Decoding of Dynamic Video Evoked by fMRI based on Dilated Long and Short Term Memory Neural Network

Yuting Wang

Integrated Program for Neuroscience

McGill University, Montreal, Canada

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

ACKNOWLEDGEMENTS

I sincerely thank my supervisor Reza and my co-supervisor HongmeiYan for their valuable guidance, help, care, and support throughout my postgraduate career. Without their support and help, I could not have successfully completed my academic task.

I would like to thank my fellow committee members Amir and Pouya, who provided invaluable expertise and advice throughout my research. Their guidance to me is invaluable and valuable, which enriches my research content.

I would also like to especially thank all the participants in our study. Their enthusiastic participation, curiosity, and patience drive our data collection. Thanks to Dr. Huang Wei for his efforts in data collection.

In addition, I would like to thank my friends, including YiningYang, Haohan Bai, Gangyao Gao, Yanling Huang, and others, for creating a pleasant and energetic research atmosphere.

Finally, I thank my parents for providing me with a positive family environment that makes me a relatively optimistic and open-minded person. Their unconditional support for me is the most solid backing for my study career.

CONTRIBUTIONS OF AUTHORS

The body of work presented in this thesis would not have been made possible without a close collaboration between myself and my M.Sc. supervisors, Dr. Wei Huang, Dr. Chong Wang. Listed below are the specific contributions:

Wei Huang: designed experiments and collected fMRI data of all subjects.

Chong Wang: provided suggestions on data calculation and analyses.

Abstract：

Background: Decoding visual perceptual content from Functional Magnetic Resonance Imaging (fMRI) is an important research topic in the field of brain-computer interaction. However, due to the small sample size, large noise, high data-dimension, and expensive acquisition cost of fMRI data, the performance of brain signal-based visual decoding is low. Therefore, the underlying visual neural encoding and decoding mechanism needs to be further explored. At present, there are many kinds of research on the visual perception decoding of static images, but there are few types of research on visual decoding of dynamic video, which is due to the time variation and content complexity of dynamic video. Dynamic perception and information integration are the basic forms of human understanding of the world. Studying the neural encoding and decoding mechanism of dynamic visual information enables us to better understand the working mode of the brain.

Methods: We designed three fMRI experiments, namely, retinal topography experiment, high-level visual cortex localization experiment, and dynamic video decoding experiment. In the retinal topography experiment, a mapping relationship between visual stimuli and fMRI signals from low-level visual cortex was established by viewing checkerboard stripe stimuli to the subjects. Through this mapping relationship, low-level visual regions can be delineated. In the high-level visual area localization experiment, the mapping relationship between complex visual stimuli and high-level visual cortex fMRI signals was established through a series of natural pictures. In the dynamic video decoding experiment, subjects watched

dynamic videos consisting of five categories for two hours and their fMRI signals were collected.

Results: Based on the visual information integration mechanism, a dynamic brain information classification and decoding model based on dilated convolutional long and short-term memory (DC-LSTM) is proposed, in which multi-scale temporal information is extracted from brain signals and fused by dilated convolution with different coefficients. By comparing the decoding accuracy corresponding to different dilation coefficients, we found that the decoding performance of short temporal sequence brain signals was best when the temporal integration scale was 4. Meanwhile, by comparing the classification and decoding performance of single time-point brain signals, averaged brain signals, and disturbed temporal sequence brain signals, it was found that brain information extracted by temporal information integration was more relevant to stimulus categories, which led to better decoding performance. We further compared the decoding accuracy of high and low visual cortex and found that high visual cortex has better information integration ability. Finally, we analyzed the brain's representational features for different types of videos by means of a representational dissimilarity matrix, providing evidence for the consistency of brain activity patterns with visual stimulus features.

Conclusions: Our results show that using dilated convolution can better integrate temporal information and improve decoding accuracy. For high-level visual regions, more information is accumulated. Therefore using information integration can significantly improve decoding performance. The less information accumulated in low-level visual regions, the better the decoding of brain information at a single time point.

Résumé :

Contexte: Le décodage du contenu visuel perceptif à partir de l'imagerie par résonance magnétique fonctionnelle (IRMf) est un sujet de recherche important dans le domaine de l'interaction cerveau-ordinateur. Cependant, en raison de la petite taille de l'échantillon, du bruit important, de la dimension élevée des données et du coût d'acquisition élevé des données IRMf, les performances du décodage visuel basé sur les signaux cérébraux sont faibles. Par conséquent, le mécanisme sous-jacent d'encodage et de décodage neuronal visuel doit être exploré plus avant. À l'heure actuelle, de nombreuses recherches ont été menées sur le décodage de la perception visuelle d'images statiques, mais peu sur le décodage visuel de vidéos dynamiques, en raison de la variation temporelle et de la complexité du contenu des vidéos dynamiques. La perception dynamique et l'intégration des informations sont les formes fondamentales de la compréhension du monde par l'homme. L'étude du mécanisme d'encodage et de décodage neuronal des informations visuelles dynamiques nous permet de mieux comprendre le mode de fonctionnement du cerveau.

Méthodes: Nous avons conçu trois expériences d'IRMf, à savoir l'expérience de topographie rétinienne, l'expérience de localisation du cortex visuel de haut niveau et l'expérience de décodage vidéo dynamique. Dans l'expérience de topographie rétinienne, une relation de correspondance entre les stimuli visuels et les signaux IRMf du cortex visuel de bas niveau a été établie en montrant aux sujets des stimuli en forme de damier. Cette relation de correspondance permet de délimiter les régions visuelles de bas niveau. Dans l'expérience de localisation des zones visuelles de haut niveau, la relation de correspondance entre les stimuli visuels complexes et les signaux IRMf du cortex visuel de haut niveau a été établie à l'aide d'une série d'images naturelles. Dans

l'expérience de décodage vidéo dynamique, les sujets ont regardé des vidéos dynamiques composées de cinq catégories pendant deux heures et leurs signaux IRMf ont été enregistrés.

Résultats: Basé sur le mécanisme d'intégration des informations visuelles, un modèle dynamique de classification et de décodage des informations cérébrales basé sur la mémoire longue et courte convolutionnelle dilatée (DC-LSTM) est proposé, dans lequel les informations temporelles multi-échelles sont extraites des signaux cérébraux et fusionnées par convolution dilatée avec différents coefficients. En comparant la précision de décodage correspondant à différents coefficients de dilatation, nous avons constaté que les performances de décodage des signaux cérébraux à séquence temporelle courte étaient meilleures lorsque l'échelle d'intégration temporelle était de 4. Parallèlement, en comparant les performances de classification et de décodage des signaux cérébraux à point temporel unique, des signaux cérébraux moyennés et des signaux cérébraux à séquence temporelle perturbée, nous avons constaté que les informations cérébrales extraites par l'intégration des informations temporelles étaient plus pertinentes pour les catégories de stimulus, ce qui a permis d'améliorer les performances de décodage. Nous avons également comparé la précision de décodage du cortex visuel haut et bas et constaté que le cortex visuel haut avait une meilleure capacité d'intégration de l'information. Enfin, nous avons analysé les caractéristiques représentationnelles du cerveau pour différents types de vidéos au moyen d'une matrice de dissimilarité représentationnelle, ce qui prouve la cohérence des schémas d'activité cérébrale avec les caractéristiques du stimulus visuel.

Conclusions: Nos résultats montrent que l'information temporelle peut être mieux intégrée et la précision de décodage peut être améliorée en utilisant la convolution dilatée. Pour les zones visuelles avancées, plus d'informations sont accumulées, et l'utilisation de l'intégration de l'information peut améliorer considérablement les performances de décodage. Moins l'

information s’accumule dans les zones visuelles de bas niveau, mieux il est possible de décoder l’information cérébrale à un moment donné.

Table of Contents

1. Introduction

In neuroscience, visual decoding plays a crucial role in understanding how the brain processes visual information. By decoding brain activity and translating it into images or videos(Kay, Naselaris et al. 2008, Miyawaki, Uchida et al. 2008, Nishimoto, Vu et al. 2011), researchers can understand the activity patterns of different neurons in response to various visual stimuli. This enables them to infer the underlying mechanisms of visual information processing in the brain. In essence, by analyzing patterns of human brain activity, researchers can determine the type of visual information being perceived, identify the multi-level semantic content encoded in the brain, and even reconstruct visual images based on the brain activity.

With the advancement of brain imaging techniques in medical applications and the rapid development of artificial intelligence algorithms, scientists have begun to explore brain decoding. However, the inherent complexity of this task imposes certain limitations on current research. For example, the performance of visual classification decoding deteriorates when confronted with scenes containing complex backgrounds or multiple objects(Haufe, Meinecke et al. 2014, Naseer and Hong 2015). Another significant limitation lies in the inability of established mapping models to effectively handle intricate temporal information(Wen, Shi et al. 2018, Wen, Shi et al. 2018).

Visual decoding provides insights into how visual information is processed in the brain. By investigating visual classification decoding, we can enhance our understanding of the response patterns of distinct brain regions to various information types(Nishimoto, Vu et al. 2011), as well as the interactions among different neurons. This knowledge is pivotal for exploring deeper into the mechanisms underlying visual processing in the brain. Moreover, visual decoding research plays a crucial role in the development of advanced brain-computer interface (BCI) technologies, which enable individuals to control

computers using their neural activity. By incorporating temporal information into real-time decoding algorithms, the advancement of BCI technologies is further enhanced (Wang, Collinger et al., 2013). This has significant implications for individuals with specific needs, particularly those with disabilities, as it offers substantial benefits in terms of improved communication and control capabilities.

The classification of fMRI brain activity patterns induced by visual stimuli is an effective approach for decoding the current cognitive state of the human brain and analyzing its working mechanisms(Gerstner, Kreiter et al. 1997). While numerous studies have been conducted on the visual perception decoding of static images such as static pictures and stripe orientation, limited research have been conducted on the visual perception decoding of dynamic videos. This is due to the challenges posed by the temporal variability and content complexity of dynamic videos, as well as the low signal-to-noise ratio of fMRI, which makes brain decoding research on dynamic video perception difficult due to the delay characteristics of the BOLD signal.

Despite these challenges, dynamic video stimuli have unparalleled advantages over static pictures as visual stimuli. Firstly, the complex daily life scenes processed by human eyes cannot be replicated by static pictures. Secondly, to model the visual function of the brain, it is necessary to rely on the processing of real-time complex visual stimuli, which requires more accurate extraction of temporal information. The temporal information provided by static pictures is overly simplistic, whereas dynamic videos provide more complex temporal information. Therefore, using dynamic videos as visual stimuli is highly significant.

Further exploration of the visual processing mechanism of dynamic visual information and the interpretation of the brain's information integration process may provide experimental evidence and methodological support for complete brain decoding in the future.

1.1 The transition from static visual stimuli to dynamic visual stimuli

In the early stages of visual decoding research, static visual stimuli, such as orientation(Haynes and Rees 2005), natural images(Kay, Naselaris et al. 2008), and pictures from dreams(Horikawa, Tamaki et al. 2013) and imagination(Hassabis, Spreng et al. 2014, Horikawa and Kamitani 2017), were predominantly used. Among them, Haynes and Rees(Haynes and Rees 2005) conducted a pioneering study in decoding the direction of invisible stimuli from the human visual cortex by monitoring subjects' brain activity, thus demonstrating the feasibility of extracting information from brain activity. However, the use of simple orientation information in their study fails to capture the complexity of visual scenes processed by the human eye. As a result, researchers have tried to replace simple orientation information with natural images. Kay et al. (Haynes and Rees 2005) successfully decoded natural images from human brain activities, albeit with a relatively small sample size, which may limit the generalizability and reliability of the decoding model. This indicates that both basic orientation information and more complex nature images can be decoded from brain activity.

Building upon this foundation, researchers have progressively increased the complexity of visual stimuli and ventured into the domain of decoding dreams and imaginary pictures. Horikawa et al. (Horikawa, Tamaki et al. 2013) achieved successful decoding of dreams using support vector machines (SVM), suggesting that specific visual experiences during sleep are represented by patterns of brain activity shared with stimulus perception. This approach offers a means of objectively deciphering the subjective content of dreams using neuroscientific measures. However, due to the inherent challenges associated with collecting dream data and the limited amount of available data, the overall decoding performance remains unstable. Subsequently, in 2017, Horikawa and Kamitani(Horikawa and Kamitani 2017) proposed a universal decoding method that leverages hierarchical visual features to identify objects observed and imagined by subjects,

significantly expanding the dataset used for decoding.

