Optical Illusions in Assessing Deep Neural Networks and Human Vision: A Study Exploring Towards Improved Brain-inspired Modeling

by

Hongtao Zhang
Student ID Number: 1258004

A dissertation submitted to the
Engineering Course, Department of Engineering,
Graduate School of Engineering,
Kochi University of Technology,
Kochi, Japan

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Assessment Committee:

Supervisor:	Shinichi Yoshida
Co-Supervisor:	Xiangshi Ren
Co-Supervisor:	Kiminori Matsuzaki
Committee Member:	Yukinobu Hoshino
Committee Member:	Liang Sun

September 2024

Abstract

Deep Neural Networks (DNNs), an approach inspired by human brain internal mechanism, have revolutionized fields like image recognition, natural language processing, and speech recognition through their advanced feature extraction and pattern recognition capabilities. Despite these advancements, DNNs still exhibit significant challenges related to robustness and interpretability, particularly when subjected to minor input variations and adversarial conditions. In other word, there is still a significant gap between DNNs and the human visual system. Thus, DNNs only can be as one of vision study models to learn potential visual mechanism. Based on this idea, we dedicated to explore how much DNNs similar to human vision that call brain-like characteristic and how improve the brain-like modeling on DNNs.

Visual illusions serve as an intriguing tool to explore the parallels and differences between human visual perception and machine vision. These illusions often exploit the ways in which humans process visual information, revealing the underlying mechanisms of our perception. Historically, visual illusions have been used to probe the workings of the human brain, offering insights into depth perception, color constancy, and the geometrical interpretation of space. Therefore, by studying how DNNs handle these illusions, researchers can uncover the extent to which neural networks simulate human-like perception and where they differ, shedding light on both the capabilities and limitations of these systems.

Thus, This study delves into the simulation of human visual perception by DNNs, using a unique and comprehensive visualization approach that integrates six classical visual illusions to probe and compare the brain-like characteristics of different DNNs architectures. Specifically, depending on the human perceptual data as benchmark, we integrated visual and analytical techniques, including representational similarity analysis and class activation maps, to provide deeper explain of internal mechanism into how DNNs process visual illusions.

Our experimental results indicate that visual illusions are widely present in DNNs. Despite differences between models, those exhibiting visual illusion effects share some common patterns, such as relatively low network complexity and relatively simple architectures. For example, classic DNNs models like VGG19 and ResNet101. However, an important gap is highlighted: in the distribution of feature attention heatmaps, DNNs primarily focus on the overall features of objects and fail to understand real physical concepts, a fundamental difference from humans. Furthermore, they are highly influenced by training dataset. For instance, pre-trained weights on the ImageNet dataset lead models to have a preference for focusing on the edges of geometric shapes. This

results in good performance on simple and singular types of visual illusions, while complex visual illusions with multiple components exhibit irregular and inconsistent visual illusion effects.

Moreover, from the RDMs of DNNs with simpler architectures that perform well, high visual illusion responses in shallow layers align with the high correlation of visual illusions in V1/V2 observed in fMRI experiments. This demonstrates the potential similarity between DNNs and the visual pathways, especially the early pathways and the shallow modules of DNNs. This suggests the importance of focusing on the architecture and feature information of primary modules in brain-like modeling.

In summary, this comparative study of DNNs architectures and classical visual illusions provides important insights into the differences between human and DNNs perception. The main contribution of this study is as following :

1. This study contributes to the understanding of DNNs behavior in visual illusions and establishes methods for further examining their brain-like processing capabilities. We provide evidence demonstrating the potential brain-like advantages and limitations of DNNs.

2. By integrating neuroscientific findings into DNNs development, this work supports targeted improvements in network architecture to more closely align with human cognitive processes. Through detailed analysis and experimental insights, this research provides the reference on improving DNNs' performance in tasks requiring complex visual processing and interpretation.

3. This study reveals the strengths and weaknesses of DNNs in handling visual illusions, offering new perspectives on their potential and limitations in practical applications. For example, in fields such as autonomous driving, medical image analysis, and human-computer interaction, understanding and improving the visual perception capabilities of DNNs can significantly enhance their performance and reliability.

Future research can further explore how optimizing training data, improving network architectures, and integrating multimodal information can enhance DNNs performance in complex visual tasks.

Acknowledgements

Time flies, and before I knew it, three years of my Ph.D. journey have passed.

First of all, I would like to express my deepest gratitude to all those who provided me with the opportunity to complete this research.

My sincere and heartfelt thanks go firstly to my supervisor, Professor Shinichi Yoshida, for his support and understanding throughout my research. His expertise, insight, and patience have greatly enriched my Ph.D. experience. Furthermore, it is my honor to benefit from his personality and diligence, which I will treasure my whole life. My gratitude to him knows no bounds.

I also want to sincerely thank my two co-supervisors, Professor Xiangshi Ren and Professor Kiminori Matsuzaki. Their guidance and suggestions have broadened my research perspective and helped me make further academic progress.

I would be grateful to professors in the committee, especially Professor Liang Sun and Professor Yukinobu Hoshino, for their instructive advice inspired me a lot.

Sincere thanks to my fellow graduates, Zhixing Guo and Peng Tan. Their support in both research and daily life has allowed me to smoothly navigate my three-year doctoral candidate's period.

Special thanks to my friends, Yang Guang and his family. Their encouragement and help have provided me with great support when I encountered difficulties.

Finally, I am deeply grateful to my parents and my younger brother for their selfless support and encouragement. Their love has been the driving force behind my perseverance.

Contents

Abstract ii

Acknowledgements v

List of Figures xi

List of Tables xiii

1 Introduction 2
- 1.1 Overview . . . 2
- 1.2 Research Objectives . . . 3
- 1.3 Structure of the Dissertation . . . 4

2 Background, Motivation and Purpose 6
- 2.1 Deep Neural Networks . . . 6
  - 2.1.1 What is Deep Neural Networks? . . . 6
  - 2.1.2 Convolutional Neural Networks . . . 7
  - 2.1.3 Recurrent Neural Networks and Long Short-Term Memory . . . 8
  - 2.1.4 3DCNN . . . 9
  - 2.1.5 PredNet . . . 10
  - 2.1.6 Transformers . . . 11
  - 2.1.7 Applications of Deep Neural Networks . . . 12
- 2.2 Optical Illusion . . . 12
  - 2.2.1 Types of Optical Illusion . . . 12
  - 2.2.2 Mechanisms Behind Optical Illusion . . . 13
  - 2.2.3 Implications of Optical Illusion . . . 14
- 2.3 Interdisciplinary Study . . . 14
  - 2.3.1 Ventral Pathway . . . 14
  - 2.3.2 Mapping DNNs to Ventral Pathway . . . 15
- 2.4 Why DNNs with Optical Illusion? . . . 16
  - 2.4.1 Related Research and Identification of Gaps . . . 18
  - 2.4.2 How do DNNs Respond to Different Types of Visual Illusions? . . . 20
  - 2.4.3 What Are the Similarities and Differences Between Human and DNN Perceptions of Visual Illusions? . . . 20
  - 2.4.4 How Do Different DNN Architectures Compare in Their Ability to Simulate Visual Illusions? . . . 20

2.4.5 What Are the Computational Principles Underlying the DNNs' Ability to Simulate Visual Illusions? . . . 20
2.4.6 Can the Findings From DNN Simulations of Visual Illusions Inform the Development of More Advanced AI Systems? . . . 20
2.4.7 Meaning and Contribution . . . 21
2.5 Proposed Method . . . 21
2.5.1 Human Perceptual Data on Optical Illusion . . . 21
2.5.2 Representational Similarity Analysis . . . 22
2.5.3 Class Activation Mapping . . . 23
2.5.4 The Main Purpose, Idea and Method . . . 24

3 Do the Illusion Performed in DNNs? 27
3.1 Step1 :Müller-Lyer Illusion on Vgg19 and ResNet101 . . . 27
3.1.1 Human Experiment on Müller-Lyer Illusion . . . 27
3.1.2 Brain-like DNNs . . . 29
3.1.3 The Preliminary Test on Two DNNs Models . . . 30
3.1.4 The Distribution of Human Perceptual Length . . . 30
3.1.5 Representational Dissimilarity Matrix . . . 31
3.1.6 The Illusion Response Changing of Different Model Depth . . . 33
3.2 Step1 :Müller-Lyer Illusion on More DNNs Models . . . 35
3.2.1 The RDM on Eight Models . . . 36
3.2.2 The L2 Distance Changes on Eight Models . . . 36
3.2.3 Grad-CAM Visualization on Eight Models . . . 38
3.3 Step 2: Five Optical Illusions . . . 39
3.3.1 The Human Experiment on Five Optical Illusions . . . 39
3.3.2 Models and Processes . . . 42
3.3.3 Color Assimilation . . . 43
3.3.4 Hermann Grid Illusion . . . 47
3.3.5 Müller-Lyer Illusion . . . 52
3.3.6 Poggendorff Illusion . . . 54
3.3.7 Zöllner Illusion . . . 56

4 Illusion Performance in DNNs by Specific Designed Dataset 58
4.1 Oblique Illusion and Human Subject Experiment . . . 58
4.2 Training Approach and Testing . . . 60
4.2.1 Model Selection and Training Strategy . . . 60
4.2.2 Permutation Test . . . 61
4.2.3 The Visualization on RDMs . . . 61
4.2.4 The Visualization on Grad-CAM . . . 61
4.3 Results . . . 62
4.3.1 Perceived Angles . . . 62
4.3.2 Model Performance . . . 62
4.3.3 Visualization Interpretations and Differences . . . 66

5 DNNs: Spatiotemporal vs Static 69
5.1 Main idea . . . 69
5.2 Video Models and Training Strategies . . . 70
5.2.1 Video Dataset . . . 70

5.2.2 Video Models . . . 72
5.2.3 Teacher-Student Self-Supervised Learning . . . 74
5.2.4 Representation Similarity Analysis (RSA) . . . 76
5.2.5 Grad-CAM . . . 77
5.2.6 Total Research Method . . . 78
5.3 Results . . . 79
5.3.1 Model Understanding of Line Length and Basis for Visual Illusions 79
5.3.2 Temporal vs. Static Characteristics on RDMs . . . 82
5.3.3 Temporal vs. Static Characteristics on Grad-CAM . . . 85

6 Mapping relationship between DNNs and visual pathway through fMRI and optical illusion 87
6.1 Experimental Introduction and Procedures . . . 87
6.1.1 Visual Illusions and DNNs . . . 88
6.1.2 fMRI Experiments . . . 88
6.2 Result . . . 90
6.2.1 Heatmaps of Visualization . . . 90
6.2.2 ROIs Response . . . 92

7 Discussion 94
7.1 How Do DNNs Respond to Different Types of Visual Illusions? . . . 94
7.2 What Are the Similarities and Differences Between Human and DNN Perceptions of Visual Illusions? . . . 95
7.3 How Do Different DNN Architectures Compare in Their Ability to Simulate Visual Illusions? . . . 96
7.4 What Are the Computational Principles Underlying the DNNs' Ability to Simulate Visual Illusions? . . . 97
7.5 Can the Findings From DNN Simulations of Visual Illusions Inform the Development of More Advanced AI Systems? . . . 98

