Exploring Parallels and Discrepancies Between the Biological Visual System and Self-Supervised Learning

Vidhi Jain[1], Pramod Kaushik[2], and Bapi Raju[2]

[1] Netaji Subhas University of Technology(NSUT), Delhi, India
[2] International Institute of Information Technology (IIIT), Hyderabad, India

Abstract. Understanding the functioning of the human visual system has been a long-standing goal in neuroscience. Similarly, the field of artificial intelligence has made significant progress in developing Self-Supervised learning algorithms that can learn representations from unlabeled data, resembling aspects of human perception. This research paper examines the parallels and discrepancies between the biological visual system and Self-Supervised learning methods, aiming to shed light on the extent to which these artificial models capture the intricacies of human visual perception. By examining various properties observed in human perception and comparing them with the behaviors of Self-Supervised learning algorithms, we aim to gain insights into the strengths and limitations of these models in emulating the complexity and efficiency of the biological visual system.

Keywords: Human Vision · Self-Supervised Learning · Alignment of deep networks with human vision · Contrastive Learning · Representational Learning

1 Introduction

The human visual system is a remarkable feat of evolution, capable of perceiving and understanding complex visual scenes effortlessly. This system, refined through millions of years of evolution, excels in processing a myriad of visual information ranging from low-level features like edges and textures to high-level object recognition and scene understanding. The brain's ability to extract meaningful information from visual stimuli has inspired the development of Convolutional Neural Network (CNN) [23] systems, particularly Self-Supervised Learning (SSL) algorithms, that attempt to mimic the efficiency and robustness of the biological visual system. Despite the advancements in CNNs and the promising direction of SSL, a fundamental question persists: how closely do these artificial systems mirror the biological visual system? While state-of-the-art results in visual perception tasks have been achieved, these systems' ability to replicate the depth and nuance of human vision remains debatable.

Recent studies reveal remarkable differences between human vision and CNNs [27, 24, 8, 4, 25, 11, 2]. Bower et al [4] argue that DNNs account for almost no results from psychological research. Jacob et al [15] focused on perceptual and

neural phenomena in terms of distance comparisons portrayed by human perception, and ask whether they are present in feed-forward deep neural networks trained for object recognition. They concluded that phenomena like Weber's law or 3D shape and depth may be only seen when models are trained differently.

Supervised CNNs trained on ImageNet seem to exhibit some biases in perception and representation that are not present in humans. Prior research found preferences for shape over color [26] and perceptual shape over physical shape [19]. Recent research has claimed that, whereas people demonstrate a shape bias, preferring to classify items according to their shape [21, 20, 9], ordinary ImageNet-trained CNNs favor texture[10].

Recent studies suggest that Self-Supervised networks mimic human vision better than the traditional supervised CNN model as it allows the model to learn from the data in a more unstructured and natural way, similar to how humans learn from their surroundings [12, 17]. These algorithms leverage the inherent structure in the data itself to generate supervisory signals, allowing the models to discover relevant features and patterns.

The theoretical foundation of our study is rooted in cognitive psychology and neuroscience, particularly in how the human brain processes visual information. Our hypotheses are formulated based on existing knowledge about human perception phenomena, such as Weber's law and the Gestalt principles. These hypotheses guided our evaluation criteria and the interpretation of the models' performance.

This study conducts a comprehensive comparative analysis between the human visual system and Convolutional Neural Networks (CNNs), particularly focusing on Self-Supervised Learning (SSL) models. The research is notable for its investigation into specific phenomena observed in human perception, such as shape-texture bias, mirror confusion, scene incongruence, correlational sparseness, and global advantage [15, 10]. A pivotal aspect of this study is the in-depth examination of various SSL methods, encompassing both Representation and Contrastive SSL networks, to assess their alignment with human visual processing. Through both quantitative and qualitative analyses, the paper offers a nuanced assessment of the strengths and limitations of SSL algorithms in replicating the sophistication and efficiency of the biological visual system. These contributions collectively advance our understanding of the intricate relationship between artificial intelligence models and the biological visual system, offering a foundation for future innovations in both neuroscience and AI.

In the following sections, we will discuss the models and experiments, present the findings obtained from our analysis, and draw conclusions based on the observed results.

