

Category-Selective Neurons in Deep Networks: Comparing Purely Visual and Visual-Language Models

Zitong Lu (lu.2367@osu.edu)

The Ohio State University

Yuxin Wang

University of Cincinnati

Abstract

Category-selective regions in the human brain, such as the fusiform face area (FFA), extrastriate body area (EBA), parahippocampal place area (PPA), and visual word form area (VWFA), play a crucial role in high-level visual processing. Here, we investigate whether artificial neural networks (ANNs) exhibit similar category-selective neurons and how these neurons vary across model layers and between purely visual and vision-language models. Inspired by fMRI functional localizer experiments, we presented images from different categories (faces, bodies, scenes, words, scrambled scenes, and scrambled words) to deep networks and identified category-selective neurons using statistical criteria. Comparing ResNet and the structurally controlled ResNet-based CLIP model, we found that both models contain category-selective neurons, with their proportion increasing across layers, mirroring category selectivity in higher-level visual brain regions. However, CLIP exhibited a higher proportion but lower specificity of category-selective neurons compared to ResNet. Additionally, CLIP's category-selective neurons were more evenly distributed across feature maps and demonstrated greater representational consistency across layers. These findings suggest that language learning increases the number of category-selective neurons while reducing their selectivity strength, reshaping visual representations in deep networks. Our study provides insights into how ANNs mirror biological vision and how multimodal learning influences category-selective representations.

Keywords: category-selectivity; functional localization; comparative neuroscience and AI; multimodal learning

Introduction

Category-selective regions in the human visual system, such as the Fusiform Face Area (FFA) (Kanwisher et al. (1997); Kanwisher & Yovel (2006)), Extrastriate Body Area (EBA) (Downing et al. (2001)), Parahippocampal Place Area (PPA) (Epstein & Kanwisher (1998)), and Visual Word Form Area (VWFA) (Dehaene & Cohen (2011); McCandliss et al. (2003)), play a crucial role in high-level visual processing. These areas exhibit hierarchical category selectivity, suggesting that categorical representations emerge progressively along the visual pathway (Grill-Spector & Weiner (2014); Haxby et al. (2001)). In recent years, artificial neural networks (ANNs) have been widely used to model human visual processing (Cadieu et al.

(2014); Cichy et al. (2016); Kubilius et al. (2019); Yamins et al. (2014); Lu & Ku (2023)). Studies have shown that ANNs can develop features resembling those in the biological visual system, such as low-level edge detectors and high-level semantic representations (Yamins & DiCarlo (2016); Margalit et al. (2024); Lu & Golomb (2023b); Huang et al. (2021)). However, it remains unclear whether ANNs naturally develop category-selective neurons and whether their selectivity patterns align with those observed in the human brain.

Recent research suggests that ANNs not only approximate human visual behavior but also exhibit neural representations akin to those in the brain (Cichy et al. (2016); Güçlü & van Gerven (2015); Kietzmann et al. (2019); Lu & Golomb (2023a); Lu et al. (2023); Yamins et al. (2014); Yamins & DiCarlo (2016)). This has led to a “reverse engineering” approach, where neuroscientists analyze ANN representations to infer principles of human vision. Understanding ANN representations can thus inform both AI research and neuroscience. While previous work has focused on purely visual models such as ResNet (He et al. (2016)), vision-language models like CLIP have demonstrated more human-like conceptual representations (Radford et al. (2021)). Since CLIP’s visual encoder shares the same architecture as ResNet but incorporates language supervision, its category representations may differ from those in purely visual models. This raises a critical question: Does language learning influence the formation of category-selective neurons? Does it increase their number or alter their selectivity patterns? Although language is known to play a role in concept learning (F. Xu (2002); Condry & Spelke (2008); Carruthers (2002)), its impact on visual representations remains unclear.

To address these questions, we adopted a “functional localizer” approach, commonly used in fMRI research to identify category-selective regions in the brain (Abassi & Papeo (2024); Li et al. (2024); Poldrack (2007)). Using the same method, we examined category selectivity in ResNet and CLIP by presenting images from different categories (faces, bodies, scenes, words, scrambled scenes, and scrambled words) and identifying category-selective neurons using statistical criteria. In addition, we analyzed their distribution, hierarchical organization, and cross-layer consistency. Our results show that category-selective neurons emerge across multiple layers of ANNs, increasing in proportion at deeper levels. Moreover, CLIP exhibits a higher proportion of category-selective neurons than ResNet, but they are less selective, more uniformly distributed, and exhibit greater representational consistency

across layers. These findings suggest that language learning increases the number of category-selective neurons while reducing their selectivity strength, shedding light on how multimodal learning reshapes visual representations.

The key contributions of our current study are as follows: (1) First systematic study of category-selective neurons in ANNs across hierarchical layers; (2) Reveals the influence of language on category selectivity in neural representations; (3) Provides new insights into ANNs as models of human vision and multimodal learning.

General Methods

In this section, we describe the stimuli used for our functional localizer experiments in ANNs, the selection of models and the rationale behind it, the method for identifying category selective neurons, and the metrics used to quantify selectivity. Additional methodological details are provided in the corresponding result section.

Stimuli

We selected 40 images each from four categories: face, body, scene, and word, with some images sourced from the fLoc functional localizer package (Stigliani et al. (2015)) (Figure 1A). However, when investigating category-selective neurons, using only these four categories could raise concerns regarding whether the selectivity is truly driven by high-level semantic information or merely by low-level visual features such as edges, textures, and spatial frequency.

To control for low-level feature-driven selectivity, we additionally included 40 scrambled scene images and 40 scrambled word images as control conditions (Figure 1A). This manipulation ensures that any observed category selectivity is more likely to reflect high-level semantic processing rather than simple visual properties. Our stimulus selection and design follow rigorous protocols from fMRI studies, ensuring that our analysis of category-selective neurons in ANNs aligns with established methodologies used in human neuroscience research.