8 Conclusion 100

Bibliography 103

List of Publication 112

List of Figures

2.1 The architecture of CNN. 7
2.2 The architecture of RNN and LSTM. 9
2.3 The architecture of 3D-CNN. 10
2.4 The architecture of PredNet. 10
2.5 The architecture of Transformer. 11
2.6 What is optical illusion? The main types of optical illusion. 13
2.7 An example of recognition and motion through two visual pathway. 15
2.8 Mapping relationship between DNNs and ventral pathway. 16
2.9 Interdisciplinary study between DNNs and Neuroscience 19
2.10 Human subject experiment 21
2.11 RDM 23
2.12 Grad-CAM 24
2.13 Main research processes 26

3.1 The main subject experiment on Müller-Lyer Illusion 29
3.2 The perceptual length on Müller-Lyer Illusion 31
3.3 The RDM between perceped group and control group on VGG19 and ResNet101 33
3.4 The illusion response about model depth of Vgg19 and ResNet101 34
3.5 The RDMs of eight models on arrow outward and inward 37
3.6 The L2 distance trend of two group on eight models 38
3.7 The heatmap of feature attention on eight models 39
3.8 Five optical illusions and their human subject experiment configuration . 41
3.9 Human color ranking and models color test on 12 colors. A: Human subjects' color ranking on three depth of each color. B: The color depth ranking from 12 DNNs models. C: The color depth ranking on different model depth. 44
3.10 The RDMs of 12 colors within 12 models 45
3.11 The CAM heatmap of green color on 12 models 46
3.12 The response of grid illusion from human subjects and DNNs illusion test. A: The distribution of four colors gird illusion from subjects. B: The DNNs test on four colors gird illusion. 49
3.13 The visualization heatmap of Hermann Grid Illusion on 12 DNNs. A: RDMs of four color on DNNs. B: The feature attention focus from 12 DNNs. 50
3.14 The illusion trend on different model depth. 51

3.15 The perceptual length of Müller-Lyer Illusion and illusion test of DNNs. A: Human subject perceptual length between arrow outward and inward. B: The heatmap of feature attention focus on 12 DNNs. C: The L2 distance of different model depth on two groups. 53
3.16 The RDMs on Müller-Lyer Illusion within perceptual gourp and control group in the DNNs. 54
3.17 Human subject perceptual illusion and DNNs testing. A: The visual bias of Poggendorff illusion. B: Different model depth's L2 distance of illusion and perceptual stimulus. C: The RDMs of 12 DNNs 55
3.18 The heatmap of feature focus on Poggendorff Illusion from 12 DNNs. . . . 56
3.19 The huamn subject perceptual angles and model test. A:The average perceptual angle on Zöllner illusion. B: The L2 distance of different model depth. C: The RDMs of 12 DNNs. 57
3.20 The CAM visulization of 12 DNNs on Zöllner illusion. 57

4.1 The main components of the Skye's Oblique Grating Illusion. 59
4.2 The preparation of the Skye's Oblique Grating Illusion. 59
4.3 The eight illusion strength of optical illusion and dataset preparation. . . 60
4.4 The human subjects' perceptual degree on 12 colors. 62
4.5 The permutation test result of 12 DNNs. 63
4.6 The models testing of C1 and C2. 64
4.7 The result of color and illusion strength on DNNs. 65
4.8 The feature attention focus of C1 and C2 on DNNs. 66
4.9 The feature attention focus of ResNet101. 66
4.10 The average L2 distance on DNNs. 67
4.11 The RDMs of pretrained and self-data trained ResNet101. 68
4.12 The RDM of the initial module of ResNet101 and mapping realtionship. . 68

5.1 The video dataset of Müller-Lyer illusion (Type A and B). 71
5.2 The main architecture of five video models. 73
5.3 The main training strategy through two steps. 76
5.4 L2 distance between same group and different group. A: The L2 distance of 10 length label on same arrow orientation from perception group and control group. B: The L2 distance of 10 length label on different arrow orientations within two groups. 81
5.5 RDMs of perception group and control group on video model and statics model. 84
5.6 The heatmaps of feature focus from four video classification models and statics DNNs 86

6.1 The main three step on exploring the mapping relationship between DNNs and ventral pathway. 88
6.2 The RDMs of six on C1 and C2. 91
6.3 The attention heatmaps of six DNNs on illusion. 92
6.4 Average response distribution under specific ROIs (V1/V2/V4/IT). 93

List of Tables

3.1 The main configuration and parameters of various DNNs Models 42

5.1 Müller-Lyer Illusion Video Dataset Overview 72
5.2 Video Model Configuration Parameters 74

Chapter 1

Introduction

1.1 Overview

The development of Deep Neural Networks (DNNs) has achieved revolutionary progress in numerous fields, especially in image recognition, natural language processing, and speech recognition [1]. DNNs achieve high-level feature extraction and complex pattern recognition through a multi-layered neuronal structure, enabling the extraction of deep features from data [1]. However, despite their outstanding performance in handling these tasks, DNNs still exhibit significant limitations in robustness and interpretability. They are highly sensitive to minor changes in input data and are easily make errors when facing adversarial attacks or extreme situations, revealing their sensitivity to inputs and insufficient understanding of complex contexts [2, 3].

These limitations have prompted researchers to seek inspiration from brain-like computing to improve neural network design and performance. The human brain exhibits high robustness and flexibility in processing visual information, making accurate judgments even in complex and ambiguous environments [4]. In particular, the brain integrates contextual information and multi-level feature processing to address various visual challenges [5, 6]. This hierarchical processing structure and feedback mechanism provide important insights for the improvement of DNNs [4].

In this process, research in artificial vision systems has increasingly focused on how to enhance the visual perception capabilities of models to make them more akin to the human visual system. For example, by enhancing the contextual understanding and temporal information processing capabilities of DNNs, introducing recurrent networks

(RNNs) and self-attention mechanisms to mimic the feedback loops and attention regulation in the visual cortex [7, 8]. These advancements have improved DNNs' performance and robustness in complex visual tasks, such as object recognition and scene understanding [9]. Moreover, models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have also demonstrated powerful potential in data generation and understanding [10].

However, despite these improvements enhancing some brain-like features of DNNs, such as processing speed and accuracy, they are still limited in simulating real neurobiological functions. Achieving brain-like characteristics is not merely through simulating a single process or mechanism; it requires a deep understanding and comprehensive simulation of various brain functions [11]. In this context, research around brain-like computing has deepened, exploring various directions, including the study of perceptual illusions.

Visual illusions, as a branch of this research, provide a unique perspective on understanding the mechanisms and limitations of DNNs. Visual illusions are a common phenomenon in the human visual system, reflecting how the brain processes visual inputs in specific contexts [12]. By studying the reactions of DNNs to images that induce visual illusions, researchers can more deeply analyze the model's deficiencies in visual recognition and information processing [13]. This research not only reveals the processing characteristics of DNNs but also helps researchers improve network structures, making them perform more similarly to the human brain in handling a broader range of visual tasks [14].

The integration of visual illusions provides a new perspective for DNN research, using brain-like mechanisms to reveal and understand the limitations of neural networks. This research can guide potential improvements in model optimization and training methods [12]. As research in this area deepens, the integration of visual illusions and DNNs will continue to drive the development of artificial intelligence technology, offering new possibilities of brain-like modeling for more efficient and reliable intelligent systems [5, 15].

1.2 Research Objectives

This study aims to explore and reveal the behavior and performance of Deep Neural Networks (DNNs) when dealing with visual illusions, and to explore and enhance the

capabilities of DNNs in simulating the human visual system, thus providing references for improving their structure and algorithms. Specifically, the main objectives of the study include:

1. Analyze the response of DNNs to visual illusions: through experiments, observe and record the responses of DNNs to various visual illusion images. Study the sensitivity of DNNs to different types of visual illusions (such as geometric illusions, color illusions), and analyze the types and frequency of errors they make in processing these illusions.

2. Compare the differences between DNNs and the human visual system: compare the responses of DNNs to those of the human visual system in handling the same visual illusions, identifying the main differences in their processing mechanisms. Through psychophysical experiments, obtain performance data of humans in visual illusion tasks and compare it in detail with the outputs of DNNs.

3. Improve the robustness and interpretability of DNNs: based on the performance analysis of DNNs with different characteristics and architectures in processing visual illusions, propose possible improvements to enhance their robustness against anomalous inputs. Especially whether DNNs inherently possess brain-like universality, reflecting common points in the brain-like mechanisms exhibited by DNNs.

1.3 Structure of the Dissertation

This dissertation is mainly divided into 8 chapters, with the main content of each chapter as follows:

- Chapter 1 summarizes the general situation of the current research and the main objectives of the experiments.

- Chapter 2 mainly introduces various backgrounds of the research, such as deep neural networks, visual illusions, and interdisciplinary studies. Also this chapter includes research questions and the main proposal method.

- Chapter 3 describes the work exploring the performance of DNNs in visual illusions.

- Chapter 4 explores the performance of visual illusions under specific visual illusion datasets.

- Chapter 5 shows the performance of DNNs in visual illusions under spatial-temporal and static characteristic.
- Chapter 6 revolves around the mapping relationship between DNNs and the ventral pathway, combining fMRI to explore the regional similarities under different modules of DNNs.
- Chapter 7 discusses based on current findings, especially the real gap between DNNs and the optimal visual model paradigm, potential advantages, and proposed improvements.
- Chapter 8 summarizes all chapters and the outlook for future research work.

Chapter 2

Background, Motivation and Purpose

2.1 Deep Neural Networks

Deep Neural Networks (DNNs) are a subset of machine learning models that have revolutionized various fields, from computer vision to natural language processing. They are characterized by their deep architectures, consisting of multiple layers that learn hierarchical representations of data [1]. This section mainly introduce DNNs, including their types, underlying mechanisms, and applications.

2.1.1 What is Deep Neural Networks?

DNNs are composed of multiple layers of neurons, where each layer transforms the input data through a series of weights and activation functions. The most common type of DNN is the feedforward neural network, which includes architectures such as Multilayer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs) [16].

A feedforward neural network can be mathematically represented as:

$$y = f(x; \theta) \tag{2.1}$$

where x is the input, θ represents the parameters (weights and biases), and f is the function representing the network's layers and activations.

2.1.2 Convolutional Neural Networks

CNNs are designed to process data with a grid-like topology, such as images. They utilize convolutional layers, pooling layers, and fully connected layers to extract spatial hierarchies of features (Fig. 2.1).

The output of a convolutional layer can be expressed as:

$$(f * g)(i, j) = \sum_{m=-k}^{k} \sum_{n=-k}^{k} f(i + m, j + n) g(m, n) \tag{2.2}$$

where:

- $f(i, j)$ is the input feature map at position (i, j),
- $g(m, n)$ is the convolution kernel (or filter) at position (m, n),
- (i, j) denotes the coordinates of the output feature map,
- k is the radius of the convolution kernel, which determines the size of the kernel. For example, if the kernel size is 3×3, then $k = 1$,
- $*$ denotes the convolution operation.

This specific structure enables CNNs to learn spatial hierarchies of features, making them widely used in image classification, object detection, and image segmentation tasks [17].