In light of these advancements, it is evident that researchers have progressively elevated the complexity of visual stimuli and developed decoding models to explore the visual mechanisms of the human eye. In recent years, there has been a gradual shift toward employing dynamic videos(Wen, Shi et al. 2018) as visual stimuli. They successfully applied deep learning models to neural encode and decode tasks for dynamic natural vision. The primary challenge in this approach lies in the complexity of content and temporal variations associated with dynamic visual stimulation. This is because real-world events are rarely presented in a sequential and isolated manner(Spiers and Maguire 2007), making it difficult to distinguish neural activity specific to a particular event from the continuous flow of complex stimuli. Due to the complexity and diversity of the real world, the decoding process faces many challenges, including the complexity of data analysis and modeling, the existence of s differences, and the complex relationship between brain activity and behavior. Therefore, there are fewer studies on dynamic video-induced brain activity decoding and fewer related datasets. But decoding the activity of the human brain in real-world experiences can improve understanding of cognition, emotions, decision-making, and behavior. Therefore, it is necessary to establish a dataset of brain activity induced by dynamic video and establish a decoding model.

1.2 Information integration can improve decoding performance

At present, most decoding models use static models to extract visual features and establish a mapping relationship between brain activity and visual stimuli. For example, Carlson et al. (Carlson, Schrater et al. 2003) used linear discriminant to analyze fMRI data induced by chairs, faces, houses, etc. The results showed that the patterns of brain activity in one class of objects and others were largely independent of each other. Kamitani et al. used a linear support vector machine to decode orientation (Kamitani and Tong 2005), which reliably predicted which of the eight stimulus directions observed in individual

experiments. In subsequent research and exploration, the team proposed a method based on Bayesian statistical analysis to reconstruct natural images from FMRI brain activity (Naselaris, Prenger et al. 2009). The team could even use linear support vector machines to decode the brain's activity when it is visually stimulated and speculate what type of dream these activities correspond to (Horikawa, Tamaki et al. 2013). This means that it is possible to understand what people are dreaming about by looking at brain activity. However, the decoder built with a simple static model cannot model the visual system of the human eye well, resulting in low decoding performance. This is because static models cannot capture dynamic changes in time and the evolution of the system. On the other hand, the human eye processes a variety of events daily, each of which encompasses a substantial amount of temporal information. Therefore, it is necessary to use dynamic models to build decoders to make full use of temporal information.

Due to the complexity and temporal variability of dynamic video content, there are currently fewer classification decoding of fMRI brain signals induced by dynamic video. In 2017, the team(Wen, Shi et al. 2018) proposed an fMRI dataset containing 15 categories of dynamic video-induced. Subsequently, the team used convolutional neural networks to extract spatial features, while also using multiple recurrent neural networks (RNNs) to fully learn temporal information, simulate hierarchical and distributed models of process memory, and further process spatiotemporal features. The model also reveals the cortical hierarchy of the temporal receptive window, which they believe has a shorter time-receptive window(Hasson, Yang et al. 2008) in the primary visual cortex and vice versa in the higher visual cortex. This indicates that the accumulation of information in the visual cortex increases gradually over time.

The vision system fuses and integrates visual information from different perceptual channels to produce a consistent, comprehensive visual perception. Real-world events

occur not only in the extended area of space but also in the extended time. Therefore, the temporal response characteristics of different brain regions should also exist in a hierarchy similar to the size of the spatial receptive field(Hasson, Yang et al. 2008). Human perception of the world unfolds over time, so it must rely on long-term information accumulation, including causal reasoning, processing linguistic information at various scales, understanding narratives, event segmentation, and human-social interaction. In most real-life processes, past information is often used to process incoming information across multiple time scales(Hasson, Chen et al. 2015), and sensation and perception use information integration to perceive the external world(Beauchamp 2005). In 2001, Rotshtein et al.(Rotshtein, Malach et al. 2001) found that the early visual area has less information accumulation, while the higher visual area has more information accumulation, so the use of information integration of the advanced visual region can better decode visual information. In most real life, information from the past is often used to process information from multiple scales. Therefore, the use of information integration can improve the decoding performance.

2. Methods and results

2.1. Rational

Due to the complex content and temporal variability of human life scenes, static images are insufficient in simulating naturalistic visual experiences, making it imperative to use dynamic video stimuli in fMRI research. However, the high cost of fMRI data acquisition coupled with a relatively low signal-to-noise ratio necessitates the need for subjects to view stimuli multiple times to obtain high-quality datasets. Therefore, it is essential to establish fMRI models for decoding visual perception evoked by dynamic video stimuli to better understand the processing mechanisms of the visual system. However, current visual perception decoding models mostly focus on static images, and

there is a lack of dynamic video-evoked fMRI datasets, posing a significant challenge in this area. A critical difficulty in dynamic visual decoding is the separation of neural activity associated with a specific event from a continuous stream of complex stimuli. To address this, we conducted behavioral experiments using five categories of dynamic videos and established a new fMRI dataset evoked by dynamic videos.

In recent years, researchers have released several datasets on brain activity patterns evoked by dynamic videos, which have advanced the development of visual perception decoding. In 2012, Hu(Hu, Li et al. 2011) randomly selected 51 photos from the sports, weather, and business categories in the TRECVID 2005(Amir, Argillander et al. 2003) dataset (20 sports, 19 weather, and 12 business) and divided them into eight sub-videos, each about 11 minutes long. These sub-videos were presented to four subjects and fMRI brain imaging data were collected. This dataset contained only three categories and was relatively small in size. Barch et al.(Barch, Burgess et al. 2013) collected the HCP-YA (Human Connectome Project) fMRI dataset in 2013, which included a large sample of young healthy adults and seven functional tasks (emotion, gambling, language, motor, relational, social, and working memory), each under different conditions. These seven tasks provided good coverage of brain activation as a whole, so classifiers trained on this dataset would help decode brain states across a wide range of functional tasks. This dataset is very large and has gradually expanded over time. However, the visual stimuli in this dataset are various functional tasks, which are somewhat different from the complex life scenes we envisage. Subsequently, Wen et al.(Wen, Shi et al. 2018) proposed a dynamic video-evoked fMRI dataset consisting of 15 categories (indoor, outdoor, people, faces, birds, insects, aquatic animals, land animals, flowers, fruits, natural scenes, cars, planes, boats, and sports), which is one of the largest datasets to date evoked by dynamic videos in fMRI. They collected data from three healthy volunteers who watched natural color video clips. The training video, which was 2.4 hours in length, contained 276 video clips and was

divided into 18 8-minute sub-videos. The testing video was an 8-minute movie containing 38 different video clips. Each subject watched the training movie twice and the testing movie ten times. The order of each category sub-clip in the movie was random. However, this dataset lacks specific sub-video category labels and is unbalanced in terms of the number of sub-categories, with only one sub-video for insects.

Analysis of existing datasets reveals that the number of datasets related to fMRI evoked by dynamic video is very small compared to those evoked by other static stimuli, and some of the dataset categories are severely imbalanced. Given the limitations of existing datasets, we designed relevant visual stimulus experiments and established a dynamic video-evoked fMRI dataset with five subjects, enriching the dataset for decoding brain visual perception and advancing the development of dynamic visual perception decoding.

2.2. Experimental approach

2.2.1 Participants

Five healthy participants took part in the experiment. All participants had a normal or corrected-to-normal vision. Before the experiment, each participant provided written informed consent. The experimental protocol was approved by the Institutional Review Board of the Institute of Biophysics, Chinese Academy of Sciences. The experiment was programmed using E-prime software. Visual stimuli were presented using a liquid crystal projector on a screen placed inside the scanner bore. During the experiment, participants were instructed to focus on the center of the screen and to refrain from moving their bodies.

2.2.2 Visual stimuli experimental design

The experiment consisted of three stages: (1) a retinotopic mapping experiment using bar stimuli, as shown in Figure 1-1 (A); (2) a functional localizer experiment of the high-level visual cortex, as shown in Figure 1-1 (B); and (3) a dynamic video experiment,

as shown in Figure 1-1 (C). The video stimuli included five categories: animals, humans, flowers, transportation, and buildings. Transportation included three subcategories: cars, boats, and planes. Animals included four subcategories: dogs, tigers, lions, and pandas. The image stimuli were dynamically presented at a resolution of 240*240. The dynamic visual stimuli used in this study can be found on Youku and iQiyi websites.

The boundaries of the brain's visual areas can be effectively localized using the stripe-based retinotopic mapping experiment to describe the low-level visual cortex. Subjects viewed black and white checkerboard stimuli with 100% contrast with a spatial length of 20° and a spatial width of 20°, and a temporal frequency of 10 Hz. The stripe stimuli had four types and moved in two different directions, resulting in a total of 8 patterns. Each stimulus stripe flickered at a position for 2 seconds and then flickered and moved in one direction for a total of 22 positions. The experiment included 12 seconds of rest time before and after, totaling 376 seconds. To improve performance, the experiment was repeated four times.

In the high-level visual cortex functional localization experiment, natural images taken from videos were used to locate the areas, consisting of two sessions. Each session had five parts, and each part contained eight images of a category, with each image flickering for 0.2 seconds followed by 1.3 seconds of rest time. There was a 12-second rest time before and after each part and at the beginning and end of the experiment. Each part lasted 10.7 seconds, and the total time was 239 seconds.

In the dynamic visual decoding process, a total of 2 hours of natural dynamic videos were shown to the subjects. To ensure sufficient rest time and concentration, the long videos were divided into 12 sub-videos of 10 minutes each. Each sub-video contained 60 10-second videos of 5 categories. There was a 12-second rest time before and after each sub-video. The experiment lasted for a total of 624 seconds. The above experiments were conducted over the course of two weeks.

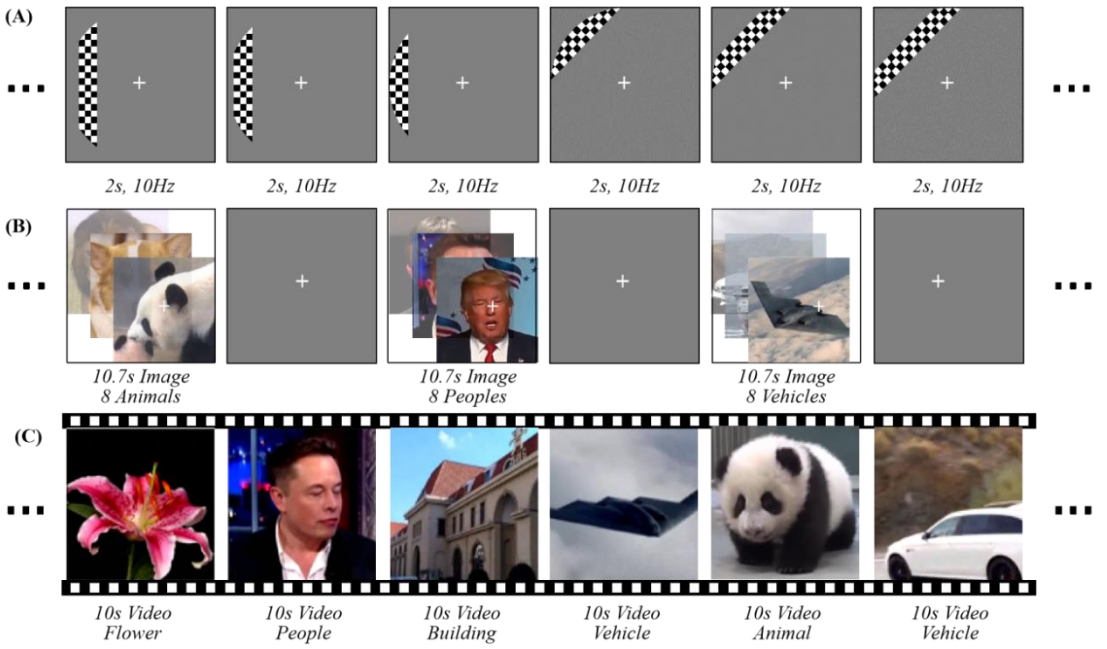

Figure 1 Experimental flow of visual stimulation. A. Bar retinal topology mapping visual stimulation; B. advanced visual cortex function positioning visual stimulation; C. visual stimulation of dynamic video