2 Methods

This section outlines the selection and implementation of Self-Supervised Learning (SSL) models used. The SSL models are evaluated using two thematic tests: Shape-Texture Bias and Quantitative Phenomena. These experiments are de-

signed to assess the capability of the models to replicate human visual perception.

2.1 Self-Supervised Learning (SSL)

SSL is a machine learning technique that has the potential to address the issues caused by the excessive reliance on labeled data, particularly for the learning of visual features. In Self-Supervised learning, the model is given an auxiliary pretext task made from the input data itself, and as it completes the auxiliary task, it learns about the underlying structure of the data. Table 1 shows all the models used in this work. These models span various architectures, including Vision Transformers (ViT) and ResNet-50, each offering unique approaches to learning and feature extraction. The theoretical basis for choosing these models is their varied handling of pretext tasks, which directly influences their learning patterns and outcome similarities to human vision.

Table 1. Self-Supervised Models

Model	Architecture	Top 1 accuracy	Year
DINOv2 [6]	ViT	86.5%	2023
iBOT [7]	ViT	82.3%	2021
SimCLR [7]	ResNet-50	76.5%	2020
SwaV [6]	ResNet-50	78.5%	2020
DeepCluster [5]	VGG-16	79.9%	2018
IPCL [18]	AlexNet	55.7%	2022
UNICOM[1]	ViT	79.3%	2023
MAE [14]	ViT	76.6%	2021
BYOL[13]	ResNet-50	78.5%	2020

Contrastive Learning Contrastive learning, a technique that aims to learn representations by contrasting similar and dissimilar examples, mimics the human ability to differentiate between objects and patterns in our visual environment. In contrast to the constant exposure to diverse visual stimuli, our brains naturally perform a form of contrastive learning, enabling us to recognize objects, perceive depth, and make sense of our surroundings.

In the context of SSL, contrastive learning attempts to replicate this aspect of human perception. By training models to distinguish between similar and dissimilar examples, meaningful representations of data can be learned. These representations have various applications, such as object recognition, image classification, and semantic understanding.

The contrastive SSL models that show the highest Top-1 accuracy for Imagenet: based on the VIT architecture model, DINOv2 [6] and iBOT [7], based on the Resnet-50 model, SWAV [6] and SimCLR [7] and based on VGG16 model,

DeepCLuster [5], are used in this work. Konkle et al. (2021) [18] argued that IPCL (Instance Prototype Contrastive Learning) achieves more human brain-like feature representations. This approach outperforms other instance contrastive learning methods and aligns with the development of visual system representation. Thus, IPCL is also used in this work.

Representational learning Representation-based SSL is an approach that aims to learn meaningful representations from unlabeled data by utilizing specific pretext tasks. These pretext tasks are designed to encourage the model to learn high-level features that capture the underlying structure and semantics of the data. By learning representations that align with human perception, representational SSL seeks to bridge the gap between the learned representations and human-level understanding.

The concept of representational SSL is closely related to human perception. It takes inspiration from the idea that humans naturally learn to recognize objects, understand scenes, and perceive meaningful patterns from visual information without explicit supervision. By designing pretext tasks that mimic certain aspects of human perception, Representational SSL methods aim to learn representations that are more aligned with human-level understanding. In representation SSL, the focus is on learning feature representations that capture the underlying structure and meaningful patterns in the data. The goal is to learn a compact and informative representation of the input data that can be used for downstream tasks. Methods such as autoencoders, Masked Autoencoders [14], Unicom [1], and BYOL [13] fall under this category and are used in this work.

2.2 Psychological Experiments

Shape Texture Bias Psychological research consistently indicates that shape plays a crucial role in human object identification [22, 10, 3, 20]. Geirhos et al. conducted a study comparing human and ImageNet-trained CNNs' performance on a dataset consisting of images with conflicting shape and texture information [10]. Based on that study, to investigate texture and shape biases, we conducted six experiments involving different datasets manipulated in various ways. The datasets corresponded to specific manipulations, including sketches, edge-filtered images, silhouettes, images with conflicting texture-shape cues, and stylized images where the original texture was replaced by a painting style. The goal is to assess whether different Self-Supervised networks exhibit biases towards shape or texture. The accuracy for each dataset and the shape bias for the cue-conflict dataset were measured to determine the presence of shape or texture biases in the networks.