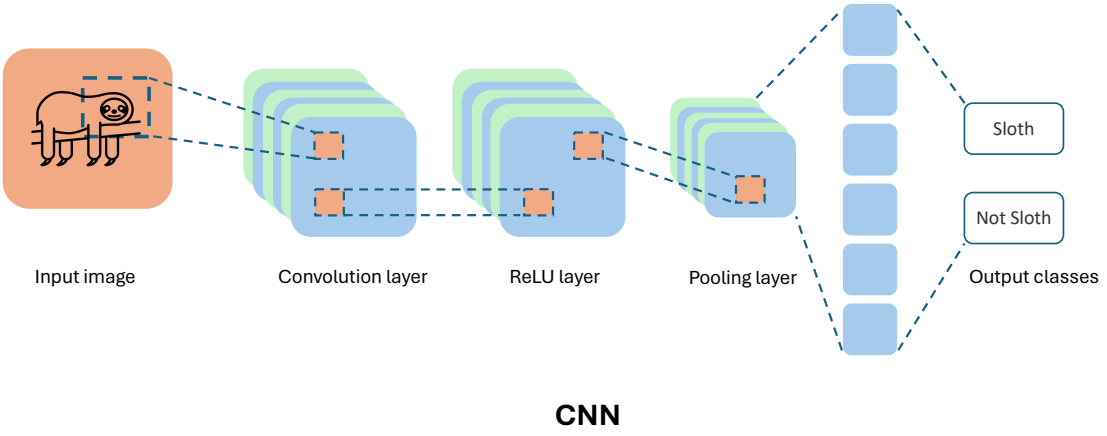


Figure 2.1: The architecture of CNN.

2.1.3 Recurrent Neural Networks and Long Short-Term Memory

For sequential data and video analysis, DNNs incorporate temporal dimensions to model dependencies and dynamics over time. This allows the network to capture not only spatial features but also the temporal evolution of those features (Fig. 2.2).

As a classical model in DNNs for temporal characteristics, Recurrent Neural Networks (RNNs) are designed to handle sequential data by maintaining a hidden state that captures information from previous time steps. The hidden state h_t at time step t is updated as:

$$h_t = \sigma(W_h h_{t-1} + W_x x_t + b) \tag{2.3}$$

where:

- σ is an activation function,
- W_h and W_x are weight matrices,
- b is a bias term.

As for LSTMs [7], a type of RNN, address the vanishing gradient problem by incorporating memory cells and gating mechanisms. The cell state c_t and hidden state h_t are updated through gates that control the flow of information:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{2.4}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{2.5}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{2.6}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{2.7}$$

$$h_t = o_t \odot \tanh(c_t) \tag{2.8}$$

where:

- i_t is the input gate,
- f_t is the forget gate,
- o_t is the output gate,
- $\odot$ denotes element-wise multiplication.

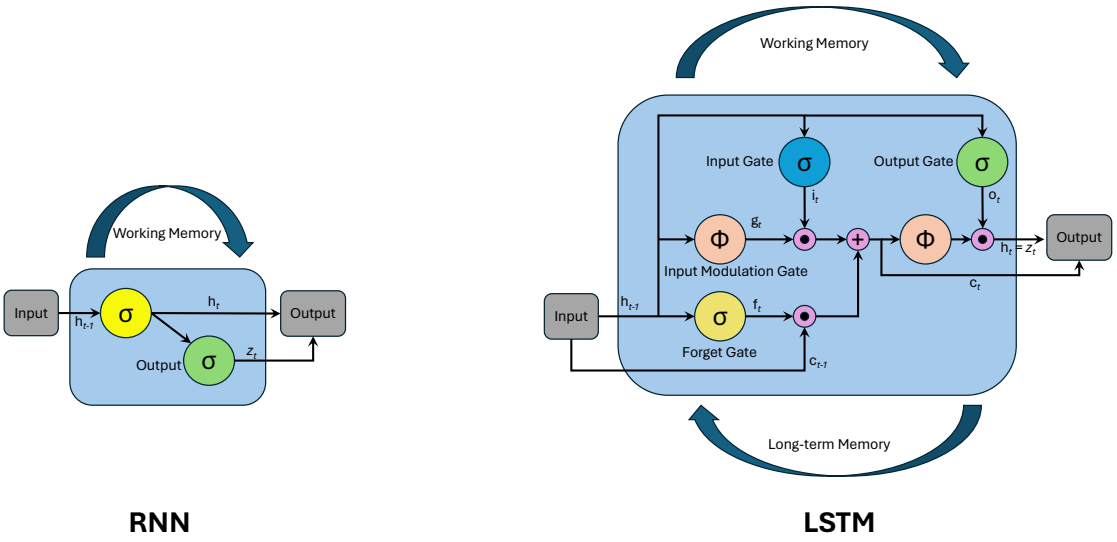


Figure 2.2: The architecture of RNN and LSTM.

2.1.4 3DCNN

3D CNNs extend traditional CNNs to spatiotemporal data by applying 3D convolutions, capturing features across both spatial and temporal dimensions [18, 19] (Fig. 2.3). The output of a 3D convolutional layer is given by:

$$(f * g)(t, x, y) = \sum_{a=-\infty}^{\infty} \sum_{b=-\infty}^{\infty} \sum_{c=-\infty}^{\infty} f(a, b, c) g(ta, xb, yc) \tag{2.9}$$

where:

- $f(t, x, y)$ is the input feature map at time t and spatial position (x, y),
- $g(a, b, c)$ is the 3D convolution kernel at position (a, b, c),
- (t, x, y) denotes the coordinates of the output feature map,
- k is the radius of the convolution kernel in each dimension, which determines the size of the kernel. For example, if the kernel size is $3 \times 3 \times 3$, then $k = 1$,
- $*$ denotes the convolution operation.

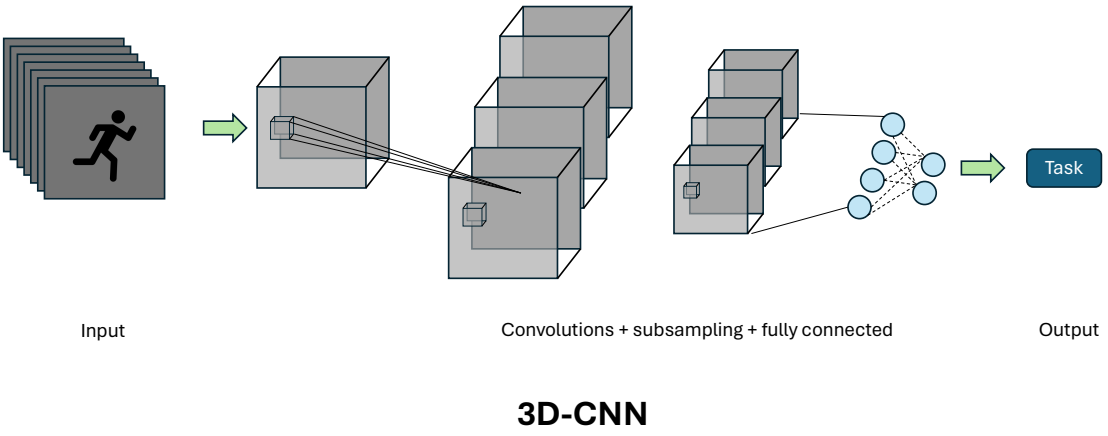


Figure 2.3: The architecture of 3D-CNN.

2.1.5 PredNet

PredNet is a type of DNN designed for predictive coding in video sequences [20] (Fig. 2.4). It consists of layers that predict the input at the next time step and then calculate the prediction error. The prediction error is used to update the model, enabling it to learn temporal dependencies in the data.

The model operates by minimizing the prediction error:

$$E_t = \|x_t \hat{x}_t\|^2 \tag{2.10}$$

where x_t is the actual input at time t and $\hat{x}_t$ is the predicted input.

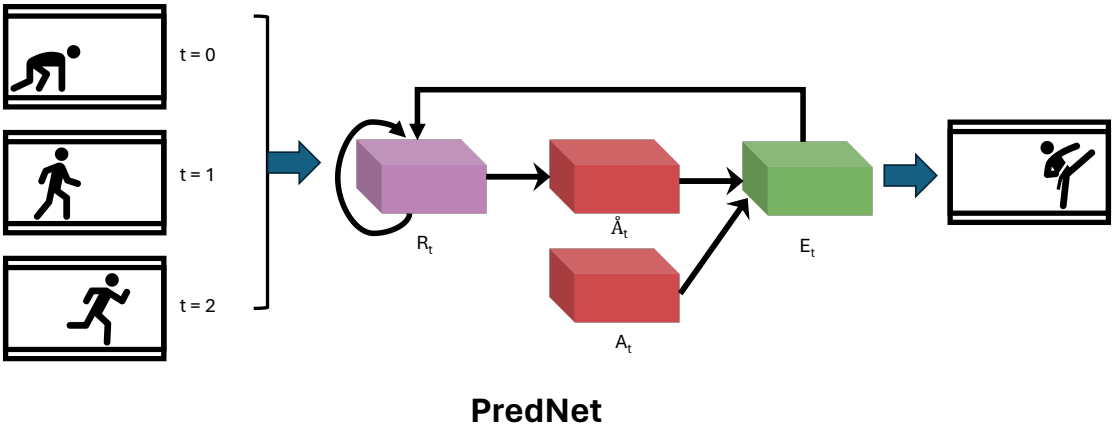


Figure 2.4: The architecture of PredNet.

2.1.6 Transformers

Transformers, initially designed for natural language processing, have been adapted for various applications, including image and video analysis [8]. They rely on self-attention mechanisms to capture long-range dependencies in the data (Fig. 2.5).

The self-attention mechanism computes the output as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.11}$$

where Q (query), K (key), and V (value) are matrices derived from the input, and d_k is the dimensionality of the key vectors.

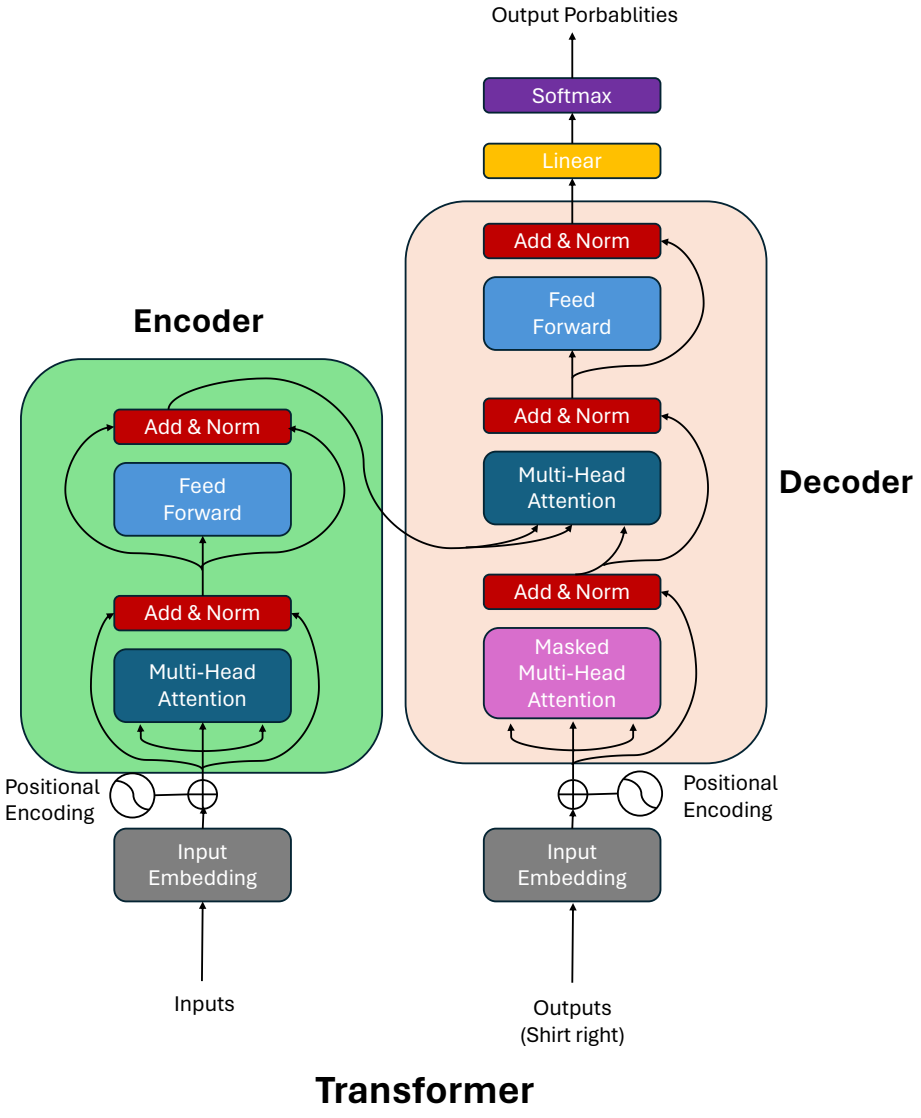


Figure 2.5: The architecture of Transformer.