Model selection

To investigate category-selective neurons in deep neural networks, we compared a purely visual model (ResNet-50) with a vision-language model (ResNet-50-based CLIP). ResNet-50 is a widely used purely visual model trained on ImageNet (Deng et al. (2009)), while ResNet-50-based CLIP has the same convolutional backbone but is trained using a contrastive learning objective with paired image-text data. We chose these two models for the following reasons: (1) CNN-based architectures better align with human vision: Prior research suggests that convolutional neural networks (CNNs) are more biologically plausible than vision transformers (ViTs). CNNs exhibit a hierarchical organization similar to the human visual cortex, have higher neural and behavioral similarity to humans (Geirhos et al. (2021)), and demonstrate human-like adversarial vulnerability (Bhojanapalli et al. (2021)). Given these advantages, we chose ResNet over ViTs. (2) Structural

control between models: To compare purely visual models with vision-language models, we controlled for architecture by selecting ResNet-50-based CLIP, which maintains the same ResNet-50 backbone. This ensures that any observed differences between the two models are primarily due to language supervision rather than architectural discrepancies.

Additionally, we focused our analysis on all layers following ReLU activations in ResNet-50, totaling 17 layers. This choice was motivated by two key considerations: First, ReLU introduces sparsity by zeroing out negative activations, making feature representations more biologically plausible. Second, nonlinear feature transformations primarily occur after ReLU, meaning category selectivity is more meaningfully assessed at these points.

"Functional localizer" in ANNs

Similar to functional localizer tasks in fMRI, we presented the 240 images to both ResNet and CLIP, recording the activation of every neuron in response to different images from different categories. A neuron was identified as category-selective if its activation was significantly stronger for one category compared to all others. Specifically, for each neuron, we conducted independent t-tests comparing activations between images from one category and each of the other categories. For instance, to be classified as a face-selective neuron, a neuron needed to show: Face>Body and Face>Scene and Face>Word and Face>Scrambled Scene and Face>Scrambled Word, with $p < 0.05$ for each comparison. This allowed us to quantify the proportion of category-selective neurons at each layer and measure the activation strength of these neurons across different visual stimuli.

Category selectivity index (CSI)

Since activation magnitudes vary across layers, we computed a Category Selectivity Index (CSI) to quantify category selectivity in a layer-independent manner:

$$CSI = \frac{R_{\text{preferred}} - R_{\text{non-preferred}}}{R_{\text{preferred}} + R_{\text{non-preferred}}} \times 100\% \quad (1)$$

where $R_{\text{preferred}}$ is the mean activation for images of the neuron's preferred category, and $R_{\text{non-preferred}}$ is the mean activation across all other categories. A higher CSI indicates stronger category selectivity. A lower CSI suggests weaker selectivity or broad tuning across categories. By comparing CSI across layers and models, we could assess how category selectivity evolves through hierarchical processing and how language supervision reshapes category representations.

Cross-layer consistency analysis

To evaluate the representational consistency of category-selective representations across layers, we performed a cross-layer representational similarity analysis (Kriegeskorte et al. (2008); Lu & Ku (2020)). For each model (ResNet and CLIP), we constructed a 40×40 representational dissimilarity matrix (RDM) for each layer and each category using the Pearson correlation distance. Specifically, given a category