2.3. Methods

2.3.1. The population receptive field(pRF) model locates the primary visual region

In the strip retinal topological mapping experiment, the population receptive field model(Dumoulin and Wandell 2008) was used to locate the lower visual cortex. The model fits neurons in each cortex, represents the receptive field of each neuron by a two-dimensional Gaussian function, and estimates pRF parameters from time series data and fMRI responses using a linear spatiotemporal model of fMRI responses. Assuming that there is a linear relationship between blood oxygen levels and MR Signals, it can be described as:

$$y(t) = p(t)\,\beta + e \tag{2-1}$$

Where p(t) is the predicted BOLD signal, beta is the scale factor that interprets the fMRI signal, and e is the measurement noise. In neuroimaging, GLM is used and the predicted BOLD signal is entered into a design matrix. The predicted response p(t) is obtained by using a Gaussian model of the neuron population. A two-dimensional Gaussian population receptive field model is defined as follows：

$$g(x,y) = e^{-\frac{(x-x_0)^2+(y-y_0)^2}{2\sigma^2}} \quad (2\text{-}2)$$

Where $(x_0, y_0)$is the center and σ is the Gaussian distribution (standard deviation). In the bar retinal topological mapping experiment, the visual stimulus is defined as $s(x,y,t)$. For a given pRF model and effective stimulus, the predicted response can be calculated. Since the pRF and the effective stimulus formula are defined in common units of visual space, the first step in predicting the fMRI time series is to calculate the overlap between the effective stimulus and the model pRF on the voxels. The pRF response (r_t) of a single voxel is defined as follows:

$$r_t = \sum_{x,y} s(x,y,t)g(x,y) \quad (2\text{-}3)$$

Then, by convolving (r_t) with a hemodynamic response function model (HRF, h(t)), time series predictions are obtained.

$$p_t = h(t) * r(t) \quad (2\text{-}4)$$

The goodness of fit is estimated by calculating the residual sum of squares (RSS) between the predicted value p(t) and the data y(t).

$$RSS = \sum_t (y(t) - p(t)\beta)^2 \quad (2\text{-}5)$$

The optimal pRF parameter minimizes RSS by searching from coarse to fine.

2.3.2. Face selectivity and scene selectivity

Back in 1997, Kanwisher et al. (Kahn, Pace-Schott et al. 1997)found that the brain's fusiform gyrus region was significantly more active when subjects saw visual stimuli such as faces than when they saw various common objects. This facial activation was used to define a specific area of interest for each subject individually. In 1998, the team further

identified three scene-selection regions in the human cortex: PPA(Epstein and Kanwisher 1998), RSC(Maguire 2010), and OPA(Dilks, Julian et al. 2013). In face processing, OFA responds strongly to the face parts (i.e. eyes, nose, mouth) regardless of their spatial arrangement, while FFA represents the face parts and the typical spatial arrangement of these parts (e.g., two eyes above the nose)(Yovel and Kanwisher 2004). The visual areas associated with scenes and the brain areas associated with faces have similar functional divisions. For example, OPA deals with scenes at the local element level, while the preceding PPA and RSC represent the overall properties of the scene. Therefore, by using the selective features of the brain regions related to "scene" and the brain regions related to "face", we can determine which brain regions have activity related to scene perception through statistical significance analysis of comparative experimental conditions or comparative stimulus conditions. Based on GLM, the contrast activation map in each scene was obtained, and the advanced visual cortex function was located for each subject.

2.3.3. Classification and decoding model of brain information based on dilated long and short term memory neural network

2.3.3.1. Dilated convolution

Dilated convolution was first proposed for time series data in 2016 by the Google team(Oord, Dieleman et al. 2016) in the WaveNet paper. The main component of WaveNet is causal convolution(Oord, Kalchbrenner et al. 2016), which ensures that the model cannot violate the order of the data being modeled: predictions made by the model at a given time step cannot depend on any future time steps. During training, because the input time steps are known, conditional predictions for all time steps can be parallelized. In models that use causal convolution, predictions are made sequentially, with each sample being predicted and then fed back into the network to predict the next sample. Because these models have no recurrent connections, they are typically easier to train than recurrent

neural networks, especially when applied to very long sequences. However, one issue with causal convolution is that it requires many layers or larger filters to increase the receptive field, which can lead to a high computational burden. Therefore, the team used dilated convolution to increase the receptive field without increasing the computational cost.

Dilated convolution applies a filter to a larger area than its length by skipping input values. It is equivalent to convolving an unfolded larger filter with the original filter replaced by zeros but with significantly higher efficiency. Dilated convolution can operate on a coarser scale than normal convolution. This is similar to pooling or hierarchical convolution, but the output size is the same as the input size. When the dilation rate is 1, it is equivalent to standard convolution. Figure 2 shows the process of dilated causal convolution with a dilation rate of 2 or 4 and 3 convolutional kernels. When the dilation rate is 2, the calculation of the convolution skips the middle value, which is determined by the previous time step and the current time step. This means that information is first integrated over a small time scale, and then over a larger time scale as the dilation rate increases. By selecting an appropriate time scale and using dilated convolution, relevant temporal information can be extracted and integrated.

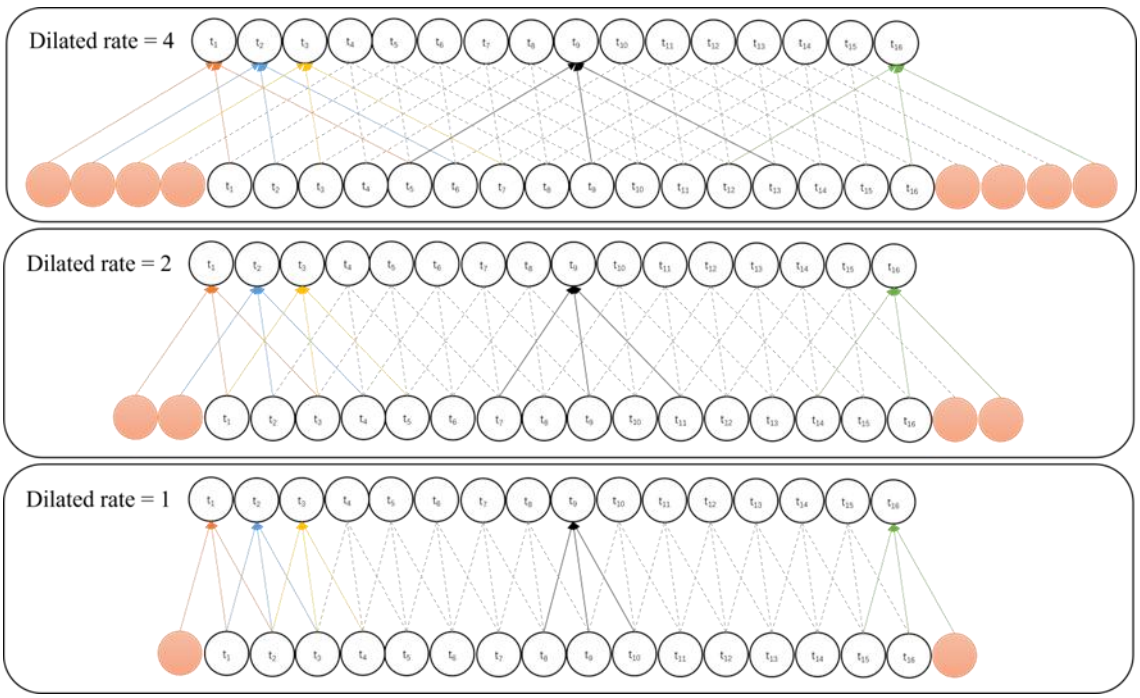

Figure 2 Visualizes the causal convolution layers with expansion coefficients of 1,2 and 4. Where the number of the convolution kernels is 3

2.3.3.2. The architecture of the dilated long and short term memory neural network

Based on the integration mechanism of visual cortex information, we have developed a new model structure for decoding dynamic video brain information, which is different from traditional decoding models that decode static stimuli. By using dilated convolution and information integration, we have established a dynamic video brain information classification decoding model, as shown in Figure 3. On the one hand, we collected the brain signals of the temporal lobe of subjects watching dynamic videos through corresponding experimental designs. After data preprocessing and feature selection, we obtained the brain responses for each visual stimulus. On the other hand, we used dilated convolution with different coefficients to extract multiscale temporal information, integrated it, and then used a long short-term memory (LSTM)(Sundermeyer, Schlüter et al. 2012) network to learn the temporal relationships between time series. LSTM can handle dynamic changes between time series well, and using LSTM can fully learn the time relationship between fMRI brain signals. Therefore, applying a dilated LSTM neural network can effectively integrate multiscale temporal information, thereby capturing the features of subjects watching dynamic videos by processing the time information between brain responses. Finally, the decoding performance of the categories is obtained.

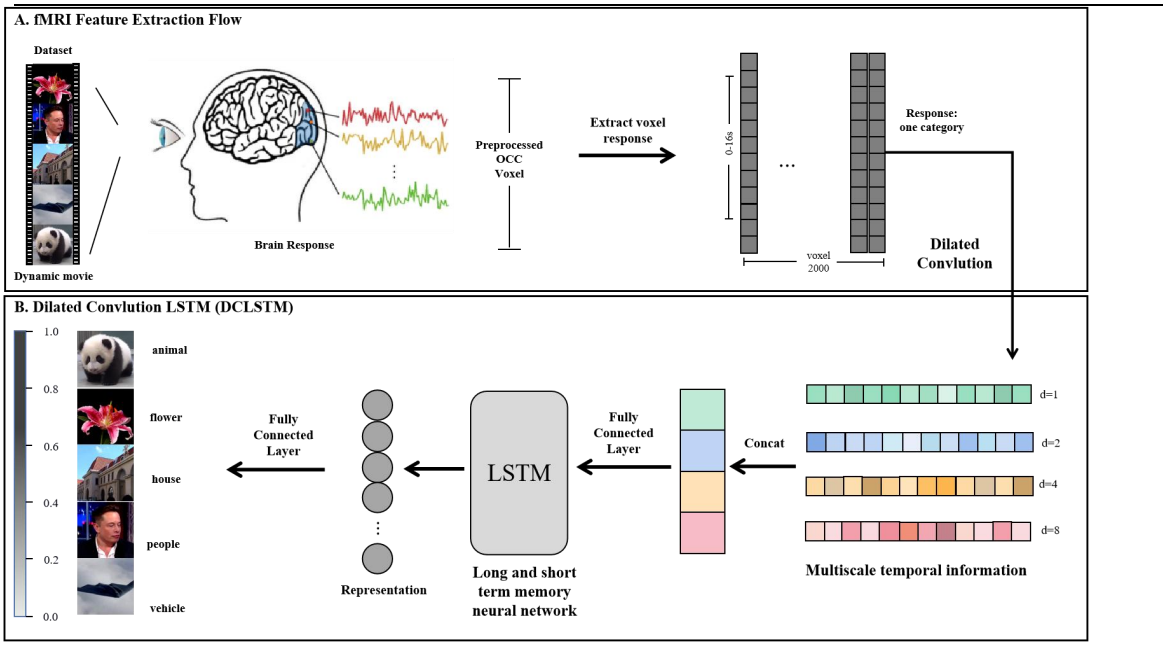

Figure 3 Model model of dynamic video brain information based on dilated long and short-term memory neural network. Different colors represent brain signals extracted from different dilated coefficients. "d=1" represents the brain signal extracted when the dilated coefficient is 1.

In the fMRI feature extraction process, the brain activity patterns of subjects watching dynamic videos are first collected, and then the response of the temporal lobe is extracted using a general linear model (GLM)(Goebel, Esposito et al. 2006). Preprocessing with time and motion correction is performed before feature selection. Since each subject's temporal lobe response contains more than 20,000 voxels, feature selection is necessary to reduce computational costs. The top 2,000 voxels are selected as visual representations. Considering the delay of the bold response and the length of the dynamic sub-videos (10 seconds), an additional 6 seconds are added to obtain a 16*2000 dimensional representation of the categories. Considering that the BOLD's peak response was around 6 seconds, we extended the time series appropriately. Secondly, because the decoding accuracy began to decline at about 14 seconds, it was not extended.

