Improving Unsupervised Task-driven Models of Ventral Visual Stream via Relative Position Predictivity

Dazhong Rong[1] **Hao Dong**[2] **Xing Gao**[3] **Jiyu Wei**[1] **Di Hong**[1]
Yaoyao Hao[1] **Qinming He**[1] **Yueming Wang**[1]*
[1]Zhejiang University, Hangzhou, China
{rdz98,weijiyu,hongd,yaoyaoh,hqm,ymingwang}@zju.edu.cn
[2]Rutgers University, New Jersey, USA
hd322@scarletmail.rutgers.edu
[3]University of Electronic Science and Technology of China, Chengdu, China
2022050903029@std.uestc.edu.cn

Abstract

Based on the concept that ventral visual stream (VVS) mainly functions for object recognition, current unsupervised task-driven methods model VVS by contrastive learning, and have achieved good brain similarity. However, we believe functions of VVS extend beyond just object recognition. In this paper, we introduce an additional function involving VVS, named relative position (RP) prediction. We first theoretically explain contrastive learning may be unable to yield the model capability of RP prediction. Motivated by this, we subsequently integrate RP learning with contrastive learning, and propose a new unsupervised task-driven method to model VVS, which is more inline with biological reality. We conduct extensive experiments, demonstrating that: (i) our method significantly improves downstream performance of object recognition while enhancing RP predictivity; (ii) RP predictivity generally improves the model brain similarity. Our results provide strong evidence for the involvement of VVS in location perception (especially RP prediction) from a computational perspective.

Keywords: ventral visual stream; task-driven models; unsupervised learning

Introduction

Ventral visual stream (VVS) is a neural pathway in the brain primarily involved in the processing of visual information (Cloutman, 2013; Zachariou, Klatzky, & Behrmann, 2014; Simic & Rovet, 2017), and its disruptions are linked to various visual disorders (*e.g.,* visual agnosia) (Greene, 2005). Task-driven methods (Nayebi et al., 2018, 2021; Yamins & DiCarlo, 2016), which aim to train artificial neural networks (ANNs) on a specific behaviorally relevant task rather than directly fitting neural data (Wei, Rong, Zhu, He, & Wang, 2024), are widely adopted to model VVS and have achieved considerable success. The outputs of intermediate layers of ANNs trained by these task-driven methods demonstrate high similarity to the neural responses in VVS, where the higher similarity indicates the architecture of the ANNs or the functionality of the learning task is more brain-like. Leveraging these task-driven methods, on the one hand, researchers can gain deeper insights into the potential neural mechanisms underlying the visual disorders; on the other hand, researchers can draw inspiration from biological vision to design ANNs that are more robust against adversarial examples (Zhu & Rong, 2024) and backdoor attacks (Rong et al., 2024).

Existing task-driven methods to model VVS can be roughly divided into two categories: supervised and unsupervised. Because of the better biological plausibility, recently unsupervised task-driven methods have received more considerable attention. Among these unsupervised methods, contrastive learning (Zhuang et al., 2021) achieves the best brain similarity in modeling VVS. However, we argue there is still room for improvement because the correspondence between the tasks of ANNs and the functions of VVS is not adequately considered.

In this paper, we conduct in-depth analysis of the correspondence in modeling VVS by contrastive learning. We first consider the core function of VVS: object recognition. Contrastive learning can endow the model with the capability of object recognition. The evidence is that after training the model by contrastive learning, the image features extracted by the model are linearly separable for object category (Chen, Kornblith, Norouzi, & Hinton, 2020; Newell & Deng, 2020; Ericsson, Gouk, & Hospedales, 2021). However, object recognition is not the only function of VVS. Some studies (Konen & Kastner, 2008; Kravitz, Kriegeskorte, & Baker, 2010; Sereno & Lehky, 2011; Zachariou et al., 2014) have suggested that the functional specialization of ventral and dorsal visual streams is not as distinct as originally proposed, with VVS also contributing to location perception, despite location perception is primarily performed by dorsal visual stream. Hence, in this paper we distill location perception into relative position (RP) prediction and additionally consider this function in modeling VVS. Unfortunately, because the constraints induced by contrastive learning are relative-position-agnostic, contrastive learning may be unable to endow the model with the capability of RP prediction.

To compensate for the above functional defect, in this paper we design RP learning which makes the model explicitly learn the RP between two sub-images, and propose a new unsupervised task-driven method to model VVS, combining contrastive learning and RP learning. Our method can handle the functions of object recognition and RP prediction simultaneously. The main contributions of this paper can be summarized as following:

1. We analyze the correspondence between two unsupervised tasks (contrastive learning and RP learning) and two functions (object recognition and RP prediction) of VVS in depth. To the best of our knowledge, we are the first to consider the function of RP prediction in modeling VVS.
2. We propose a new unsupervised task-driven method to

*Corresponding author.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

model VVS, which combines contrastive learning and RP learning. Extensive experiments demonstrate that our proposed method sets the new state-of-the-art brain similarity of unsupervised task-driven models of VVS.