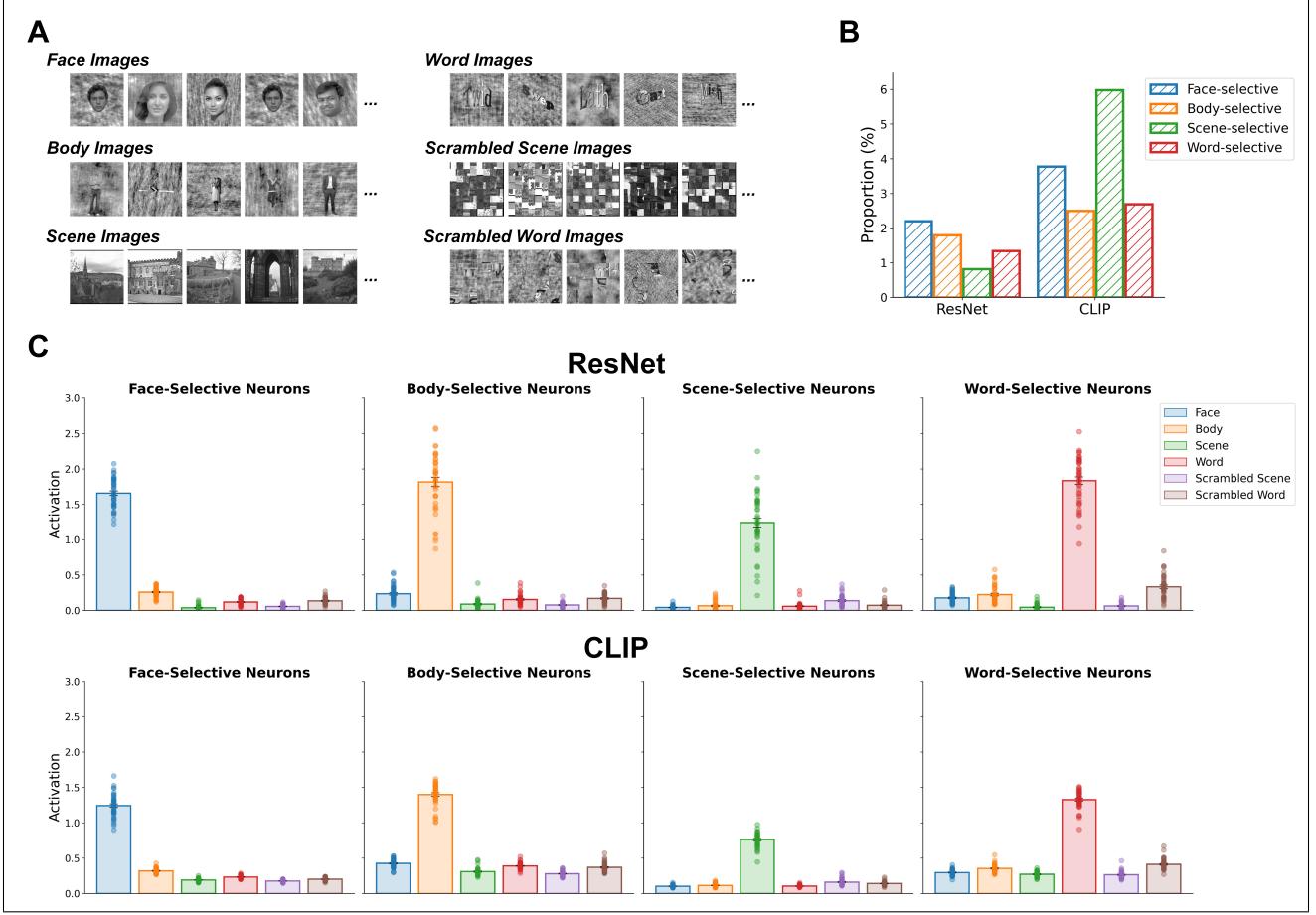


Figure 1: Category-selective neurons in layer4.2.relu from ResNet and CLIP. (A) Image stimuli used in the functional localizer experiment. Images were divided into six categories: face, body, scene, word, scrambled scene, and scrambled word. (B) Proportion of category-selective neurons in the last convolutional layer (layer4.2.relu) of ResNet and CLIP. Neurons were classified as category-selective if they exhibited significantly higher activation for one category compared to all others. (C) Response profiles of category-selective neurons from layer4.2.relu. The top row shows results for ResNet, and the bottom row for CLIP. Each individual dot corresponds to a image.

(e.g., face), we obtained the activations of all face-selective neurons in a given layer for the 40 images from that category. The pairwise dissimilarity between any two images i, j was computed as:

$$\text{RDM}_{i,j}^{(L,c)} = 1 - \text{PCorr}(\mathbf{a}_i^{(L,c)}, \mathbf{a}_j^{(L,c)}) \quad (2)$$

where \mathbf{a}_i and \mathbf{a}_j are the activation vectors of all face-selective neurons in the layer for images i and j , and PCorr is the Pearson correlation coefficient. We then computed the similarity between two different layer's RDMs using the Spearman correlation coefficient:

$$\text{Cross-layer similarity}_{(Li,Lj)} = \text{SCorr}(\text{vec}(\text{RDM}^{(Li,c)}), \text{vec}(\text{RDM}^{(Lj,c)})) \quad (3)$$

where $\text{RDM}^{(Li,c)}$ and $\text{RDM}^{(Lj,c)}$ are the RDMs from different layers in the same model and SCorr is the Spearman correlation coefficient.

Results

Category-selective neurons emerge in ANNs

To begin our analysis, we focused on a single layer within the models. Specifically, we examined the final convolutional layer after ReLU activation in both ResNet and CLIP's visual module (i.e., 'layer4.2.relu' in the ResNet-50 architecture). By statistically analyzing all neurons in this layer, we found that both models exhibited category-selective neurons for faces, bodies, scenes, and words. Figure 1B presents the proportion of category-selective neurons (number of selective neurons relative to the total number of neurons in this layer) in both models, while Figure 1C shows the activation strengths of these category-selective neurons for different image categories.

In ResNet's 'layer4.2.relu', we observed the highest proportion of face-selective neurons (2.20%), followed by body-selective neurons (1.80%), word-selective neurons (1.34%), and scene-selective neurons (0.81%). However, CLIP's