Qualitative Phenomenons To assess different qualitative properties in both brains and deep neural networks to explore the similarities and differences in their visual object representations, the SSL models were tested out for 13 different qualitative phenomena (Thatcher Effect, Mirror Confusion, Scene Incongruence, Multiple Objects, Correlational Sparseness, Weber's Law, Relative

size, Surface Invariance, 3D processing, Occulusion and Depth, Object parts and Global Advantage) [15]. For the experiments, carefully chosen sets of photos are used to characterize the network input for each attribute. It is first established whether the activations of the units in each layer for these images demonstrate that property. Since activations in this layer are ultimately used to categorize objects, the model only shows the property if it was present in the final fully linked layer. Euclidean distance is used to compare the feature representations of each layer of the network to compute the final index of the property as in the CNN hidden layer, as its representation space has a strong correlation with similarity in human perception [16, 28]. For VIT-based models, the feature representation of each block is used.

3 Results

The results of the experiments conducted in this study provide valuable insights into the performance and characteristics of different models in the context of self-supervised learning (SSL).

3.1 Experimental set-up and Datasets

The experimental framework was designed to assess the models' ability to process and interpret visual stimuli. For shape-texture bias, we used modified versions of the ImageNet dataset, where images were altered to accentuate specific visual properties like edges, textures, and colors. For the qualitative experiments, carefully chosen sets of photos are used to characterize the network input for each attribute. The datasets were processed through different augmentation techniques to challenge and evaluate the models' ability to generalize from complex visual inputs.

3.2 Evaluating Shape-Texture Bias

The results presented in Figure 1 illustrate the accuracies of the Contrastive (Black), Representational (Blue), and Base (Red) models across various image types, including Original ImageNet, Greyscale, Sketch, Silhouette, Edges, and Stylized Imagenets. When considering Greyscale versions of objects, which retain both shape and texture, all models performed equally well in recognition. Similarly, Sketches that preserved the structure and most of the texture but lacked color experienced a slight decrease in accuracy across all models. For Silhouette images where the object outlines were filled in with black, most models exhibited a decrease in accuracy. However, SWAV and iBot models surpassed the accuracy achieved with sketch-based ImageNet. This trend was even more evident for edge stimuli, suggesting that human observers are better able to classify images with minimal texture information compared to SSL models, which struggled with classifying images based solely on edges. In the case of Stylized

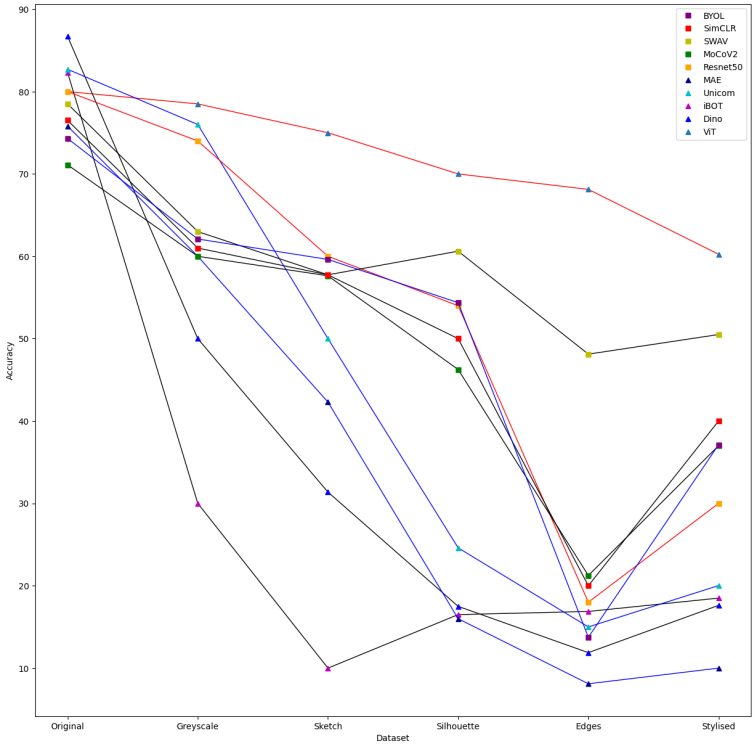

Fig. 1. SSL model accuracies across Original, Greyscale, Silhouette, Sketch, Edges and Stylised Imagenet Datasets

ImageNet, where images had the same shape but different textures, most models demonstrated higher accuracy compared to the edge-based representation of images.