2.1.7 Applications of Deep Neural Networks

DNNs have found applications across various domains:

- Computer Vision: Image classification, object detection, and segmentation.
- Natural Language Processing: Machine translation, sentiment analysis, and text generation.
- Video Analysis: Action recognition, video summarization, and anomaly detection.

2.2 Optical Illusion

2.2.1 Types of Optical Illusion

Optical illusions can be broadly categorized into three types: literal illusions, physiological illusions, and cognitive illusions [12, 21, 22] (Fig. 2.6). Literal illusions create images that differ from the objects that make them, physiological illusions are the effects on the eyes or brain of excessive stimulation of a specific type (e.g., brightness, tilt, color), and cognitive illusions are the result of unconscious inferences. Each type provides unique insights into how our brain processes visual information. The three specific categories of optical illusions are as follows:

- Literal Illusions: These illusions create images that are different from the objects that make them. A classic example is the Necker cube, a wireframe drawing that can be perceived in two different orientations. This type of illusion highlights the brain's ability to switch between different interpretations of an image.
- Physiological Illusions: These arise due to the physiological responses of the eyes or brain to certain stimuli. Examples include the Hermann grid illusion, where ghostly grey blobs appear at the intersections of a white grid on a black background, and the Mach bands, where exaggerated contrast between edges of slightly differing shades of gray is perceived. These illusions shed light on the way our visual system enhances contrast and edges to improve object detection.
- Cognitive Illusions: These illusions occur because of the brain's unconscious inferences about the world. The Müller-Lyer illusion is a well-known example, where lines of equal length appear unequal due to the orientation of arrowheads at their ends. Another example is the Ames room illusion, where the distorted shape of

the room causes people to appear to change size as they move through it. Cognitive illusions demonstrate the brain's reliance on contextual information and prior knowledge to interpret sensory input.

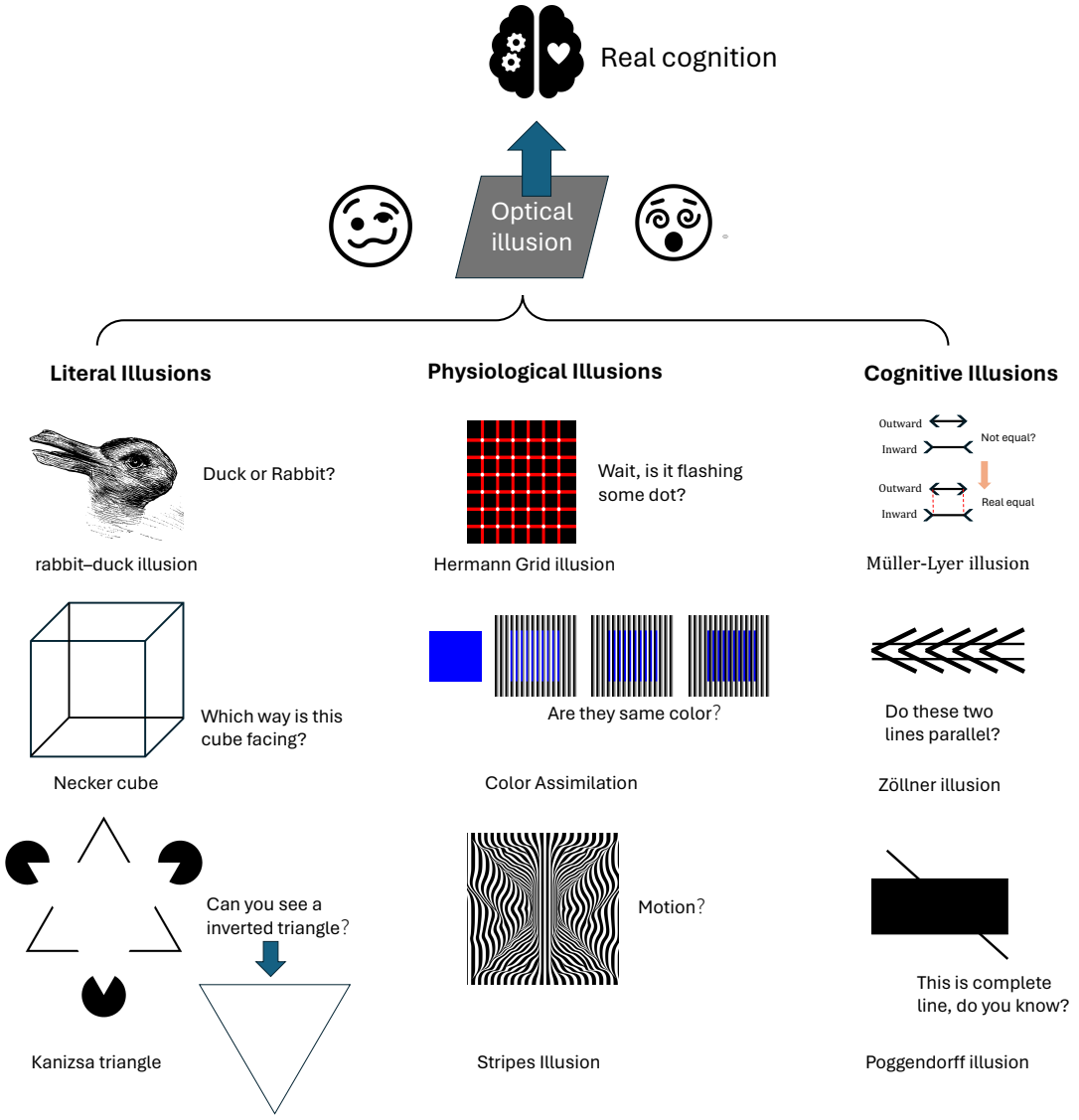


Figure 2.6: What is optical illusion? The main types of optical illusion.

2.2.2 Mechanisms Behind Optical Illusion

The study of optical illusions provides valuable insights into the cognitive and neural mechanisms of vision. Cognitive theories suggest that illusions occur because the brain interprets sensory input based on past experiences and expectations. The Gestalt principles of perception, such as similarity, proximity, and continuity, explain how we group visual elements and perceive patterns [23].

From a neural perspective, illusions are studied by examining the activity in various parts of the brain. Neuroimaging studies have shown that different types of illusions activate specific regions of the visual cortex and other related areas [24]. For example, the primary visual cortex (V1) processes basic visual features, while higher-level areas like the lateral occipital cortex (LOC) are involved in shape and object recognition [25]. The brain often prioritizes contextual information over raw sensory input, leading to perceptual discrepancies [26].

2.2.3 Implications of Optical Illusion

Optical illusions have significant implications for various fields, including neuroscience, psychology, and artificial intelligence. In neuroscience and psychology, illusions are used to probe the complexities of visual perception and the brain's interpretative processes [27]. They help researchers understand how the brain constructs a coherent representation of the world from ambiguous and often incomplete sensory information [28].

In artificial intelligence, particularly in the development of deep neural networks (DNNs), studying optical illusions can provide insights into the limitations and capabilities of these models [29]. DNNs, especially convolutional neural networks (CNNs), are designed to mimic human visual processing by learning hierarchical representations of visual data. However, similar to humans, these models can also be susceptible to optical illusions, revealing their interpretative strategies and potential areas for improvement [30].

2.3 Interdisciplinary Study

2.3.1 Ventral Pathway

The ventral pathway, often referred to as the "what pathway," is a crucial part of the human visual system responsible for object recognition and form representation [31, 32]. It extends from the primary visual cortex (V1) to the inferior temporal cortex (IT), passing through areas such as V2 and V4. Main introduction as follows:

- Primary Visual Cortex (V1): The initial stage of visual processing, where basic features such as edges and orientations are detected.
- Secondary Visual Cortex (V2): Processes information from V1, including more complex features and patterns.

- Ventral Stream (V4): Involved in processing color and simple geometric shapes.
- Inferior Temporal Cortex (IT): Responsible for high-level object recognition and visual memory.

Generally, there are two visual pathway influence each other then generate the cognition and motion, Fig. 2.7 is the example of hypothesis of double-pathway on explaining the activity. When we see a cup, then want to do the "pouring" this motivation, first we recognize the object "cup" and know catch the handle and do the "pouring" [32].

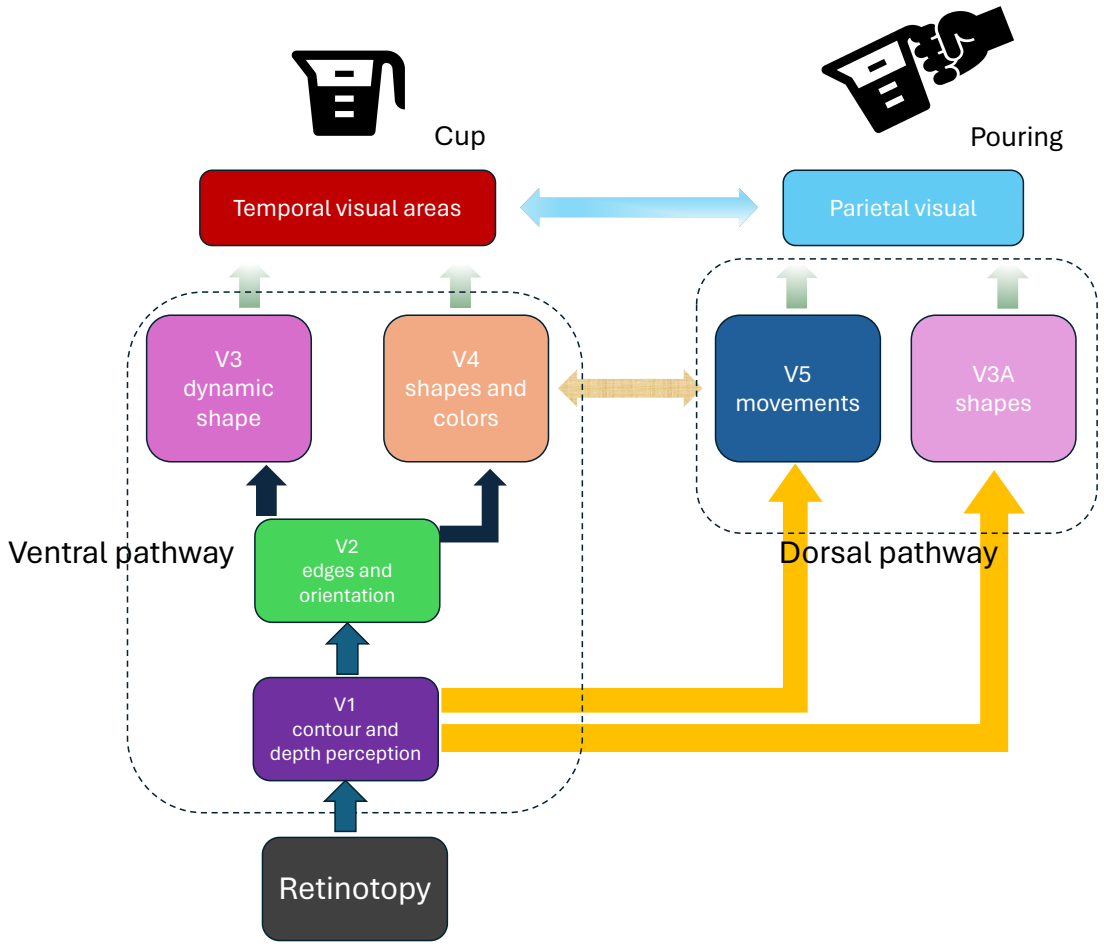


Figure 2.7: An example of recognition and motion through two visual pathway.