The Softmax function(Sundermeyer, Schlüter et al. 2012) is commonly used as an activation function for multi-classification problems. It compresses the data in the range [0,1], so it allows answering classification questions with probability. The formula is

defined below. Where Z_i is the output value of the i-th node and C is the number of output nodes, that is, the number of categories of classification.

$$\text{Softmax}(Z_i)=\frac{e^{Z_i}}{\sum_{c=1}^{C} e^{z_c}} \tag{2-6}$$

In the dilated LSTM neural network, since the time series we obtained is 16 seconds, which belongs to the medium and short type, excessively large dilation coefficients cannot extract appropriate temporal information and may introduce unrelated information. Therefore, dilation coefficients of 1, 2, 4, and 8 are set to extract multiscale temporal information. Human memory is not limited to local storage but spans the entire brain on multiple time scales. Therefore, information integration in the visual cortex is more suitable for brain decoding. After fusing the multiscale information, it is inputted into the LSTM network to learn the temporal relationships between time series. Finally, the decoding accuracy of each category is obtained through softmax.

2.4. Results

2.4.1. Primary visual cortex

After preprocessing fMRI data from the retinal experiments, retinotopic analysis was performed using SamSrf (Schwarzkopf, de Haas et al. 2018). The occipital lobe was defined as the entire visual cortex by SamSrf. As shown in Figure 4, the visual cortical boundaries were depicted on a sphere to visualize and label the visual areas of interest, with eccentricity and polar maps projected onto the sphere. V1 was found to lie quite precisely within the calcarine sulcus, extending from the green stripe of the wedge (dorsal talus) through the blue stripe deep in the talus to the red stripe of the lingual gyrus (ventral). V2 and V3 were separated into two quadrant field plots, one in the ventral cortex and one in the dorsal cortex (Burkhalter, Felleman et al. 1986). V2d extended from the green stripe at the V1 boundary to the middle of the blue stripe, while V3d followed from the blue stripe to the next green stripe. On the other hand, V2v extended from the red stripe at the V1 border to the blue stripe, and V3v extended from the blue stripe to the next red stripe.

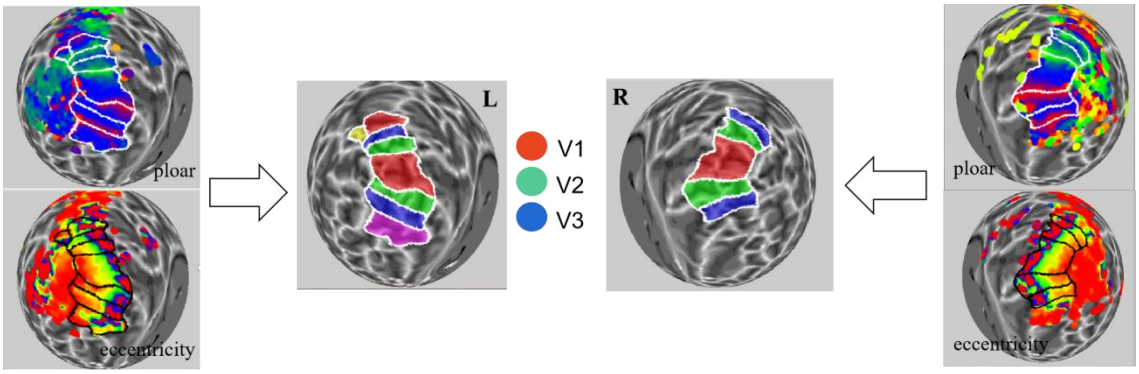

Figure 4 Primary visual cortex of subject 1.

2.4.2. High visual cortex

Functional picture mapping experiments identified FFA, OFA(Tsantani, Kriegeskorte et al. 2021), OPA, and PPA(Epstein and Kanwisher 1998) for each subject. Experimental data from the localizer were analyzed using SPM 12. Voxels with clearly significant responses to faces and scenes (two-sided t-test, uncorrected P <0.05 or 0.01) were identified by

screening for voxel clusters with high significance and defined as OFA, FFA, OPA, and PPA, respectively. There are approximately 30,000 voxels throughout the occipital cortex, but not all of them encode visual stimuli. Therefore, a rough voxel selection within the VC was performed first. An F-score feature selection algorithm was used to calculate the F value of each voxel (Chen & Lin, 2006; Huang et al., 201). The higher the F-value of the voxel, the better the ability to discriminate visual perceptual category. For high-level visual areas, the number of voxels acquired was different for each subject. To ensure the consistency of the data dimensions, the same latitude was obtained using the F-score as well. Figure 5 below delineates the inflated cerebral cortex's OFA, FFA, OPA, and PPA regions.

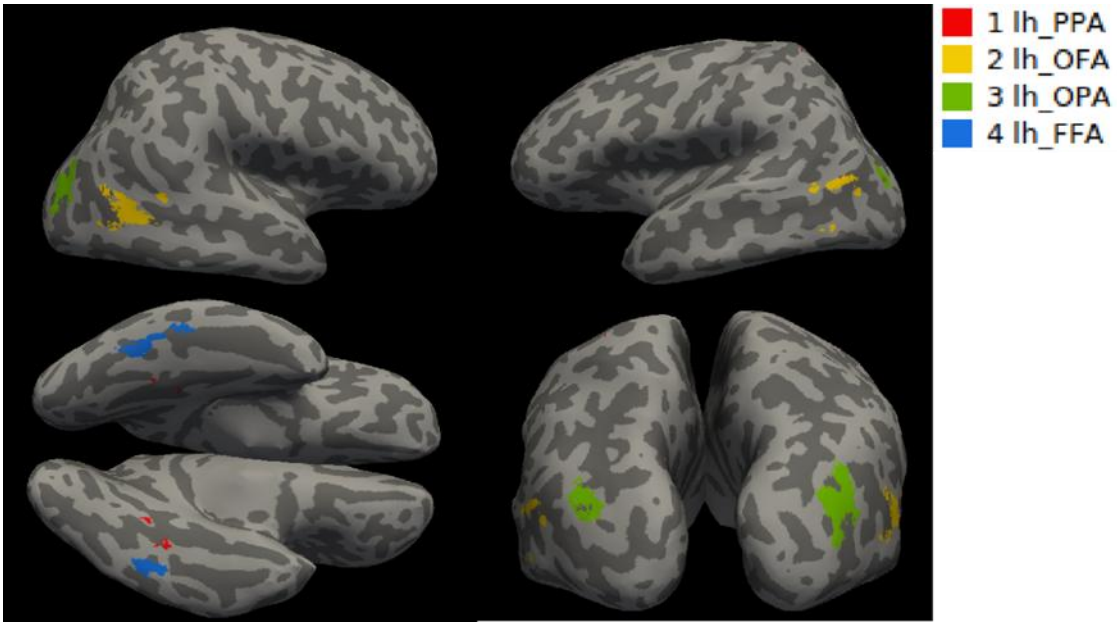

Figure 5 Subject 1's high visual cortex: OPA(occipital place area), PPA(parahippocampal place area), FFA(fusiform face area), OFA(occipital face area)

The following table shows the number of voxels in each subject's visual cortex. The "total_selected" is the number of voxels after combining the results of the five subjects and undergoing feature extraction in Table 1. Given the few voxels in higher visual areas, combine advanced brain areas: OPA and PPA into scene-related brain regions, and FFA and OFA into face-related brain regions.

Table 1 Number of voxels in each visual area per subject.

Visual cortex	sub01	sub02	sub03	sub04	sub05	total_selected
OCC	26350	26573	24406	29713	25232	2000
V1	1217	1167	1138	1452	1102	1000
V2	1118	1018	947	1291	1031	900
V3	1061	973	891	1049	773	700
FFA	700	380	608	253	255	800
OFA	402	427	486	759	666	
PPA	162	319	289	340	250	400
OPA	314	236	199	345	152	

2.4.3. The fMRI-dataset evoked by five classes of dynamic videos

We collected the brain activity patterns of five subjects while they watched dynamic videos of five categories for a total of 2 hours. The 2-hour videos were divided into 12 sub-videos of 10 minutes each. We performed 6-fold cross-validation based on the number of videos. The data details of each subject for each fold are shown in Table 2. For example, the training set for the first fold was from the 1st to the 8th dynamic videos, the validation set was from the 9th to the 10th dynamic videos, and the test set was from the 11th to the 12th videos. Here, we merged the data of the five subjects, so there were a total of 8 videos in the training set, 2 videos in the validation set, and 2 videos in the test set. Each video had 60 different visual stimuli. Therefore, there were a total of 5860=2400 samples in the training set.

Table 2 6-fold cross-validation data details

Fold	Training set	Dimensionality	Validation set	Dimensionality	Testing set	Dimensionality
k=1	1-8	5*8*60= 2400	9-10	5*2*60=600	11-12	5*2*60=600
k=2	3-10		11-12		1-2	
k=3	5-12		1-2		3-4	
k=4	7-12,1-2		3-4		5-6	
k=5	9-12,1-4		5-6		7-8	

k=6	11-12,1-6	7-8	9-10

2.4.4. Effect of the temporal integration scale on the decoding performance

Humans perceive the world through vision, which unfolds over time and requires the accumulation of information over a long period to support causal reasoning, processing of linguistic information across different scales, understanding narrative development, and various forms of social interaction. Therefore, selecting appropriate time scales is crucial for learning and integrating relevant information, as most events we encounter in our lives involve the processing of information arriving at multiple time scales. In this study, we established a decoding model using dilated convolution and extracted brain signals at different time scales by varying the dilation factors. To demonstrate the differential decoding performance of different time scales, we compared the decoding accuracy of dilated convolutions with different coefficients in extracting multi-scale temporal information and integrating them. Figure 6 shows the decoding accuracy in the OCC using different time scales. As the length of the brain signals we obtained was 16 seconds, which is considered a medium-short sequence, the decoding performance was best when the time integration scale was set to 4, as shown by the green curve in the figure. Visual information perceived by the retina is transmitted to the primary visual cortex and gradually to the higher visual cortex, which accumulates corresponding information. We observed that decoding performance gradually improved with increasing time intervals, reaching its peak at approximately 12 seconds. This is consistent with the hemodynamic response function of the BOLD signal, which has a delay of about 6 seconds. Separating neural activity related to specific events from continuous, complex stimulus streams is a major challenge in video decoding and can lead to a decrease in performance due to the introduction of responses from other categories.

In conclusion, the fusion of information at different time scales has different effects

on decoding performance, and selecting appropriate time integration scales can significantly improve decoding accuracy. Furthermore, visual information accumulates in the brain over time, and selecting the appropriate time interval can also improve decoding accuracy. Due to the complexity of dynamic video content, extracting multi-scale temporal information and integrating it can learn useful information from noisy brain signals, further improving decoding performance.