The findings from the cue-conflict experiment are depicted in Figure 2. It is evident that human observers exhibit a pronounced inclination to respond based on the shape category. In contrast, Resnet-based SSL networks demonstrate a distinct inclination towards responding based on the texture category. On the other hand, both VIT and VIT-based models exhibit an equal inclination towards both texture and shape categories.

3.3 Qualitative Phenomenona Analysis

Table 2 shows the presence or absence of different perceptual effects across various networks based on the Vision Transformer (VIT) architecture, including DINOv2, iBOT, Unicom, MAE, Auto-encoders, and VIT itself. It can be seen that only Unicom and VIT show an ability to perceive the Thatcher Effect, where the orientation of facial features in an inverted face becomes difficult to detect.

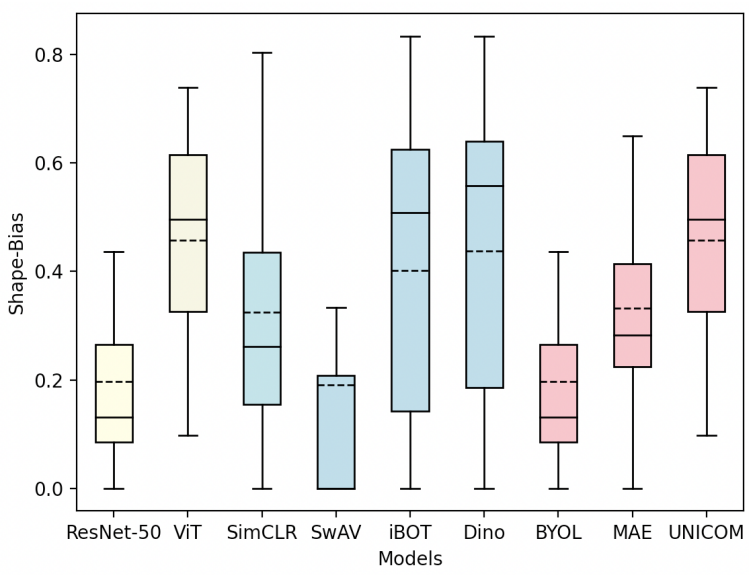

Fig. 2. Shape-Texture Bias across Base, Contrastive, and Representational SSL models

Table 2. Properties across all the VIT-based models

Perceptual effect	DINOv2	iBOT	Unicom	MAE	Auto-encoders	VIT
Thatcher Effect	No	No	Yes	No	No	Yes
Mirror Confusion	Yes	Yes	Yes	Yes	Yes	Yes
Scene Incongruence	Yes	Yes	Yes	Yes	Yes	Yes
Multiple Objects	Yes	Yes	Yes	Yes	Yes	Yes
Correlational Sparseness	Yes	Yes	Yes	Yes	Yes	Yes
Weber's Law	Yes	Yes	No	No	No	No
Relative Size	No	No	No	No	No	No
Surface Invariance	Yes	Yes	Yes	No	No	No
3D processing	No	No	No	No	No	No
Occlusion	No	No	No	No	No	No
Depth	Yes	No	Yes	No	No	No
Object parts	No	No	No	No	No	No
Global Advantage	Yes	Yes	Yes	No	Yes	No

All the networks based on VIT architecture can perceive Mirror Confusion; detect inconsistencies or oddities in a scene; perceive multiple objects in an image

and show Correlational Sparseness. However, none of the models can perform 3D processing or handle Occlusion. Moreover, VIT and SSL models do not have the capability to perceive individual object parts. Only Unicom and DINOv2 show the ability to perceive depth, allowing for the understanding of spatial relationships in a scene. DINOv2, iBOT, Unicom, and Auto-encoders exhibit a global advantage, meaning they can perceive the overall structure or global features of an image. However, MAE and VIT lack this ability. Hence most of the SSL models show global advantage. DINOv2 and iBOT can perceive Weber's Law, while MAE, Auto-encoders, and VIT cannot. Thus contrastive SSL based on VIT models show the ability to perceive differences in stimuli relative to their background. DINOv2, iBOT, and Unicom can perceive Surface Invariance, which refers to the ability to recognize objects regardless of variations in surface appearance. However, MAE, Auto-encoders, and VIT lack this ability.