2.3.2 Mapping DNNs to Ventral Pathway

Deep neural networks (DNNs), particularly convolutional neural networks (CNNs), have shown remarkable parallels to the ventral stream [33, 34]. Various layers of CNNs have been found to correspond to different stages of the ventral pathway (Fig. 2.8):

- Early Layers: Analogous to V1 and V2, detecting basic features such as edges and textures.
- Intermediate Layers: Correspond to V4, processing more complex features and shapes.
- Deep Layers: Similar to IT, responsible for high-level object and scene recognition [34].

Understanding the mapping relationship between DNNs and the ventral pathway enhances the interpretability of these models and provides insights into biological vision mechanisms.

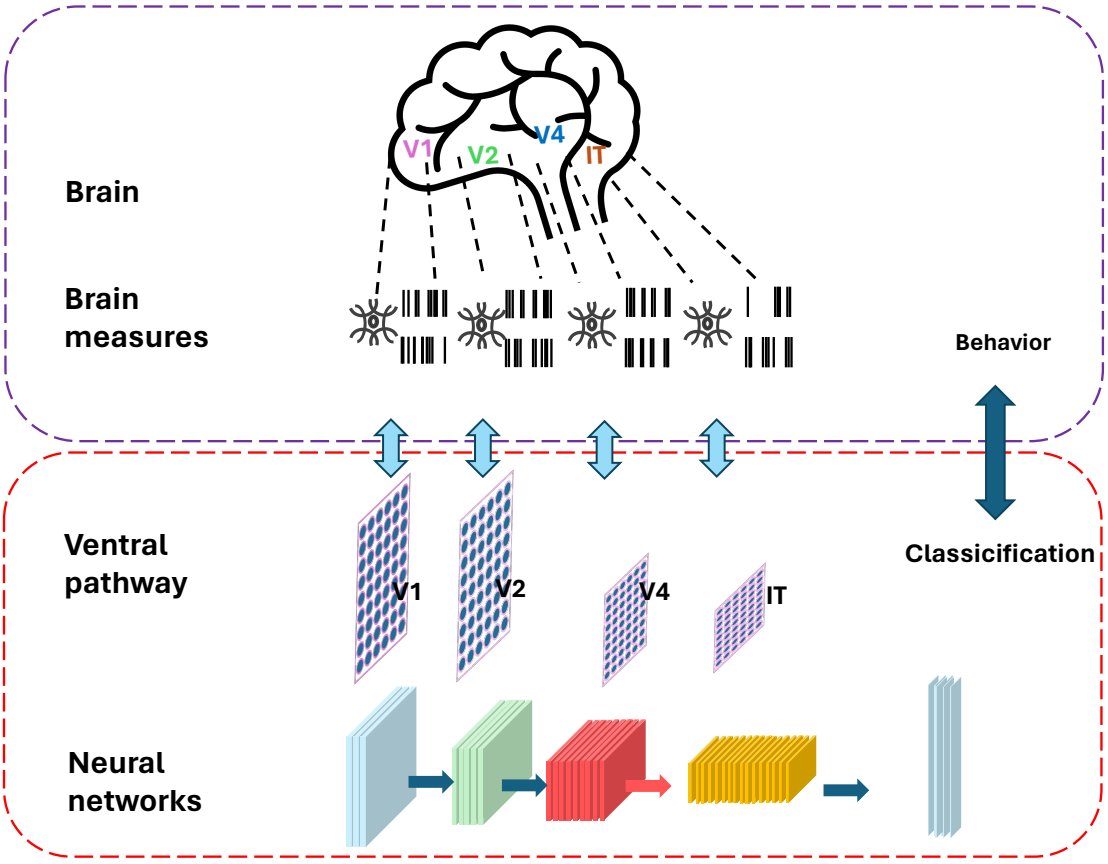


Figure 2.8: Mapping relationship between DNNs and ventral pathway.

2.4 Why DNNs with Optical Illusion?

Currently, most research on deep neural networks (DNNs) focuses on improving task performance, thereby enhancing test rates. This overlooks inherent issues with the

models: they perform poorly in terms of real-world robustness and often struggle to model and understand subjective and abstract concepts [2, 35]. Particularly, DNNs rely on statistical pattern recognition from large-scale data, primarily using methods like supervised and reinforcement learning, which require extensive labeled data or learning through interaction with the environment. However, deeper issues such as emotions, consciousness, and motivations often lack clear labels or are hard to learn through simple reward signals [36, 37].

More attention needs to be given to the inherent modeling and interpretability of DNNs rather than solely focusing on enhancing model performance. As a branch of this direction, exploring the visual pathway mappings exhibited by DNNs, as well as their exceptional performance in various visual tasks, is a viable method. In essence, by utilizing the limitations and characteristics of the human visual perception system in processing visual information, we can infer whether DNNs exhibit similar behaviors. This complementary relationship can be described as follows: DNNs learn and simulate the human brain's processing methods through multi-level complex connections, while the cognitive understanding and behavioral patterns of the human brain guide and improve the design of deep neural networks [38]. This inspiration and influence are manifestations of the integration and development in neuroscience and artificial intelligence, bringing new understanding and innovative possibilities to both fields (Fig. 2.9). Nevertheless, DNNs also face many controversies, particularly regarding their biological relevance. Kubilius et al., (2016) [39] has found that DNN models trained for image classification can explain some aspects of biological vision well. But some researchs also found the important divergences between DNNs and biological vision [40, 41]. The greatest strength of DNNs as visual learning models lies in their exploratory nature [33, 42, 43]. Moreover, some studies indicate that there are corresponding relationships between the layered visual areas and DNN layers in terms of visual feature representation, with deeper stages of DNNs showing similar perceptual levels to the brain [33, 44]. Therefore, in our research, we are inclined to use DNNs to seek and compare the differences and collaborative relationships in human visual aspects and explore the possibilities of their mechanisms. In other words, DNNs and the brain should be considered as two similar biological entities.

2.4.1 Related Research and Identification of Gaps

In the illusion research on DNNs, some relevant progress has already been made. As early as 2018, Watabave et al., (2018) [45] discovered through the PredNet test of rotating snake illusions that DNNs might exhibit motion illusions. Also, Benjamin et al., (2020) [46] found that lower-level modules are more prone to color illusions, such as color assimilation. Additionally, DNNs have shown the presence of visual illusions in phenomena like the Hermann Grid and the Müller-Lyer illusion [47, 48]. Generally, such studies use human subjects' perceptual data as a benchmark to design related visual illusion tasks and infer and test the responses of DNNs. However, this approach overlooks an important issue: relying solely on perceptual data to test is insufficient to explain the relationship between DNNs and visual pathways. A deeper exploration of the internal mechanisms and broader consideration of diversity issues are required. This involves interpretability AI, DNNs themselves, and the types of visual illusions. In terms of visual illusions, a single illusion cannot explain the widespread phenomenon of visual illusions in DNNs, and more types of visual illusions need to be considered. Previous related studies have all considered single visual illusions, neglecting the universality issue of different visual illusions. As for DNNs themselves, it is necessary to consider models with different architectures and the training of the models themselves, while current research only tests single or a few models. More brain-like DNNs need to be considered for training and testing. Furthermore, merely testing is far from enough to understand why DNNs exhibit visual illusions and how decisions are made when illusions occur. This requires the use of interpretability AI methods [49] to explain the potential mechanisms more intuitively.

