N., Peters, S., Johnsen, S., & Nowicki, S. (2018). Categorical perception of colour signals in a songbird. *Nature*, *560*(7718), 365–367.

Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2021). Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of Sciences*, *118*(12), Article e2016569118.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 1–13.

Conway, B. R. (2018). The organization and operation of inferior temporal cortex. *Annual Review of Vision Science*, *4*, 381–402.

Conway, B. R., Chatterjee, S., Field, G. D., Horwitz, G. D., Johnson, E. N., Koida, K., et al. (2010). Advances in color science: from retina to behavior. *Journal of Neuroscience*, *30*(45), 14955–14963.

Cropper, S. J., Kvansakul, J. G., & Little, D. R. (2013). The categorisation of non-categorical colours: a novel paradigm in colour perception. *PLoS One*, *8*(3), Article e59945.

Davidoff, J. (2001). Language and perceptual categorisation. *Trends in Cognitive Sciences*, *5*(9), 382–387.

de Vries, J. P., Akbarinia, A., Flachot, A., & Gegenfurtner, K. R. (2022). Emergent color categorization in a neural network trained for object recognition. *Elife*, *11*, Article e76472.

van de Weijer, J., Schmid, C., Verbeek, J., & Larlus, D. (2009). Learning color names for real-world applications. *IEEE Transactions on Image Processing*, *18*(7), 1512–1523.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Derefeldt, G., & Swartling, T. (1995). Colour concept retrieval by free colour naming. Identification of up to 30 colours without training. *Displays*, *16*(2), 69–77.

Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of Physiology*, *357*(1), 241–265.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International conference on machine learning*.

Dwivedi, K., Bonner, M. F., Cichy, R. M., & Roig, G. (2021). Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS Computational Biology*, *17*(8), Article e1009267.

Dwivedi, K., & Roig, G. (2019). Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 12387–12396).

Fagot, J., Goldstein, J., Davidoff, J., & Pickering, A. (2006). Cross-species differences in color categorization. *Psychonomic Bulletin & Review*, *13*(2), 275–280.

Franklin, A., Pilling, M., & Davies, I. (2005). The nature of infant color categorization: Evidence from eye movements on a target detection task. *Journal of Experimental Child Psychology*, *91*(3), 227–248.

Gegenfurtner, K. R., & Kiper, D. C. (2003). Color vision. *Annual Review of Neuroscience*, *26*(1), 181–206.

Gegenfurtner, K. R., & Sharpe, L. A. (1999). Color vision. Cambridge, UK: Cambridge University Press.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., et al. (2021). Partial success in closing the gap between human and machine vision. *Vol. 34*, In *Proceedings of the conference on neural information processing systems* (pp. 23885–23899).

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, *59*(1), 167–192.

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., et al. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, *114*(40), 10785–10790.

Harnad, S. (2003). Categorical perception. Nature Publishing Group: Macmillan.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Indow, T. (1988). Multidimensional studies of Munsell color solid. *Psychological Review*, *95*(4), 456.

Indow, T., & Kanazawa, K. (1960). Multidimensional mapping of Munsell colors varying in hue, chroma, and value. *Journal of Experimental Psychology*, *59*(5), 330.

Kay, P., Berlin, B., Maffi, L., Merrifield, W. R., & Cook, R. (2009). The world color survey. Citeseer.

Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist*, *86*(1), 65–79.

Kay, P., & Regier, T. (2006). Language, thought and color: recent developments. *Trends in Cognitive Sciences*, *10*(2), 51–54.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), Article e1003915.

Koida, K., & Komatsu, H. (2007). Effects of task demands on the responses of color-selective neurons in the inferior temporal cortex. *Nature Neuroscience*, *10*(1), 108–116.

Kuriki, I., Lange, R., Muto, Y., Brown, A. M., Fukuda, K., Tokunaga, R., et al. (2017). The modern Japanese color lexicon. *Journal of Vision*, *17*(3), 1.

Lindsey, D. T., & Brown, A. M. (2021). Lexical color categories. *Annual Review of Vision Science*, *7*, 605–631.

Linhares, J. M., Pinto, P. D., & Nascimento, S. M. (2008). The number of discernible colors in natural scenes. *Journal of the Optical Society of America A*, *25*(12), 2918–2924.

Martinovic, J., Paramei, G. V., & MacInnes, W. J. (2020). Russian blues reveal the limits of language influencing colour discrimination. *Cognition*, *201*, Article 104281.

Maule, J., Skelton, A. E., & Franklin, A. (2023). The development of color perception and cognition. *Annual Review of Psychology*, *74*, 87–111.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*(1), 5725.