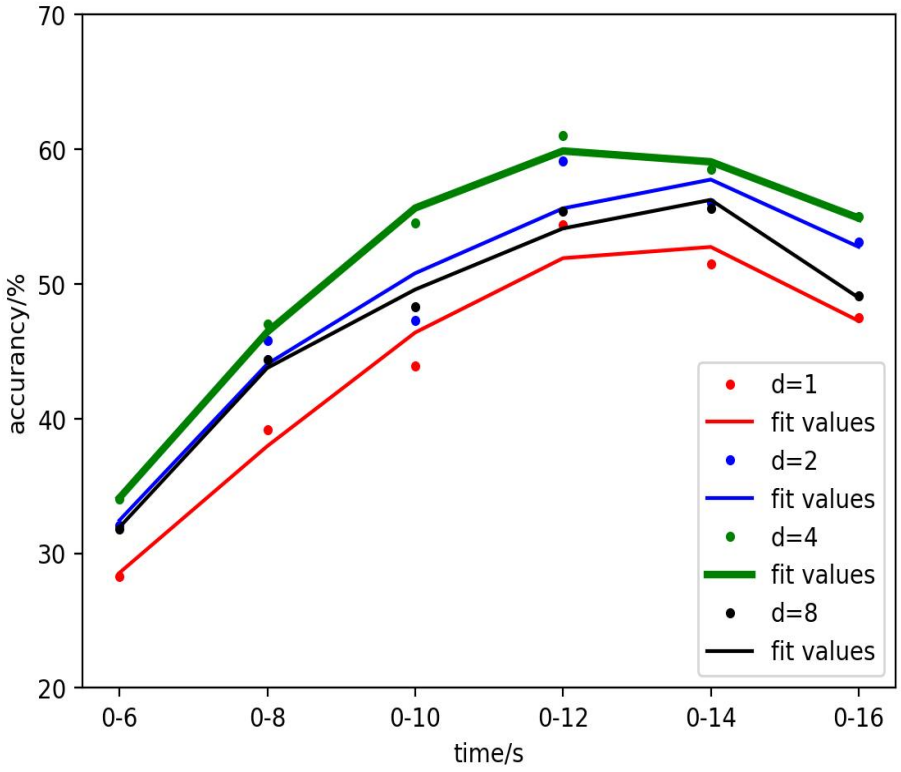

Figure 6 Decoding accuracy of different temporal integration scales in the occipital lobe. Different d represent different scales of temporal integration, and the temporal information of d=1,2, and four integrate when d=4. The horizontal axis is for different periods, and the vertical axis(20) is the decoding accuracy in%.

2.4.5. Comparison of decoding performance between DCLSTM and traditional methods

To demonstrate the decoding performance of the proposed multi-scale information integration decoding model, we compared it with five traditional machine learning models (AdaBoost, Bayes, KN, RF, and SVM). These models take the entire 16-second brain activity pattern as input, and they were implemented directly using the

scikit-learn(Pedregosa, Varoquaux et al. 2011) package.

Figure 7 shows the performance comparison of various decoding models with our proposed decoding model (DCLSTM). All results were obtained through 6-fold cross-validation. The accuracy of the DCLSTM decoding model in the entire visual area (OCC) was 56.4% (chance level = 0.2). For the other five models (AdaBoost, Bayes, KN, RF, and SVM), the decoding accuracies in the entire visual cortex were 29.45%, 32.08%, 29.57%, 36.93%, and 35.43%, respectively. The results demonstrate that our proposed DCLSTM has higher decoding performance.

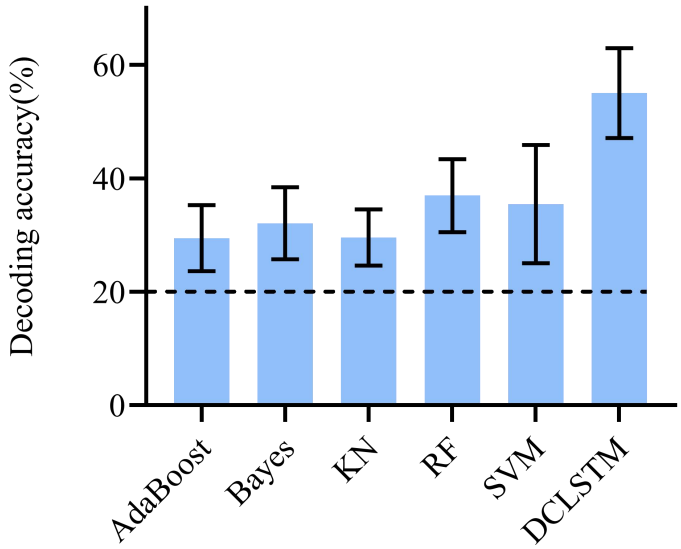

Figure 7 Compare the decoding accuracy of the six methods in the visual cortex (VC). LSTM using five traditional machine learning models (AdaBoost, Bayes, KN, RF and SVM)

In addition to comparing with commonly used machine learning algorithms, we also conducted comparisons on a publicly available dataset. The dataset is a 15-category dynamic video-induced fMRI dataset proposed by Wen et al. (Wen, Shi et al. 2018). The categories included indoor, outdoor, people, faces, birds, insects, aquatic animals, terrestrial animals, flowers, fruits, natural landscapes, cars, planes, ships, and sports. The training videos consisted of 18 stimuli, which were repeated twice by three subjects. Brain responses with correlation coefficients greater than a threshold were selected from the repeated responses and averaged. The test videos consisted of five stimuli, and the subjects

were required to watch them ten times. The final response was the average of the ten responses. Table 3 shows the results of using DCLSTM and LSTM to decode the 15-category dataset. It can be seen that the accuracy of using LSTM for 15-class classification is 25.5%, while using multiscale information fusion with time scales of 1, 2, 4, 8, or 16, can achieve an accuracy of around 30%. Therefore, using dilated convolutions with different coefficients to integrate cortical temporal information can improve decoding accuracy.

As shown above, we have demonstrated the advantages of the dynamic video decoding model based on dilated long short-term memory neural networks from two perspectives. First, compared with traditional decoding methods, using dilated convolutions to integrate temporal information can better decode category information. Second, we validated the universality of the model on a publicly available dataset. We calculated the decoding performance of 15 categories using LSTM and DCLSTM and found that our proposed DCLSTM can achieve a decoding performance about 5% higher than directly using LSTM on this dataset. Therefore, from these two perspectives, it can be seen that dilated long short-term memory neural networks can significantly improve classification decoding performance and have universality.

Table 3 Comparison of the decoding performance between DCLSTM and LSTM

Models	d=1	d=2	d=4	d=8	d=16
DCLSTM	30.1%	29.4%	29.3%	30.5%	31.3%
LSTM			25.5%		

2.4.6. Comparison of the decoding performance of DCLSTM-based temporal integrated brain information with single-time point brain information

Most previous studies(Haxby, Gobbini et al. 2001, Carlson, Schrater et al. 2003, Kamitani and Tong 2005) used peak responses of the BOLD signal as input, but Huang (Huang, Yan et al. 2020) showed that using brain signals within the period containing the peak responses could improve decoding performance. Therefore, to demonstrate that integrating brain information across different time scales can better decode category information, we compared the decoding performance of integrating occipital brain information with that of using single time points. The results are shown in Figure 8. First, for the entire visual cortex, the decoding performance of early integration is worse than that of single time points, but after 10 seconds, information integration can extract relevant information from the redundant brain signals, thereby improving decoding performance. Second, the best decoding result obtained by information integration was about 61%, while the best decoding result of single time point brain signals was about 56%, indicating that appropriate time scales should be selected to integrate brain signals across the entire visual cortex, which can significantly improve decoding performance.

In summary, the information accumulated in the early visual cortex in response to visual stimuli is limited, and more information is accumulated as time passes. Choosing an appropriate processing method to extract relevant information from the redundant brain signals is crucial. Our results demonstrate that the multi-scale fusion method can fully utilize the correlation between time series and extract relevant information. Thus, using the information integration method can significantly improve decoding performance.

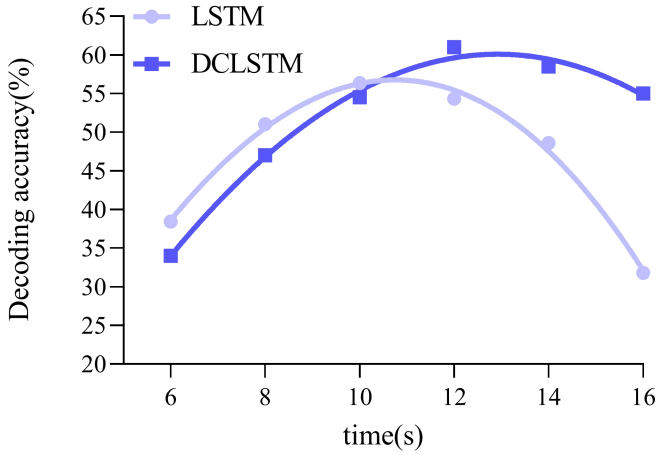

Figure 8 Decoding performance of information integration and single time point brain information in the occipital lobe. Information integration is the decoding performance of selecting the best time scale (integrating the time information of d=1,2 and 4). LSTM represents single time point brain information, and DCLSTM is the result of temporal information integration

2.4.7. Comparison of decoding performance of time-integrated brain information, average brain information, and time-scrambled brain information based on DCLSTM

In complex human environments, memory is not limited to local storage. People have long been accustomed to and skilled at extracting stored information from multiple scales, learning the correlation between information, and thus better understanding the world. To demonstrate that our proposed use of multi-scale integrated information is fully extracting relevant signals from accumulated brain signals, rather than randomly selecting, we also compared the decoding performance of time-integrated brain information with averaged brain information. Figure 9 shows the decoding differences between brain information integration and averaging on various visual cortices. Here, the information integration is selected with a time scale of 4 (integrating brain signals from 1, 2, and 4), while averaged brain information refers to the input of averaged brain information from 0 to 16 seconds. We found that on various visual cortices, information integration performed better than simple information averaging, with an overall decoding level above the 20%

random level. This suggests that our proposed method of using dilated convolution to extract multi-scale time information can extract relevant brain information from complex dynamic videos and brain signals. After fusing multi-scale brain information, it is more in line with the mechanism of processing complex events by the human eye, and can significantly improve decoding performance. Simple averaged brain information cannot fully learn relevant information in complex events. This suggests that the multi-scale time information extracted after dilated convolution is not simply averaged, but rather learns the correlation between before and after complex time series. Therefore, human memory is not simply local storage or averaging but rather spans the entire brain on multiple time scales. Proper information integration can better decode brain signals. We can clearly see that the entire occipital lobe has the highest decoding accuracy. This is mainly because the occipital lobe contains the vast majority of the visual area. Other visual regions such as V1/V2/FFA/OFA, they have fewer voxels. However, we can see that the decoding performance of the visual regions related to the face(FFA/OFA) is higher than the lower visual regions. This suggests that the voxels in high-level visual regions contain more category information, which is less related to the number of voxels.

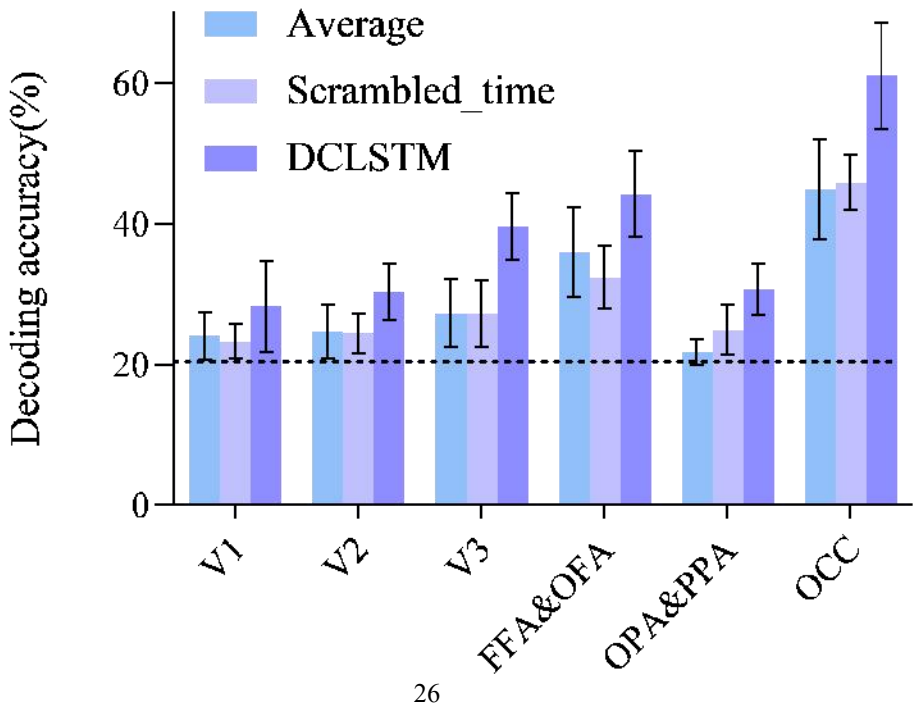

Figure 9 Compares the decoding performance of information integration, scrambled_time and average brain information across individual visual regions. Information integration is the decoding performance of selecting the best time scale (integrating the time information of d=1,2 and 4). Mean brain information was the mean brain signal value of 0-16s

2.4.8. Comparison of the decoding performance of brain signals in different brain regions

As is well known, the high-level visual cortex tends to process higher-level visual information, such as advanced semantics and categories, while the low-level visual cortex is more sensitive to simple contour information such as orientation. Therefore, we hypothesize that the high-level visual cortex has better object recognition abilities. To test this hypothesis, we compared the decoding performance of various visual cortex regions at the optimal integration scale, as shown in Figure 10. We can see that the overall decoding performance of the entire visual cortex is the best, followed by the brain regions associated with faces, with a decoding performance of approximately 44.2%. V3 comes next, with a decoding performance of about 39.6%. However, the decoding performance of the low-level visual cortex regions V1 and V2 is relatively low. It is worth noting that the decoding performance of brain regions related to scenes is also low, with an overall performance comparable to that of V2. There are several reasons for this. Firstly, the number of voxels related to scenes obtained from the functional image localization experiment we designed is relatively small, resulting in less available information and less accurate decoding of category information. Secondly, this localization method fully considers individual differences among subjects, resulting in significant differences in the number of voxels localized across subjects. Finally, compared with the brain regions related to faces, it is more difficult to localize brain regions related to scenes, mainly because scene information is more complex and varied than face information, with the vast majority of visual stimuli containing complex scene backgrounds, and dynamic video content being even more complex and varied. This results in less precise localization of

brain regions related to scenes. In summary, we can see that the high-level visual cortex can better decode category information compared to the low-level visual cortex.