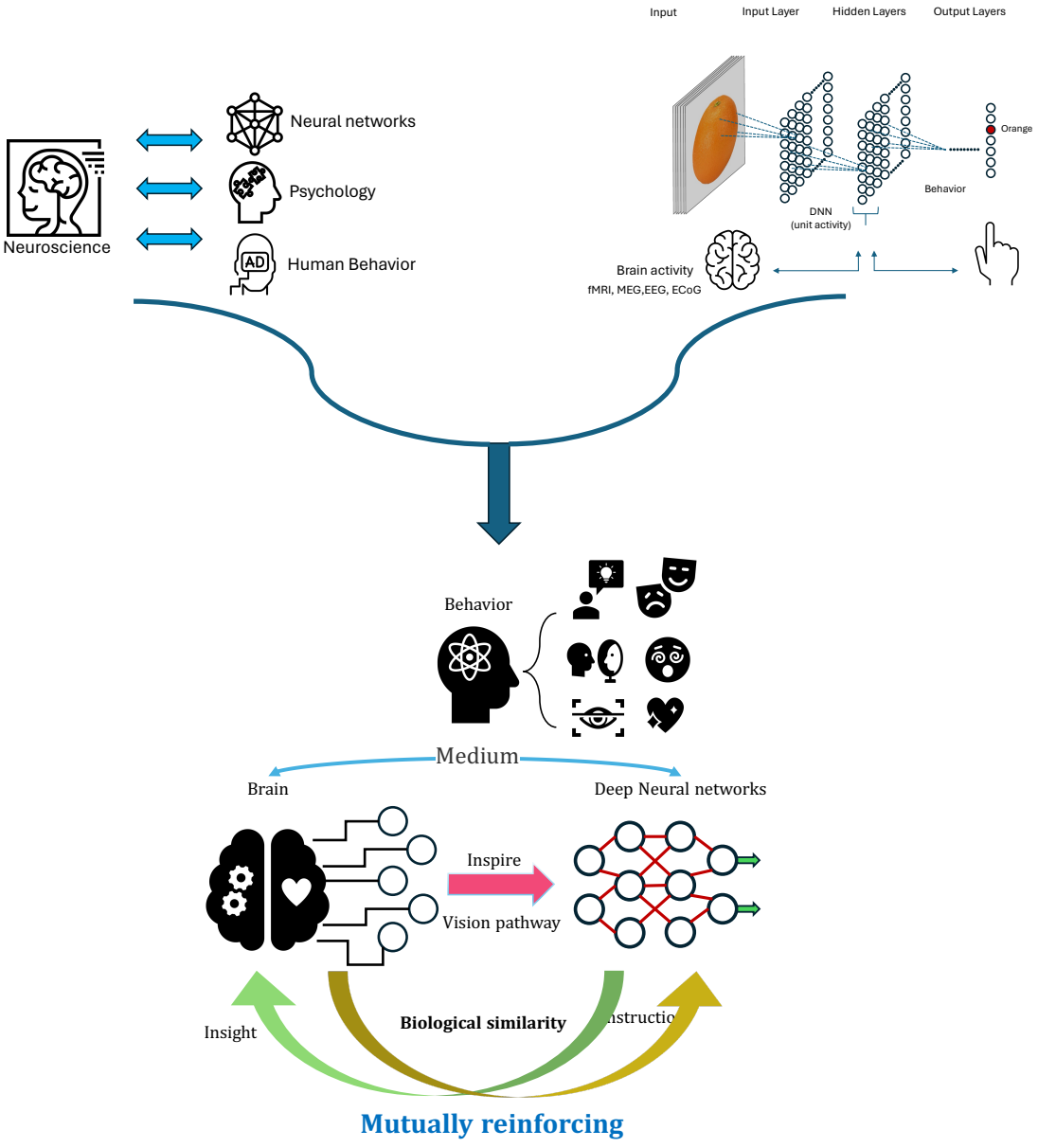


Figure 2.9: Interdisciplinary study between DNNs and Neuroscience

2.4.2 How do DNNs Respond to Different Types of Visual Illusions?

We focuses on evaluating the performance of various DNN models in simulating multiple visual illusions. By examining responses to illusions such as the Müller-Lyer, color assimilation, Hermann grid, Zöllner, and Poggendorff illusions and Ske's Oblique Grating illusion, we aim to determine the extent to which these models replicate human perceptual experiences.

2.4.3 What Are the Similarities and Differences Between Human and DNN Perceptions of Visual Illusions?

By comparing human perceptual data with DNN responses, we seek to identify both commonalities and discrepancies. This comparison will help us understand how closely DNNs mimic human visual processing and where they diverge.

2.4.4 How Do Different DNN Architectures Compare in Their Ability to Simulate Visual Illusions?

This question involves benchmarking various DNN models to assess their effectiveness in replicating visual illusions. By evaluating models with different architectures, we aim to identify which types of networks are more adept at simulating human-like visual perception.

2.4.5 What Are the Computational Principles Underlying the DNNs' Ability to Simulate Visual Illusions?

Through in-depth analysis, we aim to uncover the computational strategies that DNNs use to process and respond to visual illusions. Understanding these principles can provide insights into the fundamental mechanisms of visual cognition in both artificial and biological systems.

2.4.6 Can the Findings From DNN Simulations of Visual Illusions Inform the Development of More Advanced AI Systems?

By leveraging insights gained from studying DNN responses to visual illusions, we aim to contribute to the development of AI systems with enhanced visual processing capabilities. This includes improving the interpretability, robustness, and accuracy of these systems in complex visual tasks.

2.4.7 Meaning and Contribution

These questions guide our investigation into the intersection of human visual perception and artificial neural networks, aiming to bridge the gap between biological and artificial vision systems. By addressing these questions, we seek to advance our understanding of both the potential and limitations of DNNs in simulating complex visual phenomena.

2.5 Proposed Method

2.5.1 Human Perceptual Data on Optical Illusion

Human perceptual data is crucial for studying optical illusions and understanding how visual information is processed and interpreted. This data can be collected through psychophysical experiments where participants are asked to judge and report their perceptions of various visual stimuli. The main procedure can be set as follows (Fig. 2.10):

- Experimental Setup: Participants view images or videos containing optical illusions and provide responses regarding their perceptions.
- Measurement: Quantitative data on perceived lengths, angles, and shapes are recorded.
- Analysis: Statistical methods are used to analyze the perceptual data, identifying patterns and biases in human perception.

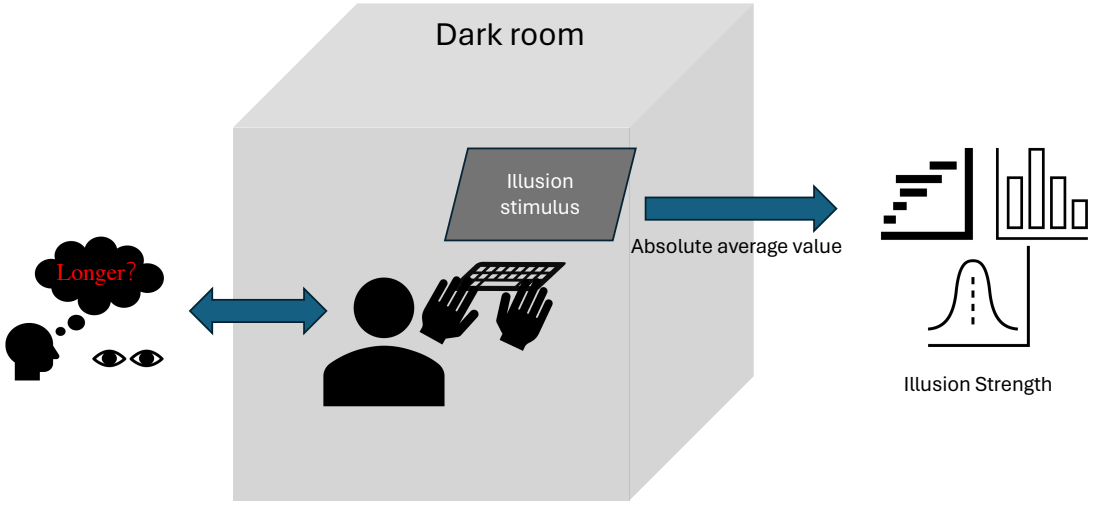


Figure 2.10: Human subject experiment

2.5.2 Representational Similarity Analysis

Representational similarity analysis (RSA) [50] is a method used to compare the similarity of representations between different systems, such as human brain activity and DNN activations.

Feature Extraction: Extract feature vectors from DNNs and human perceptual data. For DNNs, the activations from a specific layer are used as feature vectors. For human perceptual data, behavioral responses or neuroimaging data can be used.

Let $\mathbf{f}_i$ represent the feature vector of the i-th stimulus from DNNs, and $\mathbf{p}_i$ represent the feature vector from human perceptual data.

RDM Construction: Construct Representational Dissimilarity Matrices (RDMs) for both DNNs and human perceptual data. The RDMs are typically constructed by calculating pairwise dissimilarities (e.g., Euclidean distance) between feature vectors (Fig. 2.11).

For DNNs:

$$RDM_{DNN}(i,j) = \|\mathbf{f}_i\mathbf{f}_j\| \tag{2.12}$$

For human perceptual data:

$$RDM_{human}(i,j) = \|\mathbf{p}_i\mathbf{p}_j\| \tag{2.13}$$

Similarity Measures: Compare the RDMs using similarity measures such as distance metrics like Euclidean distance [51].

$$\rho = \frac{\text{cov}(RDM_{DNN}, RDM_{human})}{\sigma_{RDM_{DNN}}\sigma_{RDM_{human}}} \tag{2.14}$$

where:

- ρ is the correlation coefficient between the two RDMs,
- $\text{cov}(RDM_{DNN}, RDM_{human})$ is the covariance between the RDMs of the DNN and human perceptual data,
- $\sigma_{RDM_{DNN}}$ is the standard deviation of the RDM for the DNN,
- $\sigma_{RDM_{human}}$ is the standard deviation of the RDM for the human perceptual data.

By applying RSA, researchers can determine how well DNNs replicate human perceptual patterns when viewing optical illusions. This helps in understanding the mechanisms underlying both artificial and biological vision systems.

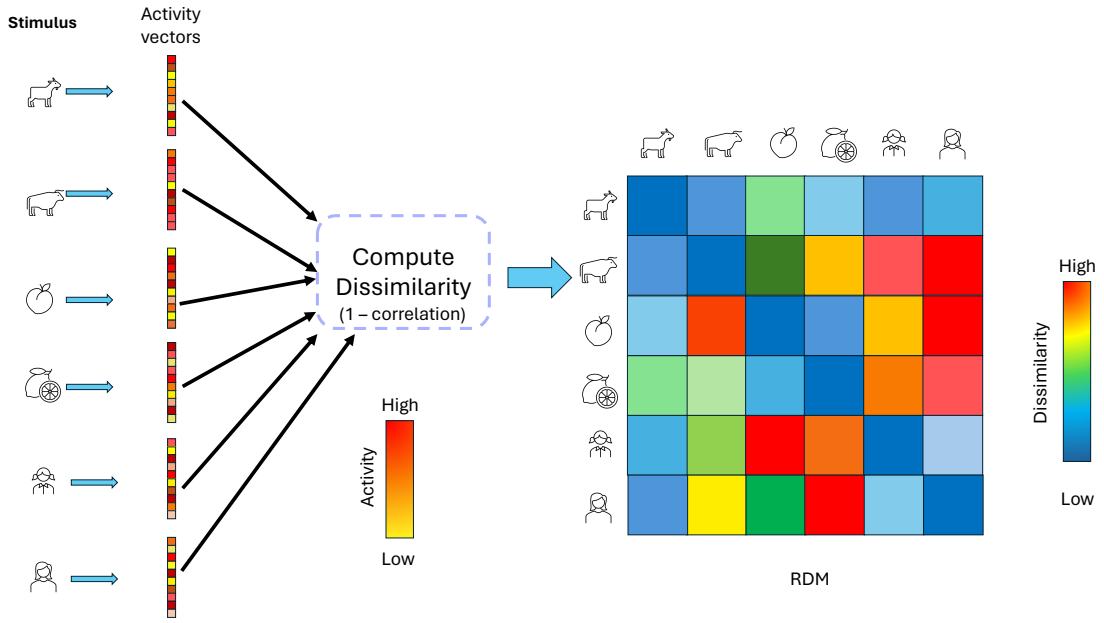


Figure 2.11: RDM

2.5.3 Class Activation Mapping

Class Activation Mapping (CAM) [52] is a technique used to visualize which regions of an input image are important for a neural network's classification decision. The main equation can be expressed as follows:

$$S^c = \sum_k w_k^c \sum_{x,y} A^k(x,y) \tag{2.15}$$

Here, $A^k(x,y)$ represents the activation of unit k at spatial location (x,y) in the activation map, and w_k^c are the weights corresponding to class c.

To generate heatmaps, the Class Activation Map CAM^c for class c is given by:

$$\text{CAM}^c(x,y) = \sum_k w_k^c A^k(x,y) \tag{2.16}$$

The heatmap is then normalized and superimposed on the input image to visualize the important regions. The main idea and process can be shown in Fig. 2.12

Totally, CAM can be used to identify which parts of an optical illusion image influence the network's perception. This provides insights into the network's decision-making process and helps in understanding how DNNs interpret complex visual stimuli.