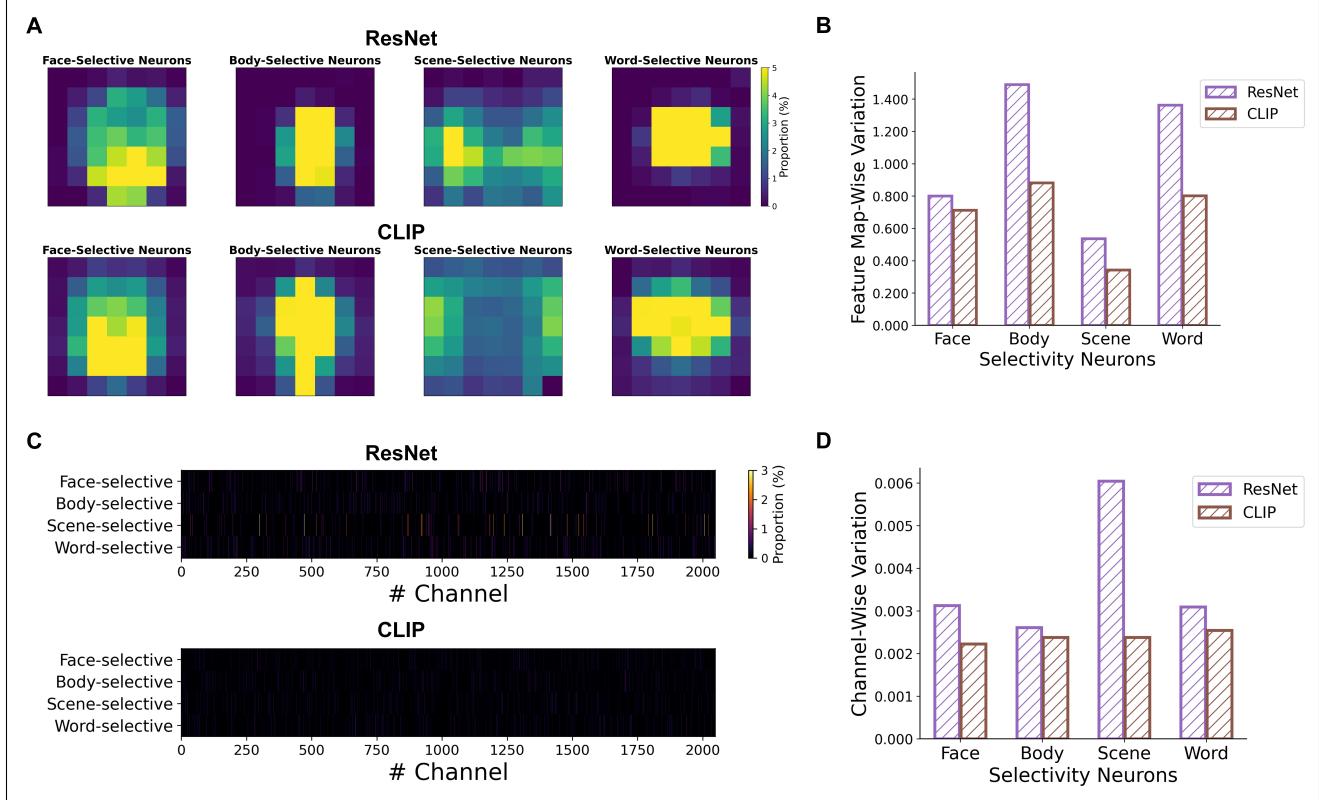


Figure 2: Spatial and channel-wise distribution of category-selective neurons in layer4.2.relu from ResNet and CLIP. (A) Feature map-wise distribution of category-selective neurons. Each heatmap represents the spatial distribution of face-, body-, scene-, and word-selective neurons across the feature map in the last convolutional layer (layer4.2.relu) for ResNet (top raw) and CLIP (bottom row). The color scale indicates the proportion of neurons at each spatial location in the feature map. (B) Quantification of feature map-wise variation in category-selective neuron distribution. (C) Channel-wise distribution of category-selective neurons. Each row represents the proportion of neurons selective to a given category across all channels in layer4.2.relu for ResNet (top) and CLIP (bottom). (D) Quantification of variation in the channel-wise distribution of category-selective neurons.

'layer.4.2.relu' exhibited significantly higher proportions of category-selective neurons across all categories—with 3.78% for faces, 2.50% for bodies, 5.98% for scenes, and 2.69% for words. This suggests that language learning enhances category selectivity in neural representations.

To further investigate the distribution of these selective neurons and how they differ between models, we analyzed their spatial arrangement within the feature maps of the convolutional layer. We considered two perspectives: feature map position and channel-wise organization.

For the feature map analysis, we would like to ask whether category-selective neurons are spatially localized. We aggregated all channels in this layer to visualize the cumulative occurrence ratio of category-selective neurons at different spatial positions in the feature map, allowing us to compare the distribution patterns between ResNet and CLIP. Figure 2A shows the proportion of different category-selective neurons at each spatial location in the feature map (i.e., the number of selective neurons appearing at a given location divided by the total number of channels in the layer). Both mod-

els exhibit location-specific preferences for different category-selective neurons, rather than a uniform distribution. To quantify whether category-selective neurons in CLIP are more evenly distributed across feature maps compared to ResNet, we calculated the variance of their spatial distribution. If a model's selective neurons are more evenly spread, different locations should exhibit similar proportions of selective neurons, resulting in a lower variance. Our analysis (Figure 2B) confirms that ResNet exhibits a higher variance, indicating that its category-selective neurons are more spatially localized, whereas CLIP's neurons are more evenly distributed across the feature map. This suggests that ResNet favors local feature selectivity, while language learning in CLIP promotes a more global representation, potentially facilitating a more holistic understanding of category information.

Similarly, to further examine how category-selective neurons are distributed across channels, we aggregated activations across all feature map locations to compute the cumulative occurrence ratio in each channel (Figure 2C). This allowed us to assess whether ResNet and CLIP exhibit differ-

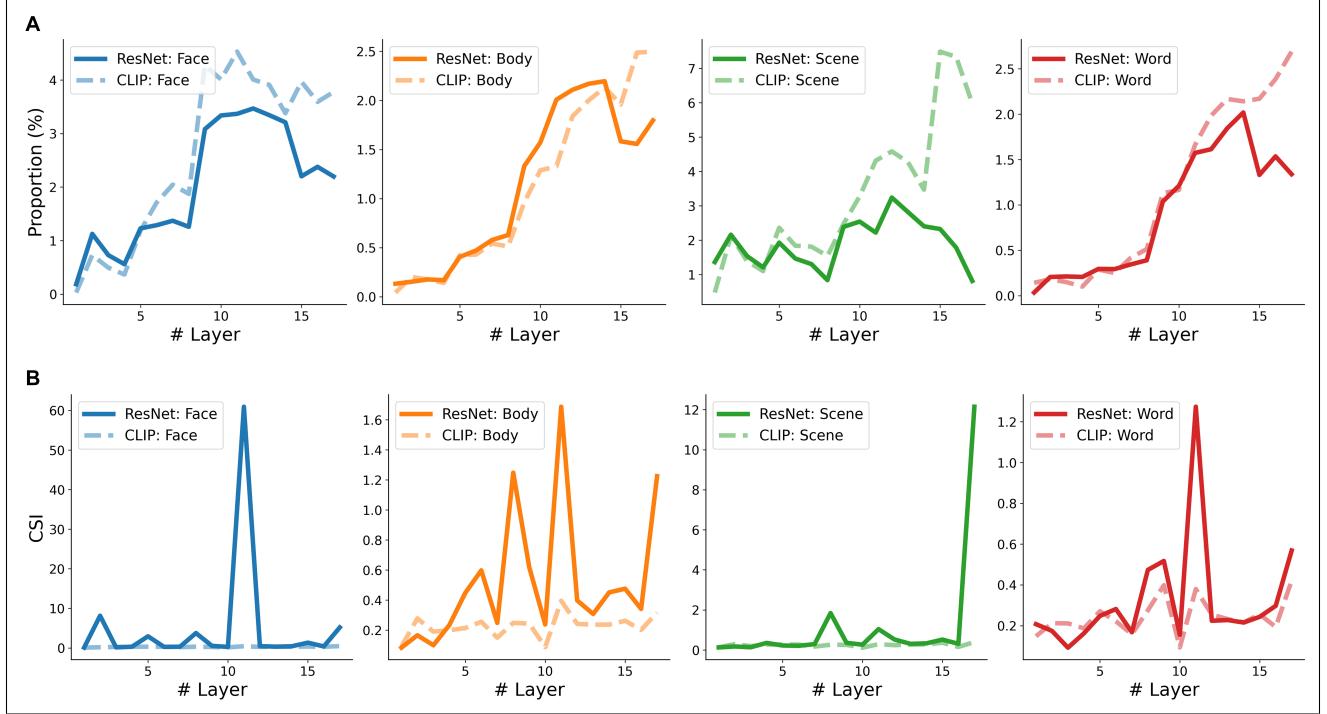


Figure 3: Layer-wise analysis of category-selective neurons in ResNet and CLIP. (A) Proportion of category-selective neurons across layers. (B) Category selectivity index (CSI) across layers. From left to right, each panel shows the proportion of neurons selective for faces (blue), bodies (orange), scene (green), and words (red) at each layer of ResNet (solid lines) and CLIP (dashed lines).