3. Furthermore, our experimental results show RP predictivity not only enhances downstream performance of object recognition, but also improves the model brain similarity to various cortical regions (V1, V2, V4, IT) on VVS, proving VVS is indeed involved in the function of RP prediction.

Related Work

Some recent work (Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014) has found that when achieving better performance of artificial neural networks (ANNs) on certain behaviorally relevant tasks (*e.g.,* object recognition), the representations of ANNs' intermediate layers simultaneously become more similar to the recorded neural activities of primate VVS. Based on this finding, various task-driven methods (Güçlü & Van Gerven, 2015; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Kubilius, Bracci, & Op de Beeck, 2016; Yamins & DiCarlo, 2016), which optimize ANNs for better task performance rather than fitting the recorded neural activities directly, are proposed to model primate VVS and obtain great success. The higher similarity between the intermediate outputs of ANNs and the neural activities of VVS indicates the model architecture or the task functionality is more brain-like.

Existing task-driven methods for modeling primate VVS can be categorized according to whether the task is supervised or unsupervised. In the following, we summarize two sorts of work respectively.

Supervised Task-Driven Methods

VVS consists of a cascade of cortical areas which are hierarchically organized and anatomically distinguishable. The low-level visual features (*e.g.,* local orientation and spatial scale) are captured in the primary visual area V1, and then are processed into some more complex features in the mid-level visual areas V2, V3 and V4 (Carandini et al., 2005; Sincich & Horton, 2005; DiCarlo, Zoccolan, & Rust, 2012; Schiller, 1995). Lastly, the most high-level features with linear separability of object category are obtained in the highest visual area IT (inferior temporal) (Brincat & Connor, 2004; Yamane, Carlson, Bowman, Wang, & Connor, 2008; Hung, Kreiman, Poggio, & DiCarlo, 2005).

Based on the above theories about the emergence of the capability to recognize objects in the brain, existing supervised task-driven methods commonly adopt object recognition as the learning task and differ in other settings. (Shi, Shea-Brown, & Buice, 2019) trains VGG-16, a multi-layer convolutional neural network (CNN), to model mouse visual cortex. (Cadena et al., 2019) trains VGG-19 to model macaque V1 area. For better matching the primate VVS, (Kubilius et al., 2019) specially designs a shallow recurrent CNN named CORnet-S consisting of four areas anatomically mapped to V1, V2, V4, and IT, and then trains CORnet-S to model macaque visual cortex. Different from conventional ANNs, spiking neural networks (SNNs) encode information with time sequences of spikes, and hence act more like biological neurons. For this better biological plausibility, (Huang, Ma, Yu, Zhou, & Tian, 2023) trains SEW ResNet, a type of SNNs, to model both mouse visual cortex and macaque visual cortex.

Unsupervised Task-Driven Methods

Despite the great success in achieving high brain similarity, all of the above supervised studies fail to explain: how the representations in the brain are learned? These studies without exception assume the brain can receive both massive visual stimuli and the corresponding category labels, hence the supervised training. However, factually there is a lack of convincing evidence to prove that the brain receives these huge numbers of category labels, especially for nonhuman primates and human infants. This implausible assumption may cause inaccuracies in subsequent research on neural mechanisms of primate visual system.

Compared to supervised learning, unsupervised learning is more in line with the real situation. In the brain, the capability to recognize objects is not obtained by training the neurons all the time to correctly categorize the input visual stimuli, but rather it emerges naturally through some unsupervised training tasks (Zhuang et al., 2021). This is consistent with recent findings in the field of computer vision, which suggest that pre-training on some unsupervised tasks can lead to better performance on downstream supervised tasks (Chen et al., 2020; He, Girshick, & Dollár, 2019; Erhan, Courville, Bengio, & Vincent, 2010). There have been several studies devoted to modeling VVS through unsupervised task learning. (Higgins et al., 2021) trains a beta-variational auto-encoder (β-VAE) on image reconstruction task to model macaque IT area. (Lotter, Kreiman, & Cox, 2020) trains a recurrent generative network (PredNet) on the task of predicting future video frames to model macaque visual cortex. (Zhuang et al., 2021) trains ResNet-18 by SimCLR, a classic framework for contrastive learning, to model macaque V1, V4, and IT areas. Note that it achieves the best brain similarity among the above studies.

However, unsupervised task-driven modeling of VVS is still under-explored. Most of existing studies just try various unsupervised tasks to pursue higher similarity score between the intermediate outputs of ANNs and the neural activities recorded in visual cortex, but lacks consideration of the correspondence between the unsupervised tasks of ANNs and the functions of VVS.