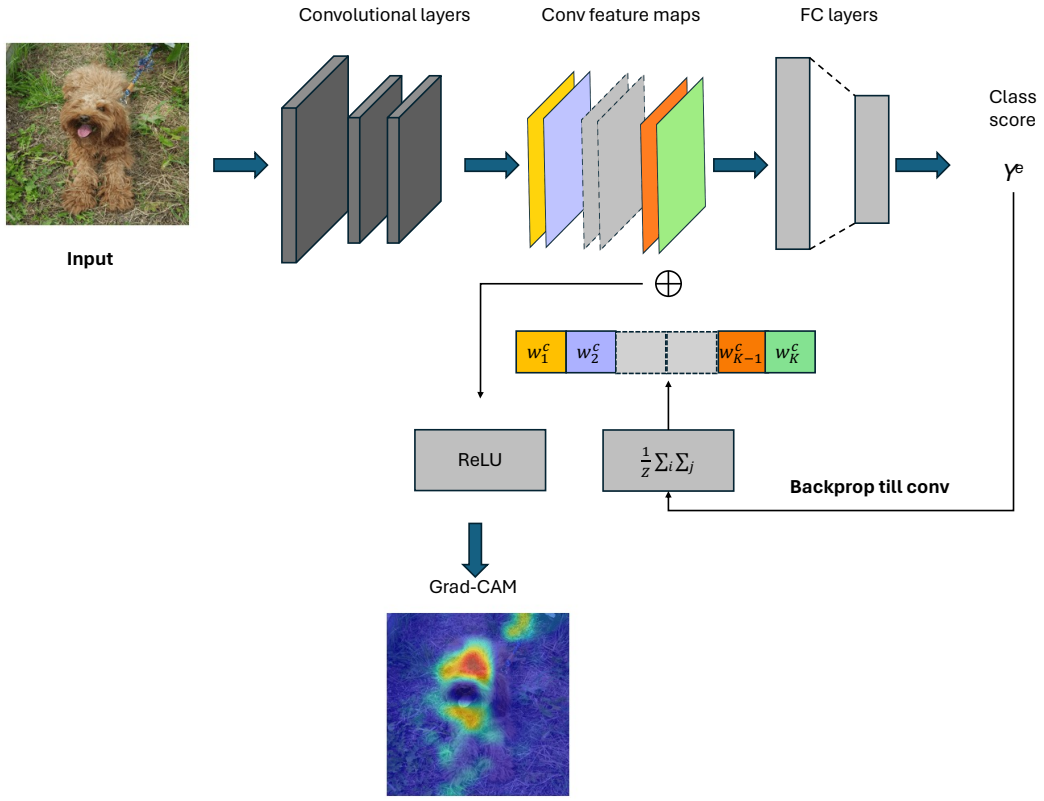


Figure 2.12: Grad-CAM

2.5.4 The Main Purpose, Idea and Method

Generally, our approach is also based on perceptual data from human subjects, which is a benchmark for studies, obtained from the most direct behavioral data processed by the brain. We proposed interpretative visualization methods based on RDM and CAM to elucidate the internal mechanisms of DNNs and the reasons for their decision dependencies and tendencies. Here is proposed visualization framework:

We present a novel framework to investigate how Deep Neural Networks (DNNs) process visual illusions, utilizing the combined strengths of Grad-CAM and the Representational Dissimilarity Matrix (RDM). This framework is designed to provide a comprehensive understanding of DNN responses to visual illusions, expressed in the following equation:

$$S_{DNN} = F(G_c(I), R_{L2}(I_{perceived}, I_{real})) \tag{2.17}$$

The components are defined as follows:

- S_{DNN} represents the sensitivity or response measure of DNNs to visual illusions.
- F is a synthesis function that combines the feature region visualization from Grad-CAM and the representational differences from RDM.
- $G_c(I)$ denotes the Grad-CAM heatmap for class c for a given image I, highlighting how the DNN focuses on specific areas of the image for decision-making.
- $R_{L2}(I_{perceived}, I_{real})$ quantifies the representational difference between human perceived images and real images using the Euclidean (L2) distance, as derived from RDM.

This framework enables the analysis of DNN responses to visual illusions by quantifying both internal representations and external responses to illusion images. Through $G_c(I)$, we can explore the differences in attention distribution in DNNs between illusion images and human perceptual images. The R_{L2} calculation quantifies perceptual differences, indicating how DNNs differentiate between actual physical attributes and human perception. The function F integrates these analyses, offering a comprehensive view of DNN processing of visual illusions.

Based on this framwork, then we referenced and extensively used various types of DNNs, including training strategies under pre-training and specific training, to comprehensively explore visual illusions in DNNs research, providing more potent ideas and insights for future brain-like modeling of neural networks, as well as potential directions for improvement. The research mainly consists of the following steps:

1. First, explore whether there are visual illusions in DNNs [47, 48, 53].
2. Then, investigate the impact of specific visual illusion datasets and training on visual illusions [54].
3. Next, deeply consider various types of DNNs, especially the performance of visual illusions in DNN models under the characteristics of both temporal and static.

4. Based on the popular mapping relationships and findings from previous steps, explore the relevance of mapping relationships through designing fMRI studies based on visual illusions.

5. Discussion and Conclusion.

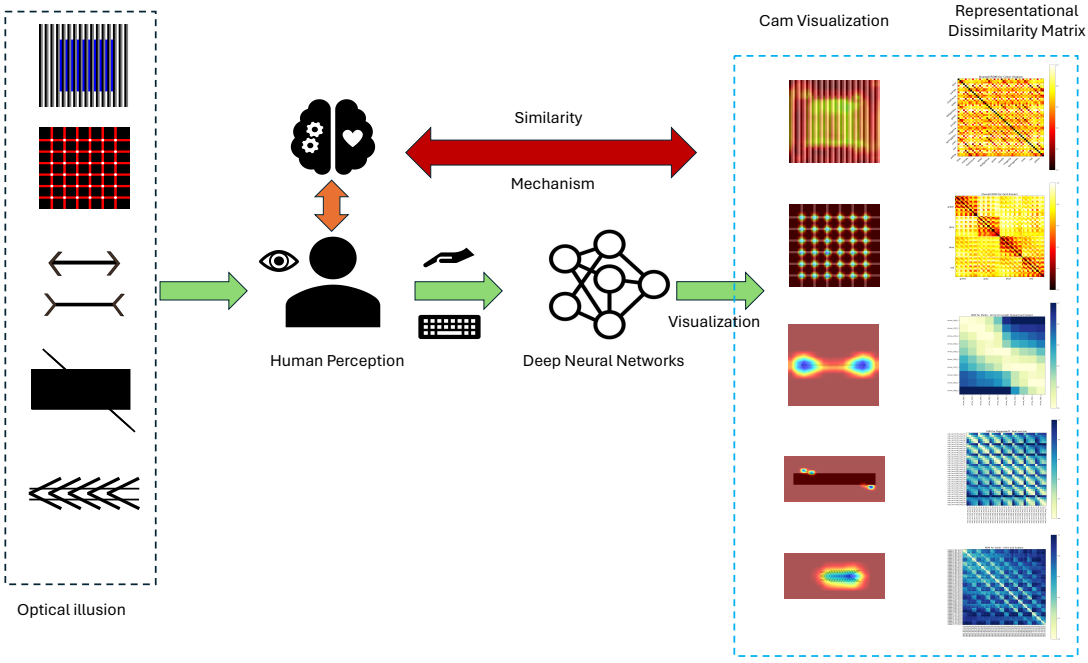


Figure 2.13: Main research processes

Chapter 3

Do the Illusion Performed in DNNs?

This chapter mainly introduces the universality of optical illusion, which set by two steps:

1. The testing of the Müller-Lyer illusion in DNNs. Around the perceived lengths of human subjects, it preliminarily explores and verifies whether DNNs exhibit visual illusions [47, 53].

2. Exploring the performance of Muilti-type illusion on DNNs [48].

3.1 Step1 :Müller-Lyer Illusion on Vgg19 and ResNet101

3.1.1 Human Experiment on Müller-Lyer Illusion

The Müller-Lyer illusion is a classic visual illusion that demonstrates how the perception of line length can be influenced by changing the direction of arrows at the ends of the lines. When the ends of the line have outward-facing arrows, the line appears shorter than its actual length; when the ends have inward-facing arrows, the line appears longer than its actual length. This illusion reveals that the surrounding contextual information significantly influences the perception of length in the visual system's processing of geometric shapes [12]

Studies have shown that the Müller-Lyer illusion is not only related to the visual processing of geometric shapes but also involves the brain's interpretation of three-dimensional space. Gillam (1998) [55] pointed out that this illusion may originate from

our understanding of spatial depth and perspective, where the shape of arrows provides clues about distance and spatial layout, thus affecting the perception of length.

To collect data on length perception, we designed ten length levels ranging from 200 to 380 pixels, with intervals of 20 pixels. In the experiments, a line without arrows randomly appeared at the top or bottom of the screen, while a line with arrows of random length appeared in the center (Fig. 3.1). Participants adjusted the length of the arrowed line to make it appear equal to the line without arrows, using the AS and KL keys on the keyboard for fine adjustments ($\pm$1 pixel) and larger adjustments ($\pm$10 pixels). Each experimental group included 40 trails, and the entire experiment was conducted in four rounds. The collected data covered lengths from 200 to 380 pixels, including perceived lengths for both arrow directions. The experiments were conducted in a dark room using a Pixio PX248 Prime monitor (resolution 1920 $\times$ 1080) with stimuli generated through Python's Psychopy software. The distance between participants and the monitor was maintained at 64 centimeters.

Then we averaged the collected perceptual length data to obtain mean perceived lengths, which will be used in pictorial form for subsequent testing phases. We set up a perceptual group and a control group for testing in these 10 length label(200 to 380), where the perceptual group was based on the adjusted perceived lengths, while the actual lengths of the lines in the control group were the same, only the arrow directions were added.

A total of 12 male volunteers participated in this study, with an average age of 24.58 years (standard deviation 2.72 years), all participants were free from color vision anomalies or other visual defects and had normal vision. This study and later all were approved by the Human Research Ethics Committee of Kochi University of Technology and conducted with written informed consent from all participants, strictly followed relevant ethical guidelines and regulations.

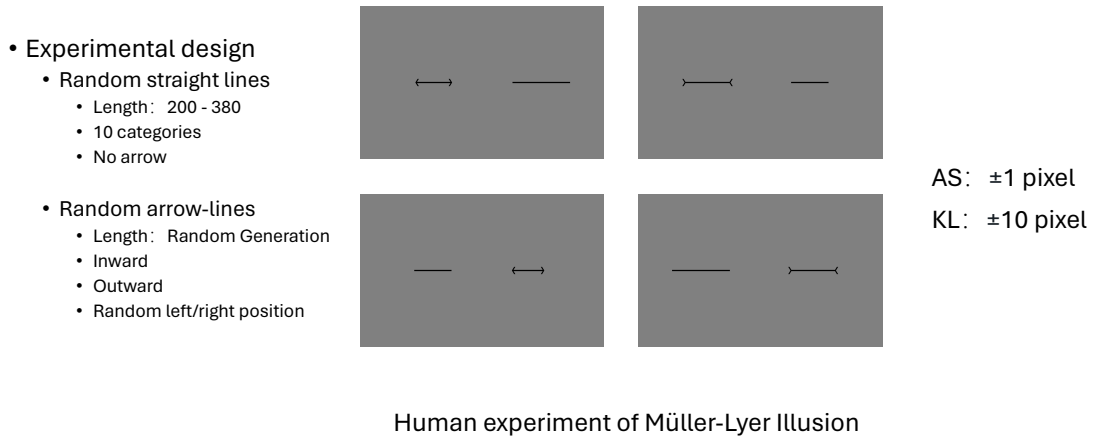


Figure 3.1: The main subject experiment on Müller-Lyer Illusion

3.1.2 Brain-like DNNs

DNNs are diverse, and Schrimpf et al. (2020) [44, 56] has explored the brain-like performance of DNNs based on mapping relationships, proposing the concept and ranking of Brain-Score. Brain-Score is used to evaluate the performance of deep neural networks in simulating brain information processing, ranking and comparing models based on different tasks and datasets. Based on this ranking, we first selected two classic models for testing: Vgg19 [57] and ResNet101 [9].