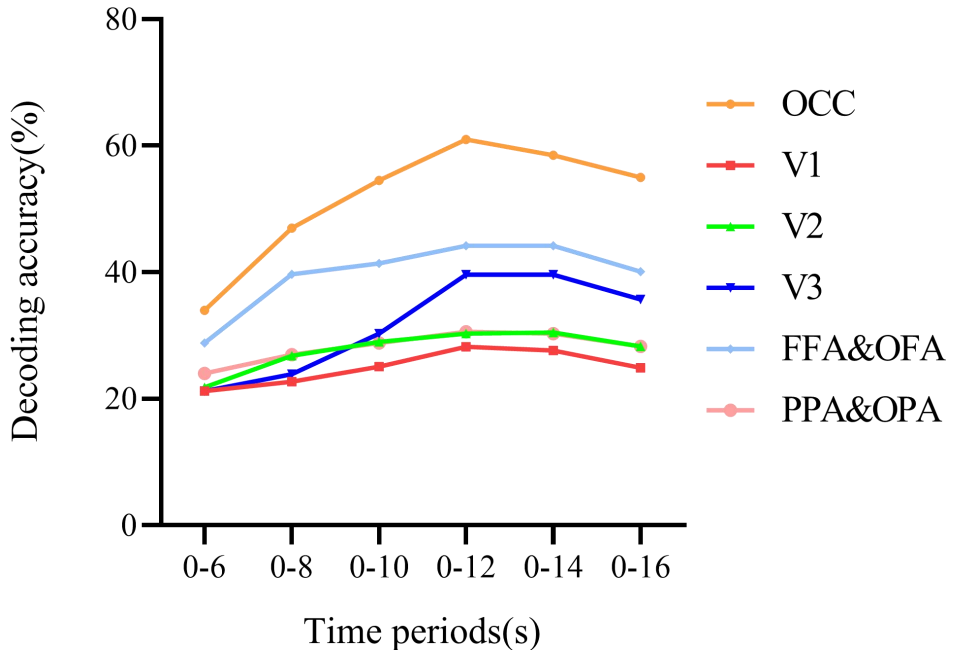

Figure 10 Decoding performance over time at a temporal integration scale of 4 (integrated brain information of d=1,2,4). The horizontal axis shows different periods, and 0-6 represents the input period of 0-6 seconds of brain information. The vertical axis is the decoding accuracy rate. Different colors represent different brain regions

2.4.9. Differences in the ability to integrate information in different brain regions

Hasson et al. (Hasson, Yang et al. 2008) used a large amount of functional magnetic resonance imaging data to discover time windows for visual processing in the visual cortex. They found that the early visual cortex had a shorter temporal reception window (TRW) while higher-order visual areas had longer TRWs. Since the TRWs of neurons in different brain regions determine the length of time for processing information, early sensory areas should have shorter TRWs to rapidly process changing sensory inputs. In contrast, higher-order visual areas should have longer TRWs to process information from perceptual and cognitive events unfolding over time. Therefore, the low-level visual cortex accumulates less information and tends to perceive and transmit visual information, while the higher-level visual cortex accumulates more information and tends to integrate

information. To prove this point, we compared the decoding accuracy of single time points and information integration over time for different visual cortical areas. As shown in Figure 11, for the low-level visual cortex (V1, V2, V3), using information from a single time point yielded better decoding performance in earlier periods. V1 had a better decoding performance with single time point information up to 14 seconds before, V2 up to 12 seconds before, and V3 up to about 11 seconds before. This indicates that information integration is not suitable for low-level visual cortex, and single time point information is better for decoding category information. In particular, in the low-level visual cortex, especially V1, information integration may introduce too much information unrelated to the stimulus, leading to decreased decoding performance. Furthermore, V1 accumulates the least amount of information among primary visual cortex regions, indicating that as cortical depth increases, information accumulation also gradually increases. Interestingly, in later periods, using information integration in the occipital cortex (OCC) and V3 significantly improved decoding performance, indicating that these two visual areas accumulate more information later and can better interpret brain activity patterns after integration. This phenomenon is particularly evident in higher-level visual areas. Brain regions related to "faces" (FFA&OFA) and "scenes" (OPA&PPA) can obtain good decoding performance with information integration in earlier periods, indicating that higher-level visual areas accumulate more information and can use information integration to extract relevant brain information earlier.

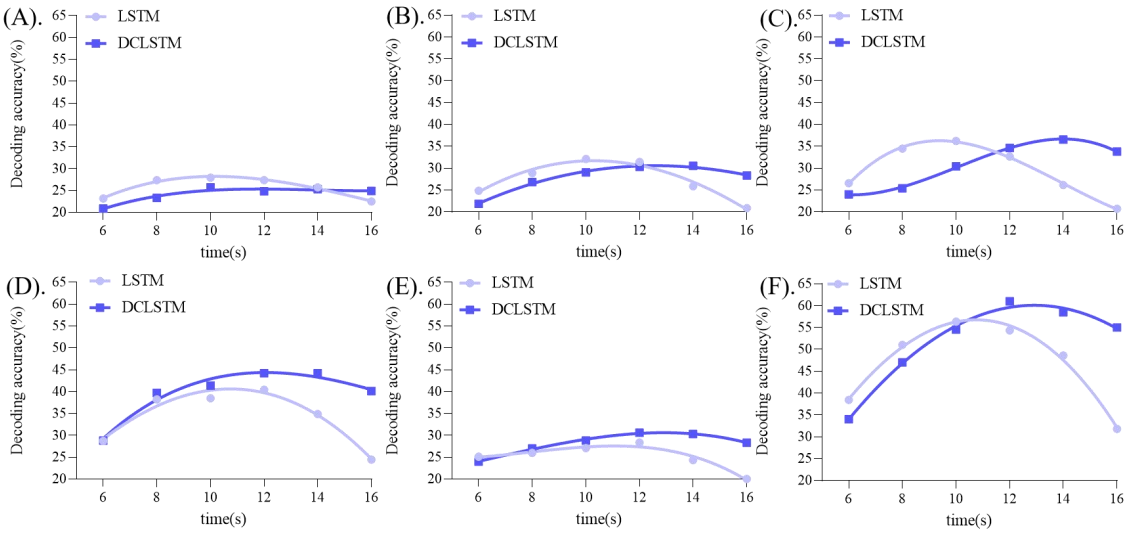

Figure 11 The figure shows the decoding performance over time for single time-point brain information and information integration (integration with d=1, 2, 4) in various brain regions. "LSTM" represents single time-point brain information and "DCLSTM" represents information integration. The horizontal axis represents different periods and time points, and the vertical axis represents decoding accuracy. A-F respectively represent V1, V2, V3, FFA&OFA, PPA&OPA, and OCC.

To further illustrate the differences in the integration capacity of visual cortical areas, we calculated the best decoding performance using single-time point brain information and integrated information for each visual cortical area, as shown in Figure 12. For lower visual cortical areas, particularly V1, the decoding accuracy of integrated information was lower than that of single-timepoint information. This suggests that there is little accumulation of information in lower-level visual cortical areas, and the integration of information introduces new neural activity from visual stimuli, leading to a decrease in decoding performance. In contrast, higher-level visual areas contain a large amount of accumulated information, particularly in face-related brain areas. After using integrated information, the decoding accuracy was found to be higher than that of single-information decoding. Therefore, the integration of temporal information in higher-level visual areas can significantly improve decoding performance.

In summary, the two figures above show that single-timepoint information can better decode category information for lower-level visual cortical areas, while integrated

information can better decode category information for higher-level visual cortical areas. This is consistent with previous cognitive findings: higher-level visual cortical areas accumulate more information over time and integrate information, but the function of lower-level visual cortical areas is mainly to perceive and transmit visual information, resulting in less accumulated content. Therefore, higher-level visual areas have a longer time window for sensory processing and require more time to process information.

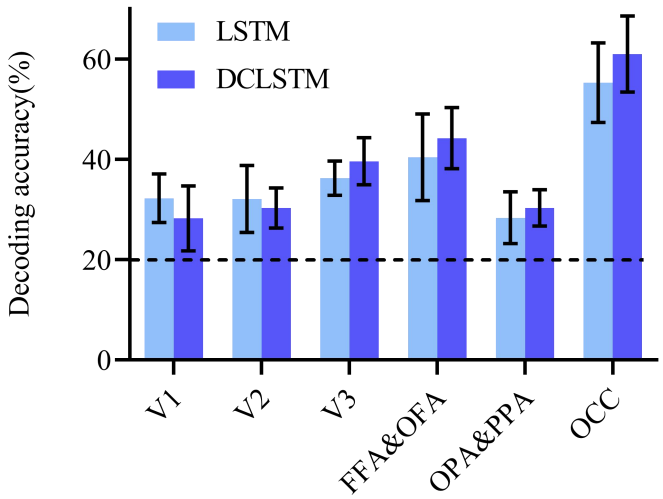

Figure 12 Compare the best decoding performance of information integration and single time-point in various visual areas. Information integration represents the decoding performance with the optimal time scale (i.e., integrating brain information with d=1,2,4), while the dashed line represents the chance level.

2.4.10. Correlation analysis between decoding accuracy and brain activity pattern

The basic assumption of brain decoding is that when subjects view the same visual stimuli, the brain responses are consistent or similar. This is a consensus and the basis of visual decoding, i.e., the brain produces similar responses to the same category of visual stimuli. However, our study found that the brain's ability to distinguish certain categories is poor. To analyze the brain representations of different categories, we first visualized the confusion matrix obtained by decoding using DCLSTM, as shown in Figure 13. We found some confusion between the five categories. For "animals", it is easily confused with "people" and "vehicles"; for "vehicles", it has the highest confusion with "animals" and "buildings". This confusion is contrary to the assumption, so we will analyze it further

from the perspective of brain activity.

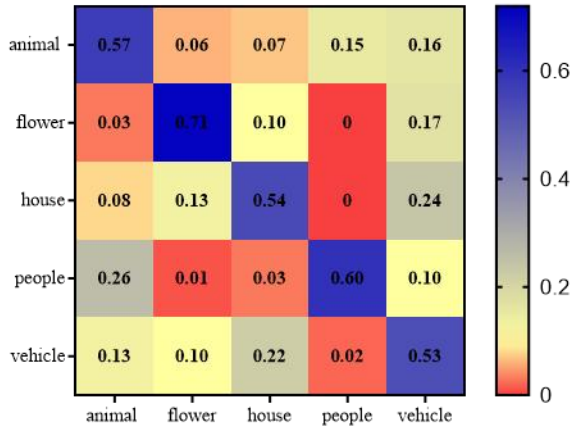

Figure 13 A confusion matrix using DCLSTM decoding (integrated d=1,2,4). The horizontal axis is the prediction result, and the vertical axis is the actual label

The neural encoding measure describes the brain's representation of different categories by mining correlations between image features and brain responses, while RSA(Kriegeskorte, Mur et al. 2008) describes the spatial similarity of different representations of stimulus images to analyze the brain's representation of different categories. We first visualized the correlation of occipital lobe (OCC) brain activity patterns using RDM, as shown in Figure 14 (A) below. It can be seen that for "people", the greater the dissimilarity with "flowers" and "houses"; As for "traffic", it is more similar to "houses" and "animals". This similarity is consistent with the confusion matrix, indicating that the confusion generated by our proposed model is due to the high similarity of brain activity patterns.