Methods

Modeling VVS involves two stages: **(i)** training a base model on image dataset, and **(ii)** evaluating the brain similarity of the trained model to four different cortical regions (V1, V2, V4, IT) of VVS on neural dataset. In this section, we focus on the

first stage and present our proposed unsupervised task-driven method to train the base model in detail.

Preliminaries

The base model takes images as inputs and extracts their features as outputs. Formally, we denote the high-dimensional Euclidean space of input images by $\mathbb{M}^{c\times k\times k}$, where c and k respectively represent the number of channels and the number of pixels along each side of the images. Besides, we denote the n-dimensional Euclidean space of the extracted feature vectors by $\mathbb{V}^n$. Then the base model can be treated as a learnable function $f:\mathbb{M}^{c\times k\times k}\to\mathbb{V}^n$.

In each round of training, a batch of images denoted by $\{\mathbf{M}_i\}_{i=1}^N$ are randomly sampled from image dataset, where $\mathbf{M}_i$ is the i-th image in the current batch and N is the batch size. In order to model VVS by unsupervised learning, we need to design a specific unsupervised learning task, indicated by a loss function $\mathcal{L}$. Then in each training round, we optimize the learnable base model f by minimizing the loss $\mathcal{L}(\{\mathbf{M}_i\}_{i=1}^N)$.

Contrastive Learning

Since object recognition is the core function of primate VVS (Zhuang et al., 2021; Huang et al., 2023), the first goal of our designed unsupervised learning task is to endow the base model with the capability to recognize objects. Recent work (Chen et al., 2020) proposes SimCLR, one of the most widely used frameworks for contrastive learning (CL), and demonstrates that after pre-training by SimCLR, a simple linear classifier trained on top of the frozen learned features can achieve significantly high image classification accuracy. This indicates the features learned by SimCLR are linearly separable for object category, proving that the base model has obtained the capability to recognize objects. This is also very consistent with the biological reality that the low-level visual features are progressively processed into the high-level visual features containing linear separability of object category through primate VVS (Yamane et al., 2008). Due to the above facts, we adopt SimCLR to train our base model.

Specifically, in each round of training, $2N$ variants are generated from the N images through data augmentations (*e.g.,* crop, flip and color distortion), denoted by $\{\mathbf{X}_i\}_{i=1}^{2N}$. Note that $\mathbf{X}_{2i-1}$ and $\mathbf{X}_{2i}$ are the variants of $\mathbf{M}_i$. An additional learnable projection head denoted by $g:\mathbb{V}^n\to\mathbb{V}^m$ is involved in SimCLR, which is a multi-layer perceptron (MLP) and projects the features extracted by our base model into the m-dimensional vector space. Based on our base model f and the projection head g, we define the similarity between two variants as $\mathbf{sim}(\mathbf{X}_i,\mathbf{X}_j)=\frac{g(f(\mathbf{X}_i))^{\mathrm{T}}g(f(\mathbf{X}_j))}{\|g(f(\mathbf{X}_i))\|\|g(f(\mathbf{X}_j))\|}$, and define $\tilde{s}_{i,j}=\frac{\exp(\mathbf{sim}(\mathbf{X}_i,\mathbf{X}_j))}{\tau}$, where τ is the temperature parameter. Furthermore, we define $\tilde{s}_i=\sum_{j=1}^{2N}\tilde{s}_{i,j}-\tilde{s}_{i,i}$. Finally, the CL loss can be represented as following:

$$\mathcal{L}_{\mathrm{CL}}=-\frac{1}{2N}\sum_{i=1}^{N}\big(\log\frac{\tilde{s}_{2i-1,2i}}{\tilde{s}_{2i-1}}+\log\frac{\tilde{s}_{2i,2i-1}}{\tilde{s}_{2i}}\big). \tag{1}$$

Relative Position Learning

Due to the considerations previously mentioned in Section , the second goal of our designed unsupervised learning task is to endow the base model with the capability to predict relative position (RP). Specifically, we abstract RP prediction into a function: taking two neighboring images $\mathbf{M}_A$ and $\mathbf{M}_B$ as input, predicting the position of $\mathbf{M}_B$ relative to $\mathbf{M}_A$ as output.

CL can only yield the capability to answer whether two images are neighboring, but not the capability to answer what the RP between two images is. In CL, two neighboring images $\mathbf{M}_A$ and $\mathbf{M}_B$ can be viewed as two variants generated by data augmentations (resize and crop) from a larger image containing $\mathbf{M}_A$ and $\mathbf{M}_B$ concurrently, and hence their representations are constrained to be of high cosine similarity by CL. In other words, the value of $\mathbf{sim}(\mathbf{M}_A,\mathbf{M}_B)$ is very close to 1. Therefore, with the model well trained by CL, we can easily estimate the probability that two images are neighboring (*i.e.,* derived from the same larger image) according to the calculated cosine similarity between their representations. However, the constraints induced by CL is completely relative-position-agnostic. Specifically, the representations of $\mathbf{M}_A$ and $\mathbf{M}_B$ are equally constrained regardless of whether $\mathbf{M}_B$ is to the left or right of $\mathbf{M}_A$. Therefore, the representations learned by CL do not carry any information about RP, ultimately resulting in the model failing to handle RP prediction. As we aforementioned in Section , this may be inconsistent with the reality reported in recent studies (Konen & Kastner, 2008; Kravitz et al., 2010; Sereno & Lehky, 2011; Zachariou et al., 2014) that VVS also contributes to location perception.