Table 3. Properties across all the Resnet50-based models

Perceptual effect	**Resnet-50**	**Stylised Resnet-50**	**SIMCLR**	**SWAV**	**BYOL**
Thatcher Effect	No	No	No	Yes	Yes
Mirror Confusion	Yes	Yes	No	Yes	Yes
Scene Incongruence	Yes	Yes	Yes	Yes	Yes
Multiple Objects	Yes	Yes	Yes	Yes	Yes
Correlational Sparseness	Yes	Yes	Yes	Yes	Yes
Weber's Law	Yes	No	No	No	No
Relative Size	Yes	No	No	No	No
Surface Invariance	No	No	No	No	No
3D processing	No	Yes	Yes	Yes	No
Occlusion	No	No	No	No	No
Depth	No	No	Yes	Yes	Yes
Object parts	No	No	No	No	No
Global Advantage	No	No	Yes	No	Yes

Table 3 compares the perceptual effects exhibited by different ResNet50-based models, including ResNet50, Stylised Resnet 50, SIMCLR, SWAV, and BYOL. To test out if inducing shape bias makes the models more compliant with phenomenons experienced by the human visual system, this study uses VGG-16 and ResNet50 trained on the basis of Stylised dataset, demonstrating that it is now less texture biased and that it learns object shapes during training rather than relying on shortcuts like memorizing texture. While all models can perceive mirror confusion, scene incongruence, multiple objects, and correlational sparseness, they differ in their capability to perceive other effects. SWAV and

BYOL show strengths in perceiving the Thatcher effect, whereas ResNet-50, Stylised ResNet-50, and SimCLR lack this ability. ResNet-50 exhibits relative size perception, while the other models do not. Only Stylised ResNet, SimCLR, and SWAV demonstrate 3D processing, and SimCLR and BYOL have a global advantage.

Table 4. Properties across all the VGG16-based models

Perceptual effect	**VGG-16 random**	**VGG-16**	**Stylised VGG-16**	**DeepCluster**	**IPCL**
Thatcher Effect	No	No	No	Yes	No
Mirror Confusion	Yes	Yes	Yes	No	Yes
Scene Incongruence	Yes	Yes	Yes	Yes	Yes
Multiple Objects	Yes	Yes	Yes	Yes	Yes
Correlational Sparseness	Yes	Yes	Yes	Yes	Yes
Weber's Law	Yes	Yes	Yes	No	No
Relative Size	Yes	Yes	Yes	No	No
Surface Invariance	No	No	Yes	Yes	Yes
3D processing	No	No	Yes	No	Yes
Occlusion	No	No	No	No	No
Depth	No	No	No	No	No
Object parts	No	No	No	No	No
Global Advantage	Yes	No	No	No	No

Table 4 presents a comparison of perceptual effects observed in various VGG16-based models, including VGG-16 random, VGG-16, Stylised VGG-16, DeepCluster, and IPCL. These models exhibit different levels of perceptual capabilities. Mirror confusion, scene incongruence, multiple objects, correlational sparseness, and Weber's Law are perceived by VGG models, but not by IPCL. Relative size perception is demonstrated by VGG-16 random, VGG-16, and Sylised VGG-16, while surface invariance is observed in Sylised Resnet-50, DeepCluster and IPCL. However, only Stylised VGG-16 possess 3D processing. None of the models show occlusion perception, depth perception, or the ability to perceive object parts. The global advantage is present in only VGG-16 random, thus training induces few properties as well as blocks also.

In summary, the tables highlight the varying perceptual abilities of different networks based on the VIT, Resnet-50, and VGG-16 architectures. Each network demonstrates strengths and weaknesses in perceiving different perceptual effects, with some networks exhibiting more comprehensive perceptual capabilities compared to others.