ent patterns of channel-wise category selectivity. Our results reveal a key distinction: In ResNet, scene-selective neurons tend to be highly concentrated in a small subset of channels, with some channels showing a markedly higher proportion of scene-selective neurons than others. In contrast, face-, body-, and word-selective neurons in ResNet, as well as all category-selective neurons in CLIP, exhibit a more dispersed distribution across channels, without clear clustering in specific channels. Also, we quantified whether category-selective neurons are more distributed across channels by calculating the variance of their channel-wise distribution. It shows similar pattern to what we found in feature map-wise variation (Figure 2D). This suggests that while scene representations in ResNet may rely on a small number of highly selective channels, other category-selective neurons—and particularly those in CLIP—are more distributed across channels. The broader distribution in CLIP further supports the hypothesis that language learning leads to a more evenly spread representation of category information within deep networks.

Layer-wise analysis of category-selective neurons

To extend our findings beyond a single layer, we systematically analyzed the layer-wise progression of category-selective neurons throughout the network (Figure 3A). Across all four categories, ResNet exhibited a pattern where the proportion of category-selective neurons initially increased with depth be-

fore declining in later layers. This decline was particularly pronounced for scene-selective neurons, which sharply decreased in the final layers. In contrast, CLIP not only exhibited a consistently higher proportion of category-selective neurons than ResNet (except for body-selective neurons in some intermediate layers) but also maintained a higher proportion of category-selective neurons in later layers.

In addition to neuron proportions, we examined the category selectivity index (CSI) across layers (Figure 3B). Unlike the neuron count results, ResNet exhibited substantially higher CSI values than CLIP in many layers. This indicates that although CLIP has more category-selective neurons overall, ResNet's category-selective neurons exhibit stronger specificity for their preferred category. Notably, face-selective neurons in ResNet showed exceptionally high CSI values, whereas body- and word-selective neurons had much weaker selectivity.

The discrepancy between the higher number of category-selective neurons in CLIP and their lower CSI values suggests a key difference in representation between the two models: In purely visual ResNet, fewer neurons are category-selective, but when they are, their selectivity is highly specific. In visual-language CLIP, a greater number of neurons exhibit category selectivity, but their responses to other categories are also stronger, leading to lower specificity. This suggests that language learning in CLIP distributes category-selective infor-

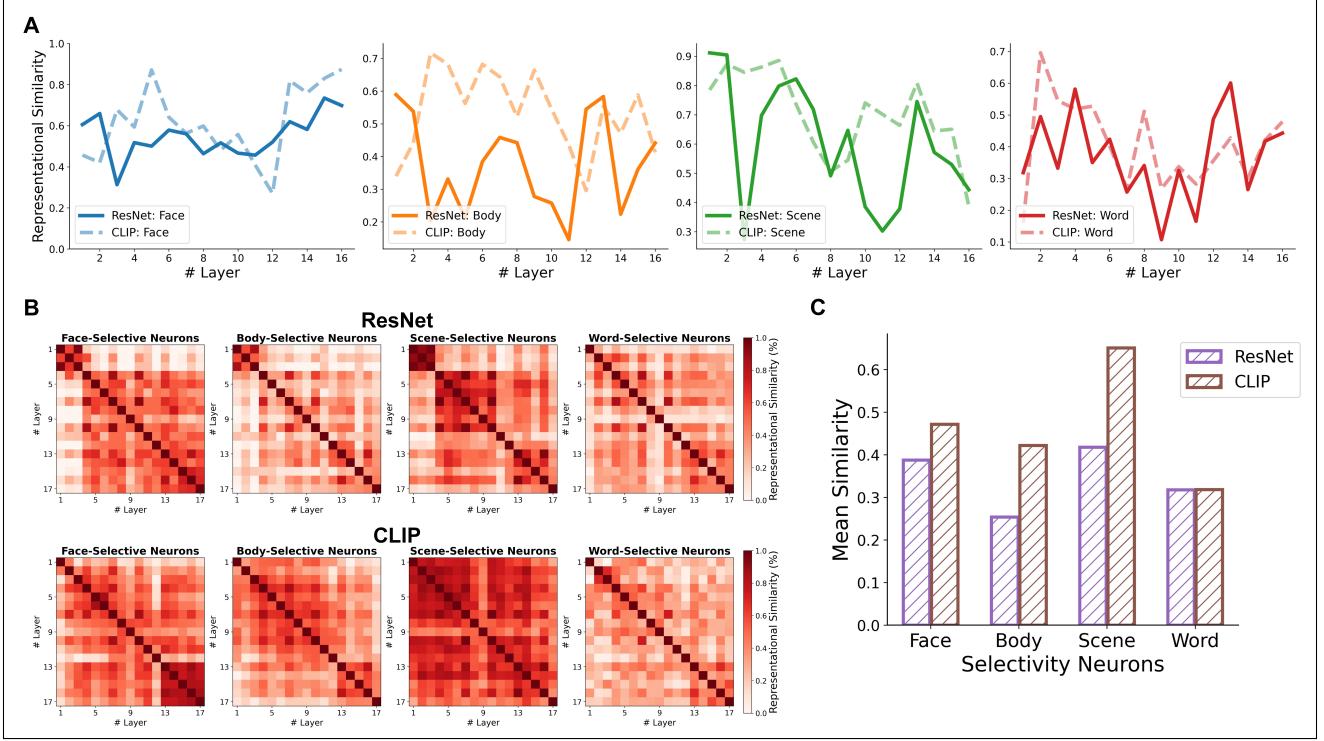


Figure 4: Cross-layer consistency of category-selective neurons in ResNet and CLIP. (A) Representational similarity between adjacent layers. (B) Fully-layer similarity matrices. Each heatmap represents the pairwise representational similarity between layers for category-selective neurons in ResNet (top) and CLIP (bottom). Darker colors indicate higher similarity. (C) Mean cross-layer representational similarity by averaging non-diagonal elements from (B).

mation more broadly across neurons, leading to a more dispersed and less sharply tuned category representation compared to ResNet.