To compensate for the above functional defect of CL, we specially design relative position learning (RPL), which makes the base model explicitly learn the knowledge about RP. Specifically, for each sampled image $\mathbf{M}_i$ in current training round, we first divide it into four equal-sized 2×2 blocks denoted by $\mathbf{X}_i^0$, $\mathbf{X}_i^1$, $\mathbf{X}_i^2$, and $\mathbf{X}_i^3$ respectively. Then we randomly select two unequal numbers from $\{0,1,2,3\}$, denoted by a_i and b_i ($a_i\neq b_i$), and obtain the label d_i indicating the RP of $\mathbf{X}_i^{b_i}$ to $\mathbf{X}_i^{a_i}$. Note that the value of d_i is an integer in the range of 0 to 7, indicating eight directional categories (left, right, upper, lower, upper-left, upper-right, lower-left, and lower-right) respectively. An additional learnable classifier head $h:\mathbb{V}^{2n}\to\mathbb{V}^8$ is involved in RPL, which is an MLP with 8-dimensional softmax outputs. The classifier takes the concatenation of two image features extracted by the base model f as inputs, and outputs the predicted probabilities of the eight directional categories. For convenience, we define $\tilde{d}_i=h(f(\mathbf{X}_i^{a_i})\oplus f(\mathbf{X}_i^{b_i}))$, where $\oplus$ represents vector concatenation. Finally, our RPL loss can be represented as following:

$$\mathcal{L}_{\mathrm{RPP}}=\frac{1}{N}\sum_{i=1}^{N}l(\tilde{d}_i,d_i), \tag{2}$$

where l is the cross entropy loss as following:

$$l(\tilde{d}_i,d_i)=\frac{1}{8}\sum_{j=0}^{7}\sigma(d_i=j)\cdot\log(\tilde{d}_{i,j}). \tag{3}$$

Note that $\tilde{d}_{i,j}$ is the j-th element of $\tilde{d}_i$, and σ is the indicator function. $\sigma(e) = 1$ if e is true; otherwise, $\sigma(e) = 0$.

After RPL, the RP information is embedded into the features extracted by the base model f, further utilized by the classifier head h to predict RP. From this aspect, the base model f has obtained the capability of RP prediction.

Method Overview

For training our base model f to concurrently handle the two visual functions (object recognition and RP prediction), we craft an unsupervised dual-task learning. Specifically, the overall loss of our proposed dual-task learning is the linear combination of the CL loss $\mathcal{L}_{\text{CL}}$ and the RPL loss $\mathcal{L}_{\text{RPL}}$, as following:

$$\mathcal{L} = \mathcal{L}_{\text{CL}} + \alpha \cdot \mathcal{L}_{\text{RPL}}, \tag{4}$$

where α is a hyper-parameter to balance the relative weight between two parts. Note that base model f, projection head g, and classifier head h are optimized simultaneously by minimizing $\mathcal{L}$.

Experiments

For thoroughly testing our proposed new unsupervised task-driven methods to model VVS, we conduct extensive experiments. In this section, we first introduce some basic experimental settings, and then demonstrate the results with detailed analyses. The source code of our experiments is available at `https://github.com/rdz98/Unsup-VVS`.

Datasets

Our experiments involve one image dataset (STL-10) and two neural datasets (Macaque-V1/V2 and Macaque-V4/IT). In the following, we introduce the three datasets respectively.

STL-10[†]. The STL-10 dataset (Coates, Ng, & Lee, 2011) is a widely used public benchmark dataset for image classification, containing 10 object classes with 5,000 labeled training images, 8,000 labeled test images, and 100,000 unlabeled images. Each image is 96×96 pixels in color.

Macaque-V1/V2[‡]. The neural responses of anesthetized macaque monkeys to 2,700 texture image stimuli were recorded in this dataset (Freeman, Ziemba, Heeger, Simoncelli, & Movshon, 2013), involving 102 V1 neurons and 103 V2 neurons. The texture image stimuli consists of 20 repetitions of 135 naturalistic or noise samples from different texture families.