4 Discussion

The present study aimed to investigate and compare the perceptual abilities of both humans and deep neural networks, specifically focusing on shape and texture bias and qualitative properties. The results shed light on the similarities and differences between human visual perception and the SSL models used in CNNs, providing valuable insights into the representational learning capabilities of these models.

The observed shape bias in human perception suggests that humans heavily rely on structural information for categorizing and identifying objects. Conversely, ResNet-based SSL networks exhibit a strong bias towards texture, highlighting a disparity between computational models and human visual perception. In contrast, Vision Transformer (VIT) and VIT-based models display an equal bias towards both texture and shape categories, bridging the gap between human perception and computational models.

Further, the ability of the SSL models to perceive various qualitative phenomena demonstrates their capacity to capture important visual attributes. However, certain limitations were observed. For example, none of the models demonstrated the ability to perform 3D processing or handle occlusion. This suggests that current self-supervised models still struggle with complex spatial relationships and object occlusion, which are fundamental aspects of human visual perception.

4.1 Contrastive vs Representational SSL

The results of this study provide valuable insights into the performance of Representation-based Self-Supervised Learning (SSL) networks and Contrastive SSL networks in capturing human perception properties. Representation-based SSL networks, such as UNICOM, BYOL, MAE, and Auto-Encoders, demonstrate a mixed performance in capturing various human perception phenomena. While they excel in aspects like mirror confusion, scene incongruence, correlational sparseness, and global advantage, they struggle with phenomena like 3D processing, Weber's Law, and relative size perception. These findings suggest that SSL approaches based on unsupervised representation learning can effectively capture and encode certain perceptual phenomena, indicating their potential in understanding complex spatial relationships and contextual information in visual scenes. Moreover, auto-encoder-based models generally don't perceive the Thatcher Effect and Depth facilitating the comprehension of spatial arrangements within a scene.

On the other hand, Contrastive SSL networks, including SimCLR, DINOv3, iBoT, SWAV, DeepCluster, and IPCL, generally exhibit better performance in capturing a wider range of human perception phenomena compared to Representation-based SSL networks. They demonstrate a better understanding of perceptual effects like Surface Invariance, Weber's Law, and 3D processing. Contrastive SSL models based on the ResNet-50 architecture, for example, show a strong capability for 3D processing. These models also demonstrate mirror confusion, scene

incongruence, and correlational sparseness, indicating their ability to capture complex relationships and context in learned representations.

The differences between the two types of SSL models are also evident when evaluating different image datasets. While all models perform equally well in recognizing grayscale versions of objects that retain both shape and texture, representational SSL models exhibit lower accuracy compared to Contrastive SSL models for datasets like silhouettes and edges. This highlights the robustness of Contrastive SSL models in handling variations in texture and their ability to generalize to different image properties.

It is important to note that both representation-based and contrastive SSL networks face challenges in understanding certain perceptual aspects such as occlusion perception, depth perception, and object parts recognition. These phenomena remain difficult for SSL models to fully comprehend, and further research is needed to address these limitations.

These insights highlight the importance of the SSL approach in expanding the perceptual capabilities of neural networks. Future research should focus on refining and developing novel SSL architectures and techniques to better capture the remaining perceptual phenomena. By addressing these limitations, SSL models have the potential to significantly advance the understanding and application of human-like perception in artificial systems.

4.2 Supervised v/s Self-Supervised Learning

Firstly, it is evident that Supervised networks, which are primarily trained on labeled data for classification tasks, have limitations in capturing a wide range of human perception phenomena (VGG16, Resnet50, and VIT). These networks tend to focus on learning discriminative features for specific object recognition tasks, resulting in a narrower understanding of perceptual effects. Properties such as the Thatcher Effect, Surface-Invariance, 3D processing, depth, perceived individual object paths, and global advantage are often not adequately represented in the learned features of Supervised networks.

On the other hand, SSL networks, which learn representations from unlabeled data using various pretext tasks, exhibit a more comprehensive understanding of human perception phenomena. By leveraging the structure and patterns present in unlabeled data, SSL networks develop more holistic representations that encompass a wider range of perceptual phenomena. Most of the SSL networks show the perception and processing of 3d and depth information allowing for the understanding of spatial relationships in a scene. They often also show Thatcher effect, surface invariance, and global advantage, thus having the ability to recognize objects regardless of variations in surface appearance and perceive the overall structure or global features of an image. As seen in Table 4, unlike other instances of VGG-16, Deep Cluster didn't show mirror confusion , or Weber's law and Relative size, however, it did show surface invariance and 3D processing.