Cross-layer consistency of category-selective representations

Our previous results suggest that CLIP has a greater number of category-selective neurons but with lower category selectivity index (CSI) values, indicating that these neurons may exhibit weaker category specificity compared to those in ResNet. Additionally, we observed that category-selective neurons in CLIP are more evenly distributed across the feature map, suggesting that category information may be encoded in a more spatially dispersed and global manner. If CLIP employs a more global encoding strategy, its category-selective neurons may also exhibit greater consistency across layers, forming a more structured and stable hierarchical representation.

To investigate this, we analyzed the cross-layer similarity of category representations. Specifically, for each model, category, and layer, we computed a 40×40 representational dissimilarity matrix (RDM) based on the model's activation patterns. Each RDM captures the dissimilarity (1 - Pearson correlation coefficient) between the responses of 40 images from the same category. We then examined the correlation between RDMs of adjacent layers to measure how consistently

category-selective representations are preserved across layers. A higher correlation between adjacent RDMs suggests a more stable and coherent category representation across hierarchical levels.

As shown in Figure 4A, CLIP exhibits higher RDM correlations across adjacent layers compared to ResNet, particularly in higher layers. This suggests that category information in CLIP remains more stable and consistent across different layers, whereas in ResNet, category information undergoes greater transformation as it propagates through the network. However, an exception to this trend occurs for word-selective neurons: CLIP does not exhibit a clear advantage over ResNet in cross-layer similarity for word representations. This suggests that, unlike scene-related representations, word processing in CLIP may be less stable across hierarchical levels.

To further confirm this observation, we computed pairwise RDM correlations between all layers within each model (Figure 4B). By averaging the off-diagonal elements of this similarity matrix, we found that CLIP exhibits overall higher representational consistency across layers (Figure 3C), reinforcing the idea that category-selective neurons in CLIP are organized in a more stable and structured manner compared to ResNet.

Discussion

Through a systematic analysis of neuronal activity in ANNs, we demonstrate that ANNs contain category-selective neurons analogous to those found in the human brain, such as face-selective neurons (akin to FFA), body-selective neurons (akin to EBA), scene-selective neurons (akin to PPA), and word-selective neurons (akin to VWFA). This finding suggests that current ANN models exhibit category-selective responses similar to those observed in biological visual systems. Importantly, our approach mirrors human fMRI studies by incorporating scrambled scene and scrambled word images as control conditions to rule out low- and mid-level visual feature-driven selectivity. This ensures that the observed category selectivity genuinely reflects semantic processing rather than basic visual properties. Notably, the ANNs examined in our study were not explicitly trained to recognize the detailed identity of these object categories — ImageNet-trained ResNet, for example, does not specifically learn face or word classification. Yet, category-selective neurons still emerged, particularly for word images, despite the absence of word-related object categories in the training set. This suggests that category-selective representations can emerge in deep networks even without direct category-level supervision.

How does language learning influence category-selective neurons? To investigate the impact of language learning on category-selective neurons, we compared the distribution of these neurons in purely visual models (ResNet) and vision-language models (CLIP). Our findings show that CLIP exhibits a greater number of category-selective neurons but lower category selectivity index (CSI) compared to ResNet. This indicates that neurons in CLIP are more evenly distributed across different categories, whereas ResNet neurons exhibit more extreme selectivity. This result suggests that language learning increases the number of category-selective neurons while reducing their specificity, possibly because language enhances category-level similarities, leading neurons to respond to multiple categories rather than being strictly selective. Further analysis revealed that category-selective neurons in ResNet are more spatially localized within feature maps, whereas those in CLIP are more evenly distributed. The higher variance in feature map localization observed in ResNet suggests that its category-selective neurons depend on specific spatial locations, whereas CLIP, influenced by language supervision, encodes category information more globally. Additionally, category-selective neurons in CLIP exhibit greater consistency across layers, meaning that category-selective representations are more stable across different processing stages. This suggests that language learning promotes hierarchical consistency in category representation, likely by reinforcing high-level semantic structures across multiple layers.

Our results suggest that ResNet relies more on local features for category discrimination, whereas CLIP encodes category information in a more global manner. This distinction is reflected in the spatial distribution of category-selective neu-

rons: In ResNet, category-selective neurons are more localized within specific regions of feature maps, indicating that object category information may be encoded in particular spatial locations (e.g., some positions may be more sensitive to faces). In CLIP, category-selective neurons are more evenly distributed, suggesting that language supervision enhances the global representation of category information, allowing neurons to encode categories more broadly across the network rather than relying on specific local features. This may contribute to CLIP’s superior generalization ability. Our cross-layer analysis further supports this interpretation. We found that face- and body-selective neurons in CLIP exhibit greater stability across layers compared to those in ResNet, likely because these categories have well-defined semantic representations in language (e.g., “face” and “body” are explicit linguistic concepts). Scene-selective neurons showed the largest improvement in cross-layer consistency in CLIP, suggesting that language supervision enhances global scene representation, whereas ResNet may rely more on local texture-based processing. In contrast, word-selective neurons in CLIP did not show significant improvements in cross-layer consistency, which may be due to two factors: (1) word-selective neurons are inherently less frequent in both models, and (2) similar to the human VWFA, word-selective neurons may primarily encode visual word forms rather than their semantic content, leading to weaker hierarchical consistency.