Macaque-V4/IT[‡]. The neural responses of alert rhesus macaque monkeys to 3,200 synthetic images were recorded in this dataset (Majaj, Hong, Solomon, & DiCarlo, 2015), involving 88 neurons in V4 and 168 neurons in IT. The images were generated by adding a 2D projection varying position, size and pose concomitantly of a 3D object model to a random natural background. There were 64 different 3D object models (8 basic-level categories, each with 8 exemplars).

[†]`https://cs.stanford.edu/~acoates/stl10/`.

[‡]Both of the two neural datasets are publicly accessible on Brain-Score (Schrimpf et al., 2018, 2020).

Metrics

After training the model by the unsupervised task on the image dataset, we adopt three metrics to respectively evaluate the model from three aspects: image classification accuracy, relative position prediction accuracy, and brain similarity. Notably, the first two are computed on the image dataset and we refer to them collectively as task accuracy, while the last is computed on the neural datasets.

Image Classification (IC) Accuracy. When evaluating IC accuracy, we drop projection head g and classifier head h while retain the base model f only. We froze the base model f, and train a new linear classifier by L-BFGS optimization algorithm on labeled training samples in the image dataset. Specifically, for each labeled training sample $(\mathbf{M}_i, y_i)$, the new linear classifier takes $f(\mathbf{M}_i)$ as inputs, and takes y_i as the ground-true category label. After training the new linear classifier, we calculate its top-1 classification accuracy on labeled test set of the image dataset as our IC accuracy.

Relative Position Prediction (RPP) Accuracy. When evaluating RPP accuracy, we drop projection head g while retain the base model f and classifier head h. For each image $\mathbf{M}_i$ in the test set of the image dataset, we adopt the same approach used in RPL to generate a test sample consisting of $\mathbf{X}_i^{a_i}$, $\mathbf{X}_i^{a_i}$, and d_i. We take d_i as the ground-true label of RP, and take $h(f(\mathbf{X}_i^a) \oplus f(\mathbf{X}_i^b))$ as the predicted probabilities of the eight RP categories. Without any further training, we directly calculate top-1 classification accuracy on the generated test samples as our RPP accuracy.

Brain Similarity. When evaluating brain similarity, we drop projection head g and classifier head h while retain the base model f only. We calculate the layer-level brain similarity for each layer of the base model respectively. Specifically, we first input all visual stimuli from the neural dataset into f to obtain the layer outputs. Afterwards, we fit the layer outputs to predict the corresponding neural responses recorded in the neural dataset by Partial Least Squares (PLS) regression. Subsequently, for each neuron, we compute the Pearson correlation coefficient between predicted responses and recorded responses, and then correct it by the neuron noise ceiling. Finally, we obtain the layer-level brain similarity as the median of the noise-corrected correlation coefficients. The brain similarity of the whole base model is the maximum of the layer-level brain similarities across the layers in the base model.

Experimental Settings

Parameters. Considering the significant success of deep residual convolutional neural networks (He, Zhang, Ren, & Sun, 2016) on various computer vision tasks, we adopt ResNet-18 as our base model f. The extracted feature dimension $n = 512$, the projection dimension $m = 128$, and the batch size $N = 512$. Besides, both the projection head g and the classifier head h are 2-layer MLPs with hidden layer dimension of 512. In our unsupervised training on the image dataset, we adopt Stochastic Gradient Descent (SGD) optimizer with learning rate 1.5, weight decay 1×10^{-6}, and mo-

Table 1: Task Accuracy and Brain Similarity with Different Balancing Weights

Balancing Weight	Task Accuracy (%)		Brain Similarity			
	IC	RPP	V1	V2	V4	IT
$\alpha = 0$ (PCL)	82.60	14.19	0.2198 ± 0.0165	0.3367 ± 0.0143	0.5236 ± 0.0056	0.4771 ± 0.0029
$\alpha = 0.00002$	85.61	35.55	0.2391 ± 0.0139	0.3330 ± 0.0192	0.5237 ± 0.0045	0.5027 ± 0.0025
$\alpha = 0.0001$	85.68	80.44	0.2217 ± 0.0152	0.3439 ± 0.0141	0.5291 ± 0.0045	0.5015 ± 0.0033
$\alpha = 0.0005$	85.74	86.05	0.2390 ± 0.0099	0.3456 ± 0.0204	0.5312 ± 0.0050	**0.5045 ± 0.0035**
$\alpha = 0.002$	**86.34**	91.43	0.2440 ± 0.0115	0.3412 ± 0.0177	0.5304 ± 0.0041	0.4998 ± 0.0032
$\alpha = 0.01$	85.23	**95.28**	**0.2500 ± 0.0113**	**0.3521 ± 0.0146**	**0.5413 ± 0.0045**	0.4932 ± 0.0035
$\alpha = 0.05$	81.91	93.25	0.2303 ± 0.0140	0.3510 ± 0.0134	0.5371 ± 0.0047	0.4817 ± 0.0027

mentum 0.9. The unsupervised training lasts for 500 epochs.