Mojsilovic, A. (2005). A computational model for color naming and describing color composition of images. *IEEE Transactions on Image processing*, *14*(5), 690–699.

Morimoto, T., Yamauchi, Y., & Uchikawa, K. (2023). Invariant categorical color regions across illuminant change coincide with focal colors. *Journal of Vision*, *23*(2), 7.

Mylonas, D., & MacDonald, L. (2016). Augmenting basic colour terms in English. *Color Research & Application*, *41*(1), 32–42.

Paggetti, G., Menegaz, G., & Paramei, G. V. (2016). Color naming in Italian language. *Color Research & Application*, *41*(4), 402–415.

Paramei, G. V. (2005). Singing the Russian blues: An argument for culturally basic color terms. *Cross-Cultural Research*, *39*(1), 10–38.

Parraga, C. A., & Akbarinia, A. (2016). NICE: A computational solution to close the gap from colour perception to colour categorization. *PLoS One*, *11*(3), Article e0149538.

Parraga, C. A., & Akbarinia, A. (2020). Color name applications in computer vision. *Encyclopedia of Color Science and Technology*, 1–7.

Peirce, J. W. (2015). Understanding mid-level representations in visual processing. *Journal of Vision*, *15*(7), 5.

Pointer, M., & Attridge, G. (1998). The number of discernible colours. *Color Research & Application: Endorsed by Inter-Society Color Council*, *23*(1), 52–54, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, the Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Rafegas, I., & Vanrell, M. (2018). Color encoding in biologically-inspired convolutional neural networks. *Vision Research*, *151*, 7–17.

Regier, T., Kay, P., & Cook, R. S. (2005). Focal colors are universal after all. *Proceedings of the National Academy of Sciences*, *102*(23), 8386–8391.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*(4), 1436–1441.

Roberson, D., Davidoff, J., & Braisby, N. (1999). Similarity and categorisation: Neuropsychological evidence for a dissociation in explicit categorisation tasks. *Cognition*, *71*(1), 1–42.

Roberson, D., Davidoff, J., Davies, I. R., & Shapiro, L. R. (2004). The development of color categories in two languages: a longitudinal study. *Journal of Experimental Psychology: General*, *133*(4), 554.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, *129*(3), 369.

Roberson, D., Hanley, J. R., & Pak, H. (2009). Thresholds for color discrimination in English and Korean speakers. *Cognition*, *112*(3), 482–487.

Seaborn, M., Hepplewhite, L., & Stonham, J. (2005). Fuzzy colour category map for the measurement of colour similarity and dissimilarity. *Pattern Recognition*, *38*(2), 165–177.

Siuda-Krzywicka, K., Boros, M., Bartolomeo, P., & Witzel, C. (2019). The biological bases of colour categorisation: From goldfish to the human brain. *Cortex*, *118*, 82–106.

Siuda-Krzywicka, K., Witzel, C., Chabani, E., Taga, M., Coste, C., Cools, N., et al. (2019). Color categorization independent of color naming. *Cell Reports*, *28*(10), 2471–2479.

Skelton, A. E., Catchpole, G., Abbott, J. T., Bosten, J. M., & Franklin, A. (2017). Biological origins of color categorization. *Proceedings of the National Academy of Sciences*, *114*(21), 5545–5550.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, *33*(10), 2044–2064.

Sturges, J., & Whitfield, T. A. (1995). Locating basic colours in the Munsell space. *Color Research & Application*, *20*(6), 364–376.

Twomey, C. R., Roberts, G., Brainard, D. H., & Plotkin, J. B. (2021). What we talk about when we talk about colors. *Proceedings of the National Academy of Sciences*, *118*(39), Article e2109237118.

Walsh, V., Kulikowski, J., Butler, S., & Carden, D. (1992). The effects of lesions of area V4 on the visual abilities of macaques: colour categorization. *Behavioural Brain Research*, *52*(1), 81–89.

Webster, M. A., & Kay, P. (2012). Color categories and color appearance. *Cognition*, *122*(3), 375–392.

Witzel, C. (2019). Misconceptions about colour categories. *Review of Philosophy and Psychology*, *10*(3), 499–540.

Witzel, C., Flack, Z., Sanchez-Walker, E., & Franklin, A. (2021). Colour category constancy and the development of colour naming. *Vision Research*, *187*, 41–54.

Witzel, C., & Gegenfurtner, K. R. (2013). Categorical sensitivity to color differences. *Journal of Vision*, *13*(7), 1.

Witzel, C., & Gegenfurtner, K. R. (2018). Color perception: Objects, constancy, and categories. *Annual Review of Vision Science*, *4*, 475–499.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., & Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3712–3722).

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.