While previous studies have identified face-selective or word-selective neurons in CNNs (Agrawal & Dehaene (2024); Baek et al. (2021); S. Xu et al. (2021)), our study is the first to systematically analyze multiple category-selective neurons (face, body, scene, and word) in structurally controlled models, comparing purely visual (ResNet) and vision-language (CLIP) architectures. Beyond confirming the existence of category-selective neurons, we examined how language supervision influences their formation, revealing its effects on neuron distribution, selectivity strength (CSI), and cross-layer stability. Additionally, we introduced the CSI (Category Selectivity Index) as a quantitative measure, allowing for rigorous comparisons across layers and models rather than relying solely on activation strength. Our study provides novel insights into: (1) The hierarchical emergence of category-selective neurons in ANNs, mirroring the increasing category selectivity observed in biological vision. (2) The role of language learning in expanding category-selective neurons while reducing their specificity, suggesting that language enhances category relationships. (3) The greater spatial uniformity and cross-layer stability of category-selective neurons in CLIP compared to ResNet, highlighting the effect of language in promoting global category representations. These findings contribute to a deeper understanding of ANNs as models of human vision and provide new insights into how multimodal learning reshapes visual representations.

Limitations and future directions

Despite providing new insights into category-selective neurons in ANNs, our study leaves several open questions: First, our study focuses on trained models, rather than examining category-selective neurons in randomly initialized networks. Some studies have suggested using untrained networks as an analogy for the “infant brain,” (Baek et al. (2021); Kim et al. (2021); Zhou et al. (2022)) but we argue that this analogy is flawed since infant brains already possess structured neural connections at birth. Future work should investigate how category selectivity emerges over different learning stages, comparing early-stage and late-stage training dynamics with human visual development. Second, while our study reveals category-selective neurons in ANNs, it does not directly compare them to category-selective regions in the human brain (e.g., FFA, PPA, VWFA). Future research could leverage fMRI or ECoG data to compare category-selective activations in ANN neurons and human cortical regions, further validating the biological plausibility of ANN representations. Third, our study characterizes the distribution of category-selective neurons but does not directly test their computational importance. Future work could employ ablation experiments to selectively remove category-selective neurons and observe their impact on object recognition or scene understanding tasks. Fourth, the ANNs in our study were trained on general object classification tasks (ImageNet for ResNet and contrastive learning for CLIP). Future research could explore how category selectivity changes when models are trained on specialized tasks. For example, does fine-tuning a model on face recognition enhance face-selective neurons while weakening other category selectivity? Understanding task-driven modifications of category selectivity could provide further insights into how different learning objectives shape visual representations. Finally, beyond classic object categories, many other categorical dimensions—such as real-world size, animacy, food-relatedness, and spikiness (Bao et al. (2020); Coggan & Tong (2023); Huang et al. (2022); Jain et al. (2023); Khaligh-Razavi et al. (2018); Khosla et al. (2022); Konkle & Oliva (2012); Konkle & Caramazza (2013); Lu & Golomb (2023b)) —have been shown to be encoded in both the brain and ANNs. Recent studies suggest that models optimized using neural data exhibit stronger food-related representations than those trained solely on images (Lu et al. (2024); Lu & Wang (2024)). Future work should explore how these categorical features are encoded differently in ANNs and the human brain, and how language learning influences their representation.

Conclusion

Our study provides the first systematic analysis of category-selective neurons in ANNs across hierarchical layers, revealing the emergence of category-selective neurons in ANNs and the influence of language learning on category selectivity in neural representations. We show that ANNs naturally develop category-selective neurons, and these neurons become more prevalent in deeper layers. Language learning increases

the number of category-selective neurons but reduces their specificity, while also promoting more uniform spatial distribution and greater hierarchical consistency. These findings offer new insights into ANNs as models of human vision and provide a theoretical foundation for understanding how multimodal learning shapes visual representations.

References

- Abassi, E., & Papeo, L. (2024). Category-Selective Representation of Relationships in the Visual Cortex. *Journal of Neuroscience*, 44(5). doi: 10.1523/JNEUROSCI.0250-23.2023
- Agrawal, A., & Dehaene, S. (2024). Cracking the neural code for word recognition in convolutional neural networks. *PLOS Computational Biology*, 20(9), e1012430. doi: 10.1371/JOURNAL.PCBI.1012430
- Baek, S., Song, M., Jang, J., Kim, G., & Paik, S. B. (2021). Face detection in untrained deep neural networks. *Nature Communications*, 12(1), 1–15. doi: 10.1038/s41467-021-27606-9
- Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, 583(7814), 103–108. doi: 10.1038/s41586-020-2350-5
- Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., & Veit, A. (2021). *Understanding Robustness of Transformers for Image Classification*.
- Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., ... DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*, 10(12), e1003963. doi: 10.1371/JOURNAL.PCBI.1003963
- Carruthers, P. (2002). The cognitive functions of language. *Behavioral and Brain Sciences*, 25(6), 657–674. doi: 10.1017/S0140525X02000122
- 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. doi: 10.1038/srep27755
- Coggan, D. D., & Tong, F. (2023). Spikiness and animacy as potential organizing principles of human ventral visual cortex. *Cerebral Cortex*, 33(13), 8194–8217. doi: 10.1093/cercor/bhad108
- Condry, K. F., & Spelke, E. S. (2008). The Development of Language and Abstract Concepts: The Case of Natural Number. *Journal of Experimental Psychology: General*, 137(1), 22–38. doi: 10.1037/0096-3445.137.1.22.SUPP
- Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15(6), 254–262. doi: 10.1016/J.TICS.2011.04.003
- Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, 248–255. doi: 10.1109/CVPR.2009.5206848
- Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A Cortical Area Selective for Visual Processing of the Human Body. *Science*, 293(5539), 2470–2473. doi: 10.1126/SCIENCE.1063414
- Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676), 598–601. doi: 10.1038/33402
- Geirhos, R., Narayananappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems* 34, 34, 23885–23899.
- Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8), 536–548. doi: 10.1038/nrn3747
- Güçlü, U., & van Gerven, M. A. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27), 10005–10014. doi: 10.1523/JNEUROSCI.5023-14.2015
- Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539), 2425–2430. doi: 10.1126/science.1063736
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.
- Huang, T., Song, Y., & Liu, J. (2022). Real-world size of objects serves as an axis of object space. *Communications Biology*, 5(1), 1–12. doi: 10.1038/s42003-022-03711-3
- Huang, T., Zhen, Z., & Liu, J. (2021). Semantic Relatedness Emerges in Deep Convolutional Neural Networks Designed for Object Recognition. *Frontiers in Computational Neuroscience*, 15, 625804. doi: 10.3389/FNCOM.2021.625804
- Jain, N., Wang, A., Henderson, M. M., Lin, R., Prince, J. S., Tarr, M. J., & Wehbe, L. (2023). Selectivity for food in human ventral visual cortex. *Communications Biology*, 6(1), 1–14. doi: 10.1038/s42003-023-04546-2
- Kanwisher, N., McDermott, J., & Chun, M. M. (1997, jun). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *Journal of Neuroscience*, 17(11), 4302–4311. doi: 10.1523/JNEUROSCI.17-11-04302.1997
- Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1476), 2109–2128. doi: 10.1098/RSTB.2006.1934
- Khaligh-Razavi, S.-M., Cichy, R. M., Pantazis, D., & Oliva, A. (2018). Tracking the Spatiotemporal Neural Dynamics of Real-world Object Size and Animacy in the Human Brain. *Journal of Cognitive Neuroscience*, 30(11), 1559–1576. doi: 10.1162/jocn_a_01290