Baselines. In (Zhuang et al., 2021), contrastive learning yields the s-o-t-a brain similarity among existing unsupervised task-driven methods to model VVS. In this paper, we call the method as pure contrastive learning (PCL), and employ it as a baseline for comparison with our approach. Actually, when setting the balancing weight $\alpha = 0$, our approach degenerates into PCL.

How does RPL affect task accuracy?

To explore how does RPL affect task accuracy (including IC accuracy and RPP accuracy), we first perform our designed unsupervised training for seven times with different balancing weights ($\alpha = 0$, 0.00002, 0.0001, 0.0005, 0.002, 0.01, and 0.05), and obtain seven models respectively. Then we evaluate task accuracy of the seven models, and the results are shown in Table 1. From the observation of task accuracy in the table, we can conclude the following three points:

1. Increasing the weight of RPL (from $\alpha = 0$ to $\alpha = 0.002$) not only improves RPP accuracy significantly (from 14.19 to 91.43) but also improves IC accuracy slightly (from 82.60 to 86.34). This is consistent with recent finding (Doersch & Zisserman, 2017; Pinto & Gupta, 2017) that combining multiple unsupervised tasks (CL and RPL) always leads to the improvement of the downstream performance (IC accuracy).

2. When the weight of RPL is exceedingly small ($\alpha =$ 0.00002), the obtained model has poor RPP accuracy (35.55%), but IC accuracy is still significantly enhanced compared to that with $\alpha = 0$. This demonstrates that even though the model is not well-trained on RPL, RPL can help CL to learn better image representations, hence the better downstream performance.

3. When the weight of RPL is too large ($\alpha = 0.05$), the interference between two tasks arises, causing drops in both IC accuracy and RPP accuracy.

Can RP predictivity improve brain similarity?

To investigate whether RP predictivity really improves the brain similarity, we further evaluate the brain similarity of the seven models with different balancing weights to four cortical regions (V1, V2, V4, and IT) of VVS. The results are shown in Table 1 as well.

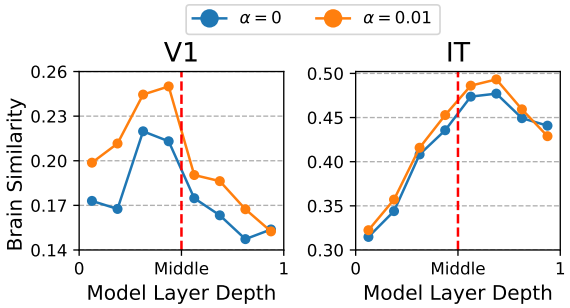

Figure 1: Layer-level Brain Similarity to V1 and IT

It is evident that once RPL is included in our designed unsupervised task, the brain similarity to four visual cortical regions will be generally higher. When setting $\alpha = 0.01$, our method significantly outperforms PCL ($\alpha = 0$), which enhances the brain similarity to V1 by 13.74% from 0.2198 to 0.2500, enhances the brain similarity to V2 by 4.57% from 0.3367 to 0.3521, enhances the brain similarity to V4 by 3.38% from 0.5236 to 0.5413, and enhances the brain similarity to IT by 3.37% from 0.4771 to 0.4932. Through the comparison, we can conclude our proposed method sets the new state-of-the-art.

Despite the above significant improvement of brain similarity, we can not directly conclude RP predictivity can improve the brain similarity from it. As we demonstrated previously, RPL also improves IC accuracy while enhancing RP predictivity. The better IC accuracy indicates the stronger ability of object recognition of the model, which may also result in higher brain similarity. Hence, the improvement of brain similarity may not be caused by the improvement of RPP accuracy, but be caused by the improvement of IC accuracy. Fortunately, when $\alpha = 0.05$, due to the relatively large α, the interference between two tasks makes current IC accuracy (81.91) lower than baseline IC accuracy (82.60). This case rules out IC accuracy improvement but current brain similarity is still comprehensively higher than the baseline to