In addition, Figure 14 (B) shows the semantic distance of brain activity patterns between the five categories. We use cosine similarity to measure the semantic distance between categories. For cosine similarity, the larger the value, the smaller the corresponding distance, and the closer the semantic distance in space. Therefore, it can be clearly seen from the figure that "traffic" and "animal" and "house" have very similar semantic distance. In other words, for the fMRI responses obtained from the brain, the

semantic distance between "traffic" itself and "animal" and "house" is small, so these three categories have a high degree of similarity and confusion. From the perspective of semantic distance, it is also consistent with the confusion matrix. Therefore, we believe that the confusion of this model (DCLSTM) may be mainly due to the fact that the collected brain signals are too similar to distinguish very similar signals well.

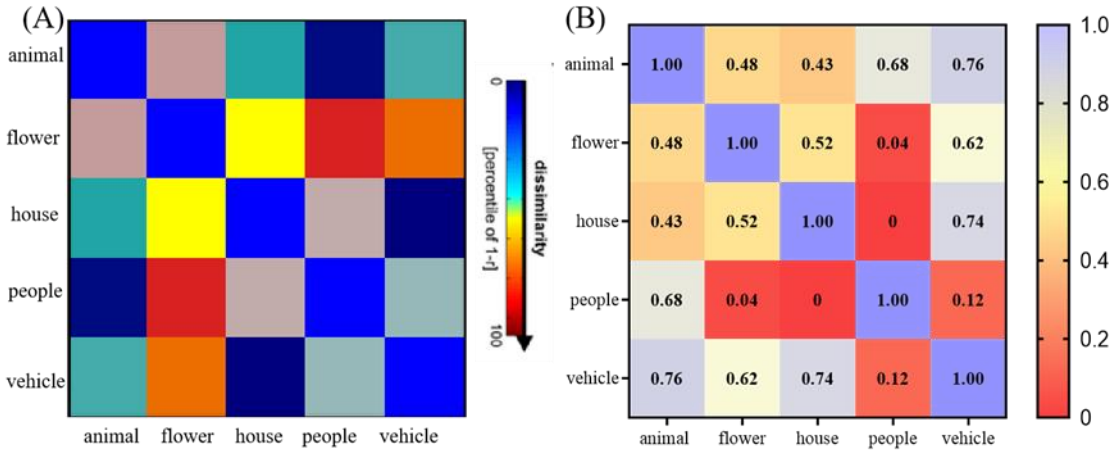

Figure 14 **(A)** Characterational dissimilarity matrix of occipital activity patterns. The horizontal and vertical axes are the visual stimuli. The color column indicates that the bluer the color, the more similar the corresponding brain activity pattern is, whereely, the redder the color indicates that the corresponding brain activity pattern is less similar, the greater the difference. **(B)** Cosine similarity matrix of occipital activity patterns induced by five categories of videos, with more blue color representing smaller semantic distance and more similarity. Redder indicates greater semantic distance and greater dissimilarity

2.4.11. Hierarchical feature similarity analysis of video categories

In 2001, Haxby et al.(Haxby, Gobbini et al. 2001) found that when subjects were presented with pictures of faces, cats, five kinds of objects and other chaotic images, the fMRI response patterns corresponding to different types of stimuli were different, but the response patterns of the same type of stimuli were similar. Therefore, according to the differences and similarities in brain activity patterns, the types of visual stimuli perceived by the brain can in turn be identified and decoded, which is the fMRI brain signal classification decoding principle. However, as shown in the 2.4.10 analysis, the brain activity we obtained was very similar in the "house" and "traffic" categories. So we went further to analyze how visual stimuli induce patterns of brain activity. We use

VGG16(Simonyan, Zisserman et al. 2014) to extract different levels of visual features and draw the representation dissimilarity matrix of each level of visual features. Figure 15 shows the representational dissimilarity matrix of image-level visual features, and image-level RDM can reflect the distance between images of any stimulus. We can see that the advanced visual features can clearly see the "block" structure, which means that the five categories can be clearly distinguished, indicating that different stimuli of the same category are still highly similar even if they are different. The intermediate visual feature is very fuzzy when distinguishing "traffic". Low-level visual features fail to distinguish categories. This is because low-level visual features, such as contour direction, have no categorical information and are not sufficient for classification, but high-level visual features contain more and richer semantic information and can be well classified.

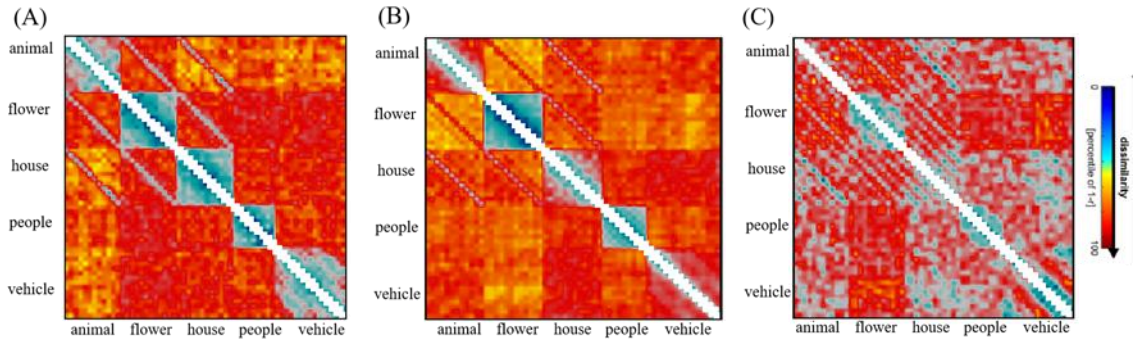

Figure 15 The representational dissimilarity matrix of image-level video features extracted from five categories of video based on VGG16, where the image-level represents each visual stimulus compared. (A) Representation dissimilarity matrix of high visual features; (B) Representation dissimilarity matrix of intermediate visual features; (C) Representation dissimilarity matrix of low-level visual features. The bluer the color column, the more similar the brain activity pattern, and the redder the color, the less similar the corresponding brain activity pattern, and the greater the difference

Class-level RDM represents the distance between the average representations of different classes of stimuli. For low, medium and high visual features, "vehicle" is easily confused with "house". Figure 16 shows the red box. This shows that "house" and "vehicle" have great similarities in terms of visual characteristics. This similarity is consistent with patterns of brain activity. As shown in Figure 16. To further quantify this similarity, we calculated the similarity between different levels of class-level visual

features and brain activity patterns, as shown in Figure 16 (E). It can be seen that there are high similarities between the obtained brain activity patterns and low, intermediate and high visual features, which also indicates that the brain activity patterns are driven by visual stimuli. Secondly, the pattern of brain activity is more similar to the intermediate visual features. It can also be clearly seen from the figure that different levels of visual features are very different.

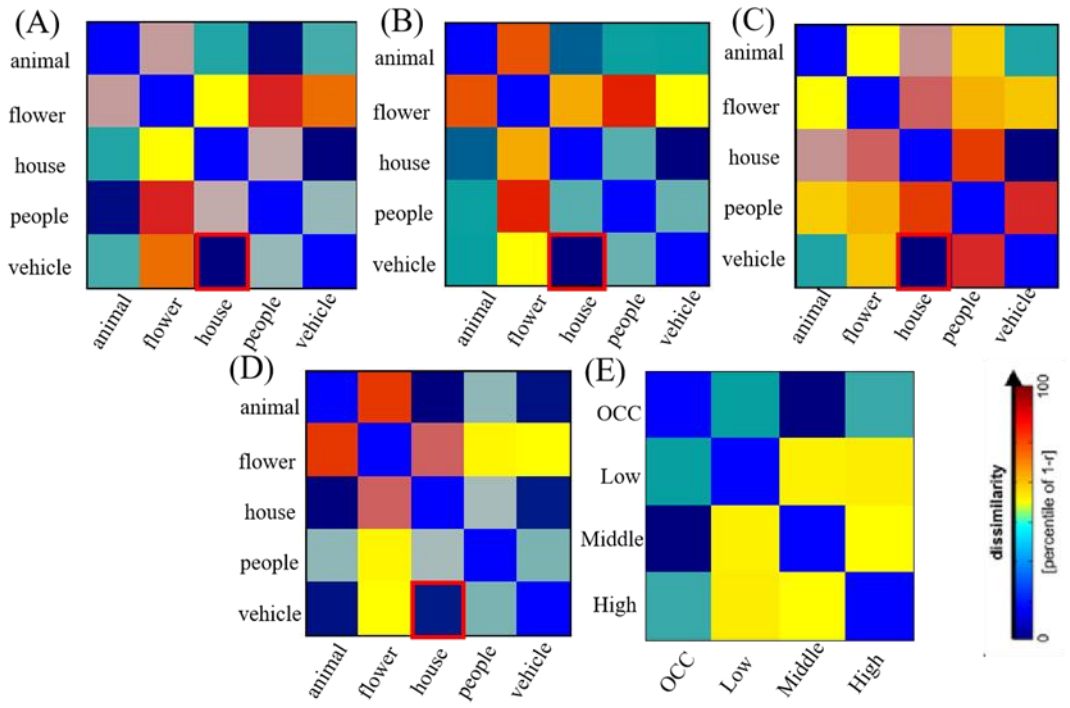

Figure 16 The representation dissimilarity matrix of five categories of video features extracted based on VGG16. (A) A representation dissimilarity matrix of brain activity patterns; (B) Representation dissimilarity matrix of low-level visual features; (C) Representation dissimilarity matrix of intermediate visual features; (D) Representation dissimilarity matrix of higher visual features; (E) Second-order representation dissimilarity matrix between brain activity patterns and visual features at different rank class levels. The color column showed that the bluer the color, the more similar the brain activity pattern, and the redder the color, the less similar the brain activity pattern, and the greater the difference

Reviewing the video stimuli, we found that the backgrounds of some of the vehicles and the backgrounds of the houses were very similar. As shown in Figure 17, (A) and (B) are different types of houses with blue sky as the background; (C) and (D) are different vehicles with blue sky and sea water in the background. The background of both subvideos

is the sky and the surrounding landscape, which also leads to the similarity of visual features extracted from the two categories. When comparing the image-level video features and class-level video features horizontally, it can be found that after the visual features of any video of each category are averaged. Many detailed features are blurred, which may lead to the confusion of class-level visual features to some extent. For example, when we distinguish Figures 17 (A) and (B) separately, it is obvious that they are two different types of buildings, but when the two "house" are averaged, some visual features may be blurred. At the same time, the background of "house" is highly similar to the background of "traffic", which is likely to further lead to the confusion between categories from the visual feature level analysis at the class level.