However, it is important to note that even SSL networks face challenges in capturing certain perceptual aspects such as occlusion perception and object parts recognition. These phenomena remain difficult for both Supervised and

Self-Supervised models to fully comprehend. They also see changes in size based on absolute size (Weber's Law) and don't relate to the object part sizes.

Overall, the findings suggest that SSL approaches hold great potential in expanding the perceptual capabilities of neural networks. By leveraging unlabeled data and adopting various pretext tasks, SSL models can capture a broader range of human perception phenomena. The ability of these models to go beyond simple object recognition tasks and develop more nuanced perceptual understanding opens up avenues for applications in fields such as computer vision, robotics, and artificial intelligence.

4.3 Feature representation of models

The presence or absence of specific perceptual properties in a network can provide insights into the features represented by the model.

In the case of Contrastive SSL models, which are trained on instances with random crop and rotations, they exhibit robustness to the inversion effect, processing inverted and upright faces similarly. This is because most of these models do not show the Thatcher effect. These models learn features based on shape and boundaries, as evidenced by their ability to demonstrate 3D effects, surface invariance, and depth processing. Additionally, they successfully capture high-level features and prioritize learning beyond spatial hierarchies of features. This indicates a deeper understanding of the underlying semantics in the data. The lower layers of these models learn to detect simple edges and textures, while the higher layers learn to combine these lower-level features to form more complex shapes and contours. However, it should be noted that contrastive learning, with its choice of instance discrimination task, can introduce bias and influence the learned features. Different instance discrimination tasks in models like DINOv2, iBoT, IPCL, SWAV, SimCLR, and DeepCluster lead to variations in feature representation and the presence or absence of certain properties.

Representation-based SSL models, including auto-encoders, lack surface invariance and depth processing. However, they excel in capturing high-level visual representations that rely on global form and contour processing. Auto-encoders can consider both local and global shape similarities when grouping objects, similar to how the human brain encodes shape elements and their combinations during lower levels of visual processing. However, models based on masked auto-encoders, which predict missing patches, do not produce feature representations resembling human vision, even though they achieve high accuracy in image classification tasks. Other representation-based SSL models are capable of capturing fine details and spatial relationships of features, allowing them to identify distorted or inverted regions that create illusions. They can also capture information about relative distances, object occlusion, and perspective, which are essential for perceiving depth in a scene. These models exhibit the ability to discern spatial positioning and infer the three-dimensional structure of objects. Furthermore, they can represent visual features invariant to changes in surface properties, such as color, texture, or lighting conditions, enabling robust perception and recognition across different appearances.

The choice of base models also influences the properties exhibited by the SSL models. Notably, models based on the Vision Transformer (VIT) and VIT-based architectures demonstrate a lower bias towards texture compared to other models. This suggests that the architectural design of the base models plays a role in shaping the biases and capabilities of the models. The attention-based architecture of VIT models, which captures global relationships among image patches, contributes to their effective capture of both shape and texture information. On the other hand, models based on convolutional neural networks (CNNs) like ResNet may have a tendency to capture local texture patterns due to their hierarchical and spatially localized feature extraction processes.

It is evident that, in contrast to models trained on Imagenet, VGG-16 and Resnet-50 trained on Stylised Imagenet displayed Surface Invariance and 3D processing. It might be argued that Stylised Imagenet models, which favor shape, more accurately reflect the connections between 3D shape and object attribute.

5 Conclusion

In conclusion, the analysis of various models across different perceptual properties provides valuable insights into the capabilities of Supervised and Self-Supervised networks. Supervised networks, which focus on downstream tasks, tend to neglect high-level features and prioritize spatial hierarchies. On the other hand, SSL networks, such as Representation-based SSL and Contrastive SSL, demonstrate a better understanding of human perception phenomena. These models exhibit strengths in capturing spatial relationships, recognizing scene incongruence, representing correlational sparseness, and leveraging global advantages. They also showcase the ability to perceive the Thatcher effect, depth, and surface invariance. These findings suggest that Self-Supervised learning approaches have the potential to learn rich and meaningful representations that align with human perception, opening new avenues for advancing computer vision and artificial intelligence. However, further research is needed to explore the underlying mechanisms and enhance the performance of Self-Supervised models in diverse perceptual tasks.