Table 2: Layer-level Brain Similarity to V1, V2, V4 and IT

Model Layer	V1		Model Layer	V2	
	PCL	Ours		PCL	Ours
layer1.0	0.1730 ± 0.0156	**0.1987 ± 0.0094**	layer1.0	0.1945 ± 0.0197	**0.2283 ± 0.0167**
layer1.1	0.1676 ± 0.0152	**0.2116 ± 0.0088**	layer1.1	0.2248 ± 0.0230	**0.2636 ± 0.0159**
layer2.0	0.2198 ± 0.0165	**0.2445 ± 0.0128**	layer2.0	0.3130 ± 0.0162	**0.3429 ± 0.019**5
layer2.1	0.2131 ± 0.0179	**0.2500 ± 0.0113**	layer2.1	0.2990 ± 0.0164	**0.3270 ± 0.0187**
layer3.0	0.1748 ± 0.0148	**0.1904 ± 0.0117**	layer3.0	0.3249 ± 0.0148	**0.3369 ± 0.0192**
layer3.1	0.1633 ± 0.0143	**0.1863 ± 0.0141**	layer3.1	0.3367 ± 0.0143	**0.3521 ± 0.0146**
layer4.0	0.1473 ± 0.0125	**0.1674 ± 0.0116**	layer4.0	0.3028 ± 0.0131	**0.3304 ± 0.0114**
layer4.1	**0.1538 ± 0.0088**	0.1525 ± 0.0077	layer4.1	0.3096 ± 0.0160	**0.3185 ± 0.0136**
Model Layer	V4		Model Layer	IT	
	PCL	Ours		PCL	Ours
layer1.0	0.5058 ± 0.0046	**0.5180 ± 0.0056**	layer1.0	0.3148 ± 0.0023	**0.3224 ± 0.0034**
layer1.1	0.5189 ± 0.0039	**0.5338 ± 0.0046**	layer1.1	0.3440 ± 0.0021	**0.3571 ± 0.0025**
layer2.0	0.5236 ± 0.0056	**0.5293 ± 0.0042**	layer2.0	0.4082 ± 0.0030	**0.4158 ± 0.0041**
layer2.1	0.5204 ± 0.0042	**0.5413 ± 0.0045**	layer2.1	0.4356 ± 0.0037	**0.4528 ± 0.0040**
layer3.0	0.4509 ± 0.0055	**0.4600 ± 0.0066**	layer3.0	0.4737 ± 0.0031	**0.4861 ± 0.0036**
layer3.1	0.4456 ± 0.0064	**0.4586 ± 0.0062**	layer3.1	0.4771 ± 0.0029	**0.4932 ± 0.0035**
layer4.0	**0.3404 ± 0.0067**	0.3208 ± 0.0072	layer4.0	0.4493 ± 0.0024	**0.4593 ± 0.0039**
layer4.1	**0.3314 ± 0.0059**	0.2746 ± 0.0076	layer4.1	**0.4408 ± 0.0035**	0.4289 ± 0.0046

all of the four visual cortical regions. Therefore, from this case, we can conclude that RP predictivity can really improve the brain similarity.

Layer-level Brain Similarity Analysis

To explore the influence of our proposed method on brain similarity in more detail, we select the final activation layers of each residual block in our base model (ResNet-18 totally contains $4 * 2$ residual blocks) and demonstrate their layer-level brain similarity to four visual cortical regions with PCL ($\alpha = 0$) and our method ($\alpha = 0.01$) respectively. The results are shown in Table 2. Note that from "layer1.0" to "layer4.1", the layer depth increases progressively. From the observation of the results we can conclude that, compared to PCL, our method can generally enhance the brain similarity of all base model layers to all visual cortical regions, with only very few exceptions occurring in the two deepest layers.

Besides, for better understanding the variation of layer-level brain similarity as the layer depth increases in our base model, we focus on the most primary visual area V1 and the highest visual area IT among the four visual areas, and we can find that: (i) the higher brain similarity to V1 is mostly achieved in the shallow layers of our base model (the layers before the middle), while (ii) the higher brain similarity to IT is mostly achieved in the deep layers of our base model (the layers after the middle). This means our base model roughly matches VVS structurally. The shallow layers correspond to the low-level visual areas, and the deep layers correspond to the high-level visual areas. Furthermore, it is evident in the figure that, compared to PCL ($\alpha = 0$), our method ($\alpha = 0.01$) preserves good model-layer-to-brain-area correspondence while generally enhancing the brain similarity across all layers.

Conclusion and Future Work

Based on existing findings that ventral visual stream also contributes to location perception, in this paper we propose a new task-driven method to model ventral visual stream unsupervisedly, which integrates relative position learning with contrastive learning. The experimental results demonstrate that relative position predictivity can significantly improve the brain similarity, which corroborates the previous findings from a computational perspective.

Despite the outstanding brain similarity we achieved, there is still room for improvement. In our future work, we will explore some more biologically plausible base models such as spiking neural networks instead of conventional ResNet.

Ethics Statement

The biological datasets utilized in this study are either publicly accessible or obtained from published papers with the authors' consent. All data usage complies with respective data sharing and ethical guidelines. No new data involving human or animal subjects were collected for this study.

Acknowledgments

This work was supported in part by STI 2030—Major Projects (2022ZD0208903), in part by the National Natural Science Foundation of China under Grant 62336007, in part by the Starry Night Science Fund of the Zhejiang University Shanghai Institute for Advanced Study under Grant SN-ZJU-SIAS-002. Yueming Wang is the corresponding author.