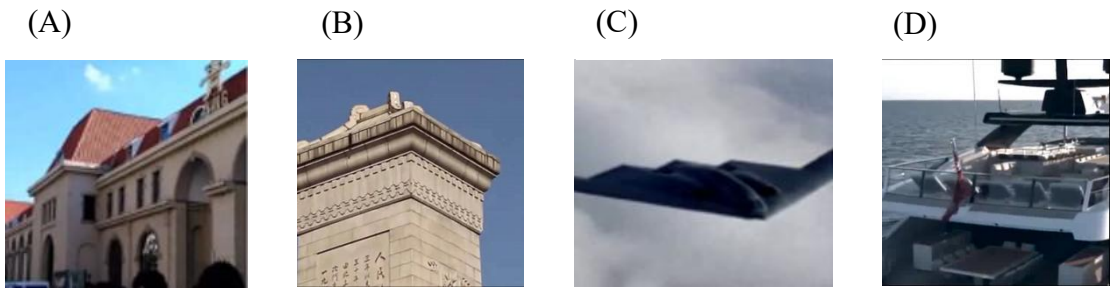

Figure 17 Schematic of visual stimulation for "building" and "vehicle".**(A)** and **(B)** are the buildings. **(C)** and **(D)** are the means of vehicles

3. Conclusions and discussion

It is well known that neurons in visual cortical pathways have increasingly large spatial receptive fields(Einevoll and Heggelund 2001). This is the basic organizing principle of the visual system. Neurons in higher visual areas receive input from smaller neurons in early vision, accumulating information from most of the space occupied by the objects and scenes they process(Hubel and Wiesel 1968). Events in the real world occur not only in an extended region of space, but also in an extended time. Therefore, the temporal response characteristics of different brain regions should also have a hierarchical structure similar to the size of spatial receptive fields (Hasson, Yang et al. 2008). Humans experience the world through visual perception as it unfolds

over time, and therefore must rely on the accumulation of information over a long period of time, including causal reasoning, processing linguistic information on various scales, understanding narratives, event segmentation, and human social interaction. In most real-life processes, past information is often used to process incoming information across multiple time scales(Hasson, Chen et al. 2015), and sense and perception use information integration to experience the external world. From a large amount of functional imaging data, scientists believe that visual cortex circuits can accumulate information gradually over time and have different time levels(Beauchamp 2005). In 2001, Rotshtein et al.(Rotshtein, Malach et al. 2001) found that there was less information accumulation in the early visual area, while there was more information accumulation in the advanced visual area. Therefore, the advanced visual area with information integration could better decode visual information. In 2008, Hasson et al.(Hasson, Yang et al. 2008) found from a large number of FMRI data that almost all cortical circuits can accumulate information over time and have different hierarchies on time scales. By having subjects watch both forward and backward silent movies and calculating the correlation coefficients, they found that the early visual cortex had a shorter time window and higher order visual areas had a longer time window. Specifically, a neuron's Temporal Receptive Windows (TRWs)(Hasson, Yang et al. 2008) is defined as the length of time before sensory information influences the response. Since the TRWs of neurons in brain regions determine the length of time in which information can be processed in the past, it is assumed that the range of TRWs in each region must correspond to its functional role. The TRWs in the early sensory areas should be short and able to quickly process changing sensory input. In contrast, TRWs in some higher-level regions should be longer, allowing them to process information in perceptual and cognitive events that unfold over time. Thus, memory is not limited to some local storage, but rather an intrinsic feature of information

processing that runs throughout the brain on multiple timescales. As a result, information integration can better decode brain signals.

Based on the data set proposed in this thesis, the advantages of DCLSTM in information integration are analyzed from multiple perspectives. We first compare the decoding performance of the model under different time integration scales, and find that choosing the appropriate time integration scale according to the length of the time series is helpful to improve the decoding performance. Next, we compare the decoding performance of traditional machine learning algorithms and DCLSTM, and verify the universality of the model on public data sets. Then, we analyzed the accuracy of time information integration and decoding of brain information of single time point, average brain information and disturbed time series, and found that multi-scale fusion can make full use of the correlation between time series and extract the relevant information to improve the decoding accuracy. We also compared the differences in decoding between different brain regions, and found that higher visual regions contain more category information and can be better decoded. Then, we further compare the differences of information integration in different visual areas, and find that the low-level visual areas have less information accumulation and less integration. More information is accumulated in higher visual areas, and more information is integrated accordingly. Since there is a high degree of confusion in the results of "building" and "traffic", we analyzed the brain activity pattern and visual features, and found that it is likely that the highly similar background of these two types of visual stimuli caused the confusion.

4. Bibliography

Amir, A., J. Argillander, M. Campbell, A. Haubold, G. Iyengar, S. Ebadollahi, F. Kang, M. R. Naphade, A. Natsev and J. R. Smith (2003). IBM Research TRECVID-2003 Video Retrieval System. TRECVID.

Barch, D. M., G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit and C. J. N. Feldt (2013). "Function in the human connectome: task-fMRI and individual differences in behavior." **80**: 169-189.

Beauchamp, M. S. J. C. o. i. n. (2005). "See me, hear me, touch me: multisensory integration in lateral occipital-temporal cortex." **15**(2): 145-153.

Burkhalter, A., D. Felleman, W. Newsome and D. J. V. r. Van Essen (1986). "Anatomical and physiological asymmetries related to visual areas V3 and VP in macaque extrastriate cortex." **26**(1): 63-80.

Carlson, T. A., P. Schrater and S. J. J. o. c. n. He (2003). "Patterns of activity in the categorical representations of objects." **15**(5): 704-717.

Dilks, D. D., J. B. Julian, A. M. Paunov and N. J. J. o. N. Kanwisher (2013). "The Occipital Place Area Is Causally and Selectively Involved in Scene Perception." **33**(4): 1331-1336.

Dumoulin, S. O. and B. A. J. N. Wandell (2008). "Population receptive field estimates in human visual cortex." **39**(2): 647-660.

Einevoll, G. T. and P. J. V. N. Heggelund (2001). "Mathematical models for the spatial receptive-field organization of nonlagged X-cells in dorsal lateral geniculate nucleus of cat." **17**(6): 871-885.

Epstein, R. and N. J. N. Kanwisher (1998). "A cortical representation of the local visual environment." **392**(6676): 598-601.

Epstein, R. and N. J. N. Kanwisher (1998). "Epstein, R. & Kanwisher, N. Cortical representation of the local visual environment. Nature 392, 598601." **392**(6676): 598-601.

Gerstner, W., A. K. Kreiter, H. Markram and A. V. J. P. o. t. N. A. o. S. Herz (1997). "Neural codes: firing rates and beyond." **94**(24): 12740-12741.

Goebel, R., F. Esposito and E. J. W. S. S. Formisano, Inc., A Wiley Company (2006). "Analysis of FIAC data with BrainVoyager QX: From single-subject to cortically aligned group GLM analysis and self-organizing group ICA." **27**(5): 392-401.

Hassabis, D., R. N. Spreng, A. A. Rusu, C. A. Robbins, R. A. Mar and D. L. J. C. C. Schacter (2014). "Imagine all the people: how the brain creates and uses personality models to predict behavior." **24**(8): 1979-1987.

Hasson, U., J. Chen and C. J. J. T. i. c. s. Honey (2015). "Hierarchical process memory: memory as an integral component of information processing." **19**(6): 304-313.

Hasson, U., E. Yang, I. Vallines, D. J. Heeger and N. J. J. o. N. Rubin (2008). "A hierarchy of temporal receptive windows in human cortex." **28**(10): 2539-2550.

Haufe, S., F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz and F. J. N. Bießmann (2014). "On the interpretation of weight vectors of linear models in multivariate neuroimaging." **87**: 96-110.

Haxby, J. V., M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten and P. J. S. Pietrini (2001). "Distributed and overlapping representations of faces and objects in ventral temporal cortex." **293**(5539): 2425-2430.

Haynes, J.-D. and G. J. N. n. Rees (2005). "Predicting the orientation of invisible stimuli from activity in human primary visual cortex." **8**(5): 686-691.

Horikawa, T. and Y. J. N. c. Kamitani (2017). "Generic decoding of seen and imagined objects using hierarchical visual features." **8**(1): 15037.
Horikawa, T., M. Tamaki, Y. Miyawaki and Y. J. S. Kamitani (2013). "Neural decoding of visual imagery during sleep." **340**(6132): 639-642.
Hu, X., K. Li, J. Han, X. Hua, L. Guo and T. J. I. T. o. M. Liu (2011). "Bridging the semantic gap via functional brain imaging." **14**(2): 314-325.
Huang, W., H. Yan, C. Wang, J. Li, X. Yang, L. Li, Z. Zuo, J. Zhang and H. J. H. B. M. Chen (2020). "Long short-term memory-based neural decoding of object categories evoked by natural images." **41**(15): 4442-4453.
Hubel, D. H. and T. N. J. T. J. o. p. Wiesel (1968). "Receptive fields and functional architecture of monkey striate cortex." **195**(1): 215-243.
Kahn, D., E. F. Pace-Schott and J. A. J. N. Hobson (1997). "Consciousness in waking and dreaming: the roles of neuronal oscillation and neuromodulation in determining similarities and differences." **78**(1): 13-38.
Kamitani, Y. and F. J. N. n. Tong (2005). "Decoding the visual and subjective contents of the human brain." **8**(5): 679-685.
Kay, K. N., T. Naselaris, R. J. Prenger and J. L. J. N. Gallant (2008). "Identifying natural images from human brain activity." **452**(7185): 352-355.
Kriegeskorte, N., M. Mur and P. A. J. F. i. s. n. Bandettini (2008). "Representational similarity analysis-connecting the branches of systems neuroscience." 4.
Maguire, E. A. J. S. J. o. P. (2010). "The retrosplenial contribution to human navigation: A review of lesion and neuroimaging findings." **42**(3): 225-238.
Miyawaki, Y., H. Uchida, O. Yamashita, M.-a. Sato, Y. Morito, H. C. Tanabe, N. Sadato and Y. J. N. Kamitani (2008). "Visual image reconstruction from human brain activity using a combination of multiscale local image decoders." **60**(5): 915-929.
Naseer, N. and K.-S. J. F. i. h. n. Hong (2015). "fNIRS-based brain-computer interfaces: a review." **9**: 3.
Naselaris, T., R. J. Prenger, K. N. Kay, M. Oliver and J. L. J. N. Gallant (2009). "Bayesian reconstruction of natural images from human brain activity." **63**(6): 902-915.
Nishimoto, S., A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu and J. L. J. C. b. Gallant (2011). "Reconstructing visual experiences from brain activity evoked by natural movies." **21**(19): 1641-1646.
Oord, A. v. d., S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. J. a. p. a. Kavukcuoglu (2016). "Wavenet: A generative model for raw audio."
Oord, A. V. D., N. Kalchbrenner and K. Kavukcuoglu (2016). "Pixel Recurrent Neural Networks."
Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. J. t. J. o. m. L. r. Dubourg (2011). "Scikit-learn: Machine learning in Python." **12**: 2825-2830.
Rotshtein, P., R. Malach, U. Hadar, M. Graif and T. Hendler (2001). "Feeling or Features: Different Sensitivity to Emotion in High-Order Visual Cortex and Amygdala - ScienceDirect."
Rotshtein, P., R. Malach, U. Hadar, M. Graif and T. J. N. Hendler (2001). "Feeling or features: different sensitivity to emotion in high-order visual cortex and amygdala." **32**(4): 747-757.
Schwarzkopf, D., B. de Haas and I. J. O. S. F. Alvarez (2018). "SamSrf 6-Toolbox for pRF modelling." **10**.
Spiers, H. J. and E. A. J. T. i. c. s. Maguire (2007). "Decoding human brain activity during

real-world experiences." **11**(8): 356-365.
Sundermeyer, M., R. Schlüter and H. Ney (2012). LSTM Neural Networks for Language Modeling. Interspeech.
Tsantani, M., N. Kriegeskorte, K. Storrs, A. L. Williams, C. McGettigan and L. J. J. o. N. Garrido (2021). "FFA and OFA encode distinct types of face identity information." **41**(9): 1952-1969.
Wen, H., J. Shi, W. Chen and Z. J. S. r. Liu (2018). "Deep residual network predicts cortical representation and organization of visual features for rapid categorization." **8**(1): 3752.
Wen, H., J. Shi, Y. Zhang, K.-H. Lu, J. Cao and Z. J. C. c. Liu (2018). "Neural encoding and decoding with deep learning for dynamic natural vision." **28**(12): 4136-4160.
Yovel, G. and N. J. N. Kanwisher (2004). "Face Perception : Domain Specific, Not Process Specific." **44**(5): 889-898.
Chen, Y.-W., & Lin, C.-J. (2006). Combining SVMs with various feature selection strategies. Feature extraction (pp. 315–324). Berlin, Heidelberg:Springer.
Huang, W., Yan, H., Liu, R., Zhu, L., Zhang, H., & Chen, H. (2018). F-score feature selection based Bayesian reconstruction of visual image from human brain activity. Neurocomputing, 316, 202–209.
Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. ArXiv Preprint ArXiv:1409.1556, 2014.