- Khosla, M., Ratan Murty, N. A., & Kanwisher, N. (2022). A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, 32(19), 4159–4171.e9. doi: 10.1016/J.CUB.2022.08.009
- Kietzmann, T. C., Spoerer, C. J., Sørensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences of the United States of America*, 116(43), 21854–21863. doi: 10.1073/PNAS.1905544116/SUPPL_FILE/PNAS.1905544116.SM06.MP4
- Kim, G., Jang, J., Baek, S., Song, M., & Paik, S. B. (2021). Visual number sense in untrained deep neural networks. *Science Advances*, 7(1). doi: 10.1126/SCIADV.ABD6127
- Konkle, T., & Caramazza, A. (2013). Tripartite Organization of the Ventral Stream by Animacy and Object Size. *Journal of Neuroscience*, 33(25), 10235–10242. doi: 10.1523/JNEUROSCI.0983-13.2013
- Konkle, T., & Oliva, A. (2012). A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron*, 74(6), 1114–1124. doi: 10.1016/J.NEURON.2012.04.036
- Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 4. doi: 10.3389/NEURO.06.004.2008
- Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N. J., ... Dicarlo, J. J. (2019). Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.
- Li, J., Hiersche, K. J., & Saygin, Z. M. (2024). Demystifying visual word form area visual and nonvisual response properties with precision fMRI. *iScience*, 27(12), 111481. doi: 10.1016/J.ISCI.2024.111481
- Lu, Z., Doerig, A., Bosch, V., Krahmer, B., Kaiser, D., Cichy, R. M., & Kietzmann, T. C. (2023). End-to-end topographic networks as models of cortical map formation and human visual behaviour: moving beyond convolutions. *arXiv*. doi: 10.48550/arXiv.2308.09431
- Lu, Z., & Golomb, J. D. (2023a). Generate your neural signals from mine: individual-to-individual EEG converters. *Proceedings of the Annual Meeting of the Cognitive Science Society* 45.
- Lu, Z., & Golomb, J. D. (2023b). Human EEG and artificial neural networks reveal disentangled representations of object real-world size in natural images. *bioRxiv*. doi: 10.1101/2023.04.26.538469
- Lu, Z., & Ku, Y. (2020). NeuroRA: A Python Toolbox of Representational Analysis From Multi-Modal Neural Data. *Frontiers in Neuroinformatics*, 14, 61. doi: 10.3389/FNINF.2020.563669
- Lu, Z., & Ku, Y. (2023). Bridging the gap between EEG and DCNNs reveals a fatigue mechanism of facial repetition suppression. *iScience*, 26, 108501. doi: 10.1016/j.isci.2023.108501
- Lu, Z., & Wang, Y. (2024). Teaching CORnet Human fMRI Representations for Enhanced Model-Brain Alignment. *arXiv*. doi: 10.48550/arXiv.2401.17231
- Lu, Z., Wang, Y., & Golomb, J. D. (2024). Achieving More Human Brain-Like Vision via Human Neural Representational Alignment. *arXiv*. doi: 10.48550/arXiv.2401.17231
- Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., & Yamins, D. L. (2024). A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 112(14), 2435–2451.e7. doi: 10.1016/J.NEURON.2024.04.018
- McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, 7(7), 293–299. doi: 10.1016/S1364-6613(03)00134-7
- Poldrack, R. A. (2007). Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience*, 2(1), 67–70. doi: 10.1093/SCAN/NSM006
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the international conference on machine learning (icml)*.
- Stigliani, A., Weiner, K. S., & Grill-Spector, K. (2015). Temporal Processing Capacity in High-Level Visual Cortex Is Domain Specific. *Journal of Neuroscience*, 35(36), 12412–12424. doi: 10.1523/JNEUROSCI.4822-14.2015
- Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, 85(3), 223–250. doi: 10.1016/S0010-0277(02)00109-9
- Xu, S., Zhang, Y., Zhen, Z., & Liu, J. (2021). The Face Module Emerged in a Deep Convolutional Neural Network Selectively Deprived of Face Experience. *Frontiers in Computational Neuroscience*, 15, 1–12. doi: 10.3389/fncom.2021.626259
- Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3), 356–365. doi: 10.1038/nn.4244
- Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014, jun). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23), 8619–8624. doi: 10.1073/PNAS.1403112111
- Zhou, L., Yang, A., Meng, M., & Zhou, K. (2022). Emerged human-like facial expression representation in a deep convolutional neural network. *Science Advances*, 8(12), 4383. doi: 10.1126/SCIADV.ABJ4383