References

Brincat, S. L., & Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature neuroscience*, *7*(8), 880–886.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, *15*(4), e1006897.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., ... Rust, N. C. (2005). Do we know what the early visual system does? *Journal of Neuroscience*, *25*(46), 10577–10597.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597–1607).

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), 27755.

Cloutman, L. L. (2013). Interaction between dorsal and ventral processing streams: where, when and how? *Brain and language*, *127*(2), 251–263.

Coates, A., Ng, A., & Lee, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 215–223).

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Doersch, C., & Zisserman, A. (2017). Multi-task self-supervised visual learning. In *Proceedings of the ieee international conference on computer vision* (pp. 2051–2060).

Erhan, D., Courville, A., Bengio, Y., & Vincent, P. (2010). Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 201–208).

Ericsson, L., Gouk, H., & Hospedales, T. M. (2021). How well do self-supervised models transfer? In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 5414–5423).

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature neuroscience*, *16*(7), 974–981.

Greene, J. (2005). Apraxia, agnosias, and higher visual function abnormalities. *Journal of Neurology, Neurosurgery & Psychiatry*, *76*(suppl 5), v25–v34.

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.

He, K., Girshick, R., & Dollár, P. (2019). Rethinking imagenet pre-training. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 4918–4927).

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).

Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., & Botvinick, M. (2021). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, *12*(1), 6456.

Huang, L., Ma, Z., Yu, L., Zhou, H., & Tian, Y. (2023). Deep spiking neural networks with high representation similarity model visual pathways of macaque and mouse. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 37, pp. 31–39).

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866.

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

Konen, C. S., & Kastner, S. (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nature neuroscience*, *11*(2), 224–231.

Kravitz, D. J., Kriegeskorte, N., & Baker, C. I. (2010). High-level visual object representations are constrained by position. *Cerebral Cortex*, *20*(12), 2916–2925.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, *12*(4), e1004896.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., ... others (2019). Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, *32*.

Lotter, W., Kreiman, G., & Cox, D. (2020). A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature machine intelligence*, *2*(4), 210–219.

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, *35*(39), 13402–13418.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., ... Yamins, D. L. (2018). Task-driven convolutional recurrent models of the visual system. *Advances in neural information processing systems*, *31*.

Nayebi, A., Sagastuy-Brena, J., Bear, D. M., Kar, K., Kubilius, J., Ganguli, S., ... Yamins, D. L. (2021). Goal-driven recurrent neural network models of the ventral visual stream. *bioRxiv*, 2021–02.

Newell, A., & Deng, J. (2020). How useful is self-supervised

pretraining for visual tasks? In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 7345–7354).

Pinto, L., & Gupta, A. (2017). Learning to push by grasping: Using multiple tasks for effective learning. In *2017 ieee international conference on robotics and automation (icra)* (pp. 2161–2168).

Rong, D., Yu, G., Shen, S., Fu, X., Qian, P., Chen, J., ... Wang, W. (2024). Clean-image backdoor attacks. In *International conference on artificial neural networks* (pp. 187–202).

Schiller, P. H. (1995). Effect of lesions in visual cortical area v4 on the recognition of transformed objects. *Nature*, *376*(6538), 342–344.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*. Retrieved from https://www.biorxiv.org/content/10.1101/407007v2

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*. Retrieved from https://www.cell.com/neuron/fulltext/S0896-6273(20)30605-X

Sereno, A. B., & Lehky, S. R. (2011). Population coding of visual space: comparison of spatial representations in dorsal and ventral pathways. *Frontiers in computational neuroscience*, *4*, 159.

Shi, J., Shea-Brown, E., & Buice, M. (2019). Comparison against task driven artificial neural networks reveals functional properties in mouse visual cortex. *Advances in Neural Information Processing Systems*, *32*.

Simic, N., & Rovet, J. (2017). Dorsal and ventral visual streams: Typical and atypical development. *Child Neuropsychology*, *23*(6), 678–691.

Sincich, L. C., & Horton, J. C. (2005). The circuitry of v1 and v2: integration of color, form, and motion. *Annu. Rev. Neurosci.*, *28*(1), 303–326.

Wei, J., Rong, D., Zhu, X., He, Q., & Wang, Y. (2024). Speed-enhanced subdomain alignment for long-term stable neural decoding in brain-computer interfaces. In *2024 ieee international conference on bioinformatics and biomedicine (bibm)* (pp. 3832–3837).

Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., & Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature neuroscience*, *11*(11), 1352–1360.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365.

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.

Zachariou, V., Klatzky, R., & Behrmann, M. (2014). Ventral and dorsal visual stream contributions to the perception of object shape and object location. *Journal of Cognitive Neuroscience*, *26*(1), 189–209.

Zhu, H., & Rong, D. (2024). Multiview consistent physical adversarial camouflage generation through semantic guidance. In *2024 international joint conference on neural networks (ijcnn)* (pp. 1–8).

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118.