References

1. An, X., Deng, J., Yang, K., Li, J., Feng, Z., Guo, J., Yang, J., Liu, T.: Unicom: Universal and compact representation learning for image retrieval (04 2023)
2. Azulay, A., Weiss, Y.: Why do deep convolutional networks generalize so poorly to small image transformations? (05 2018)
3. Baker, N., Lu, H., Erlikhman, G., Kellman, P.: Deep convolutional networks do not classify based on global object shape. PLOS Computational Biology **14** (12 2018). https://doi.org/10.1371/journal.pcbi.1006613
4. Bowers, J., Malhotra, G., Dujmović, M., Llera, M., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J., Heaton, R., Evans, B., Mitchell, J., Blything, R.: Deep problems with neural network models of human vision. Behavioral and Brain Sciences pp. 1–74 (12 2022). https://doi.org/10.1017/S0140525X22002813

5. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features (2018)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments (06 2020)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations (02 2020)
8. Fawzi, A., Frossard, P.: Manitest: Are classifiers really invariant? (07 2015). https://doi.org/10.5244/C.29.106
9. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**, 665–673 (11 2020). https://doi.org/10.1038/s42256-020-00257-z
10. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness (11 2018)
11. Geirhos, R., Temme, C., Rauber, J., Schütt, H., Bethge, M., Wichmann, F.: Generalisation in humans and deep neural networks (08 2018)
12. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations (03 2018)
13. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Pires, B., Guo, Z., Azar, M., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning (06 2020)
14. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners (11 2021)
15. Jacob, G., Pramod, R., Katti, H., Arun, S.: Qualitative similarities and differences in visual object representations between brains and deep networks. Nature communications **12**(1), 1–14 (2021)
16. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. vol. 9906, pp. 694–711 (10 2016). https://doi.org/10.1007/978-3-319-46475-6_43
17. Konkle, T., Alvarez, G.: Beyond category-supervision: instance-level contrastive learning models predict human visual system responses to objects (05 2021). https://doi.org/10.1101/2021.05.28.446118
18. Konkle, T., Alvarez, G.: A self-supervised domain-general learning framework for human ventral stream representation. Nature Communications **13** (01 2022). https://doi.org/10.1038/s41467-022-28091-4
19. Kubilius, J., Bracci, S., Op de Beeck, H.: Deep neural networks as a computational model for human shape sensitivity. PLoS computational biology **12**, e1004896 (04 2016). https://doi.org/10.1371/journal.pcbi.1004896
20. Kucker, S., Samuelson, L., Perry, L., Yoshida, H., Colunga, E., Lorenz, M., Smith, L.: Reproducibility and a unifying explanation: Lessons from the shape bias. Infant Behavior and Development **54** (10 2018). https://doi.org/10.1016/j.infbeh.2018.09.011
21. Landau, B., Smith, L., Jones, S.: The importance of shape in early lexical learning. Cognitive Development - COGNITIVE DEVELOP **3**, 299–321 (07 1988). https://doi.org/10.1016/0885-2014(88)90014-7
22. Landau, B., Smith, L., Jones, S.: The importance of shape in early lexical learning. Cognitive Development - COGNITIVE DEVELOP **3**, 299–321 (07 1988). https://doi.org/10.1016/0885-2014(88)90014-7
23. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)

24. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (12 2014)
25. RichardWebster, B., Anthony, S., Scheirer, W.: Psyphy: A psychophysics driven evaluation framework for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP** (11 2016). https://doi.org/10.1109/TPAMI.2018.2849989
26. Ritter, S., Barrett, D., Santoro, A., Botvinick, M.: Cognitive psychology for deep neural networks: A shape bias case study (06 2017)
27. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks (12 2013)
28. Zhang, R., Isola, P., Efros, A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric (01 2018)