- Vgg19 is a deep convolutional neural network consisting of 19 layers, mainly composed of 16 convolutional layers and 3 fully connected layers. It has shown remarkable effects on the ImageNet dataset [58]. Studies have indicated that the activation patterns of certain intermediate layers of Vgg19 have high similarities with the neural activities of the human visual system, particularly sensitive to details and local features in images when processing visual tasks, similar to the human visual system's reliance on these processes for object recognition.

- ResNet101 is a deep neural network with residual modules, comprising 101 layers. Its introduced residual blocks solve the problems of gradient vanishing and explosion in deep network training, allowing for deeper network structures. ResNet101 performs excellently in visual tasks such as object recognition and scene classification, and its information processing methods have been found similar to the brain's, particularly in understanding complex scenes and extracting global features, simulating the human brain's mechanisms in processing high-level visual tasks .

Evidence suggests that Vgg19 and ResNet101 display distinct brain-like characteristics in different visual tasks [39, 59]. Vgg19 is better at simulating the human visual system's processing methods when recognizing objects in complex scenes. This is evident in its higher sensitivity to details and local features within images, which are crucial for the human visual system when recognizing objects. On the other hand, due to its deep structure, ResNet101 excels in understanding complex scenes and extracting global features, similar to how the human brain processes high-level visual tasks.

3.1.3 The Preliminary Test on Two DNNs Models

Both models use pretrained weights, specifically those trained on the ImageNet dataset. Using pretrained models, especially those trained on large datasets like ImageNet, has proven to perform excellently in simulating human brain processing of visual information [4, 60]. These models have undergone extensive training on a wide range of visual tasks, enabling them to exhibit activity patterns similar to those of the human visual cortex when processing new, unseen visual data. By utilizing these pretrained models, we can directly leverage their existing learning achievements to accelerate the validation of their performance in new brain-like studies. Moreover, because these models have been trained on a wide variety of images, they have better generalization capabilities for various visual features, allowing for minimal adjustments to achieve good results in specific tasks.

3.1.4 The Distribution of Human Perceptual Length

The average perceived lengths adjusted by participants for outward and inward facing arrows are shown in Fig. 3.2. From the figure, adjustments for outward-facing arrows are positive, while those for inward-facing arrows are negative, consistent with the principles displayed by the Müller-Lyer illusion. Outward-facing arrows appear shorter, while inward-facing arrows appear longer.

Essentially, participants exhibited a visual illusion of length changes: the line with outward-facing arrows (light blue) required a greater adjustment in length than the actual corresponding line length to appear visually equal in length. Similarly, the line with inward-facing arrows (orange) required a shorter adjustment than the actual corresponding length. Additionally, the average changes for both directions are inconsistent,

with the average sensory length change for inward-facing arrows around 10 pixels, while for outward-facing arrows, it is around 15 pixels.

Based on this distribution, the corresponding perceived lengths are used as the dataset for the perception group. In contrast, the actual lengths of the lines (same length, just arrows added) are used as the dataset for the control group. Both groups correspond to ten length labels.

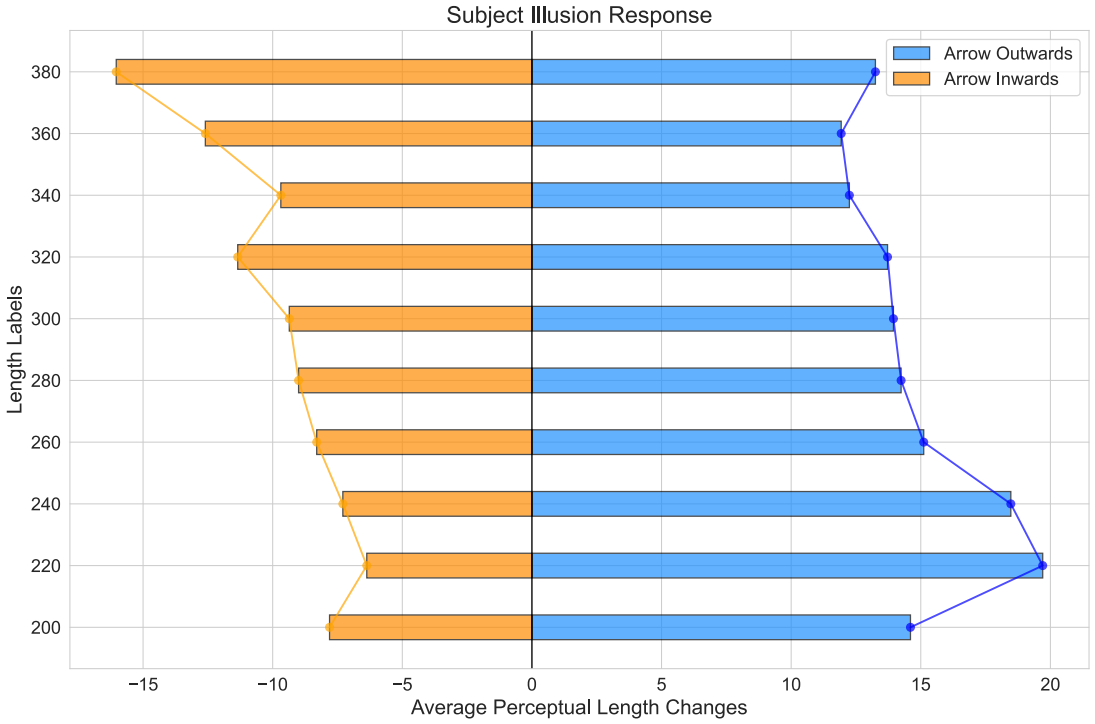


Figure 3.2: The perceptual length on Müller-Lyer Illusion

3.1.5 Representational Dissimilarity Matrix

Before testing the visual illusion performance of Vgg19 and ResNet101, we mainly based our assessment on constructing dissimilarity matrices. Specifically, we extracted feature vectors of lines with outward-facing and inward-facing arrows from the perception group at the last convolutional layer before the decision module of the models, constructing the dissimilarity matrix by calculating the Euclidean distance. The RDM size is 10x10, corresponding to ten length labels. RDMs are displayed as heatmaps; the darker the color, the higher the similarity.

From the RDM heatmaps (Fig. 3.3 "Perceived length"), both models show a similar distribution: the RDM heatmaps of the two models exhibit high similarity along the diagonal. This indicates that for the same length labels, lines with inward-facing

arrows and lines with outward-facing arrows are consistent in both models. The perception group represents the visual length adjustments of human subjects, and the high representational similarity displayed by the models shows that the models exhibit visual illusions similar to humans.

Moreover, we also constructed RDMs using images from the control group for both models (Fig. 3.3 "Real length"). Compared to the high similarity along the diagonal in the perception group, the control group's high similarity distribution appears shifted upwards from the diagonal. In the control group, the line lengths under the same length labels are consistent with the line lengths corresponding to the length labels. Interestingly, the lengths of the lines with outward-facing and inward-facing arrows are the same. The upward-shifted high representational similarity shows that the models consider the lines with outward-facing arrows to be different from those with inward-facing arrows, need to be longer to match the inward-facing arrows. This is consistent with the Müller-Lyer illusion: lines with outward-facing arrows appear shorter, and lines with inward-facing arrows appear longer.

The RDMs for the perception group and control group demonstrate that Vgg19 and ResNet101 exhibit visual illusions similar to those observed in humans.

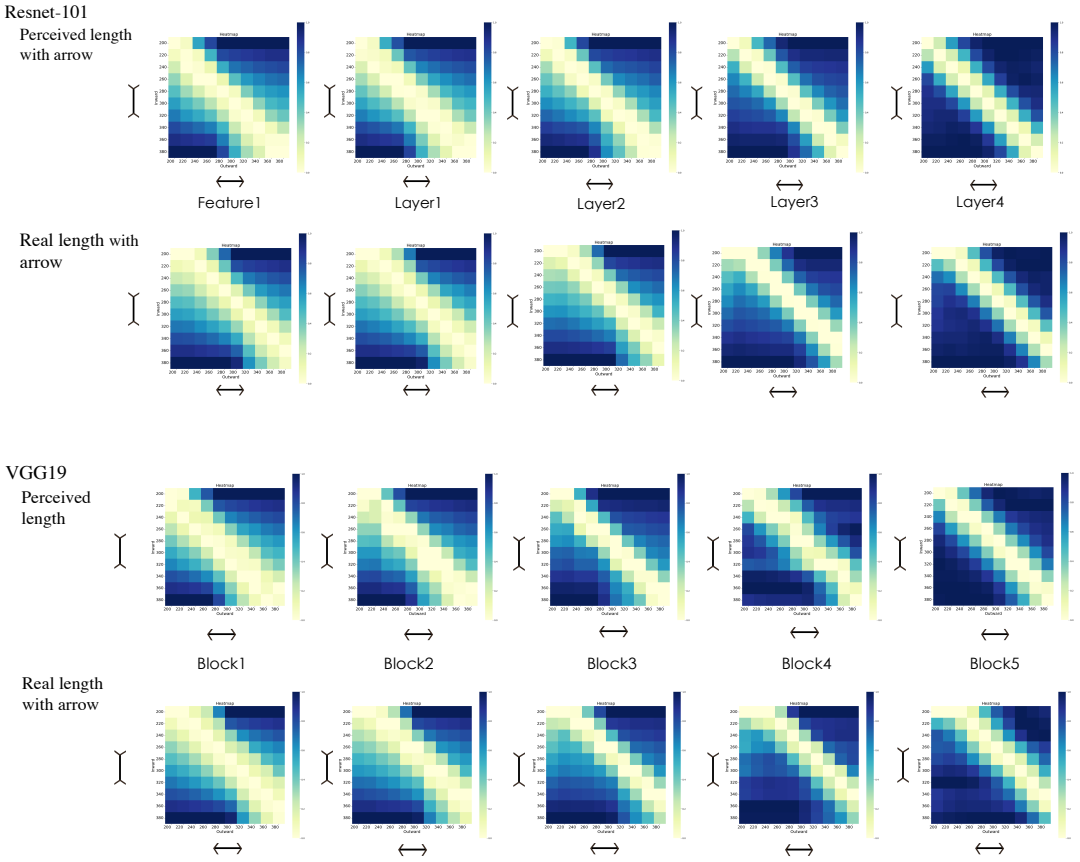


Figure 3.3: The RDM between perceped group and control group on VGG19 and ResNet101

3.1.6 The Illusion Response Changing of Different Model Depth

We then continued to explore within the models the specifics under different network depths. Figure 3.4 shows the distribution of the L2 distances for the two orientations of lines in Vgg19 and ResNet101 across both groups. The distribution trends are generally consistent between the two groups, which indicates the models' understanding of length and the real representation of visual illusions. Furthermore, the shallower the depth of the model, i.e., the lower layers, the more pronounced the visual illusion(lower L2 distances). This may suggest the potential areas in DNNs where visual illusions occur.

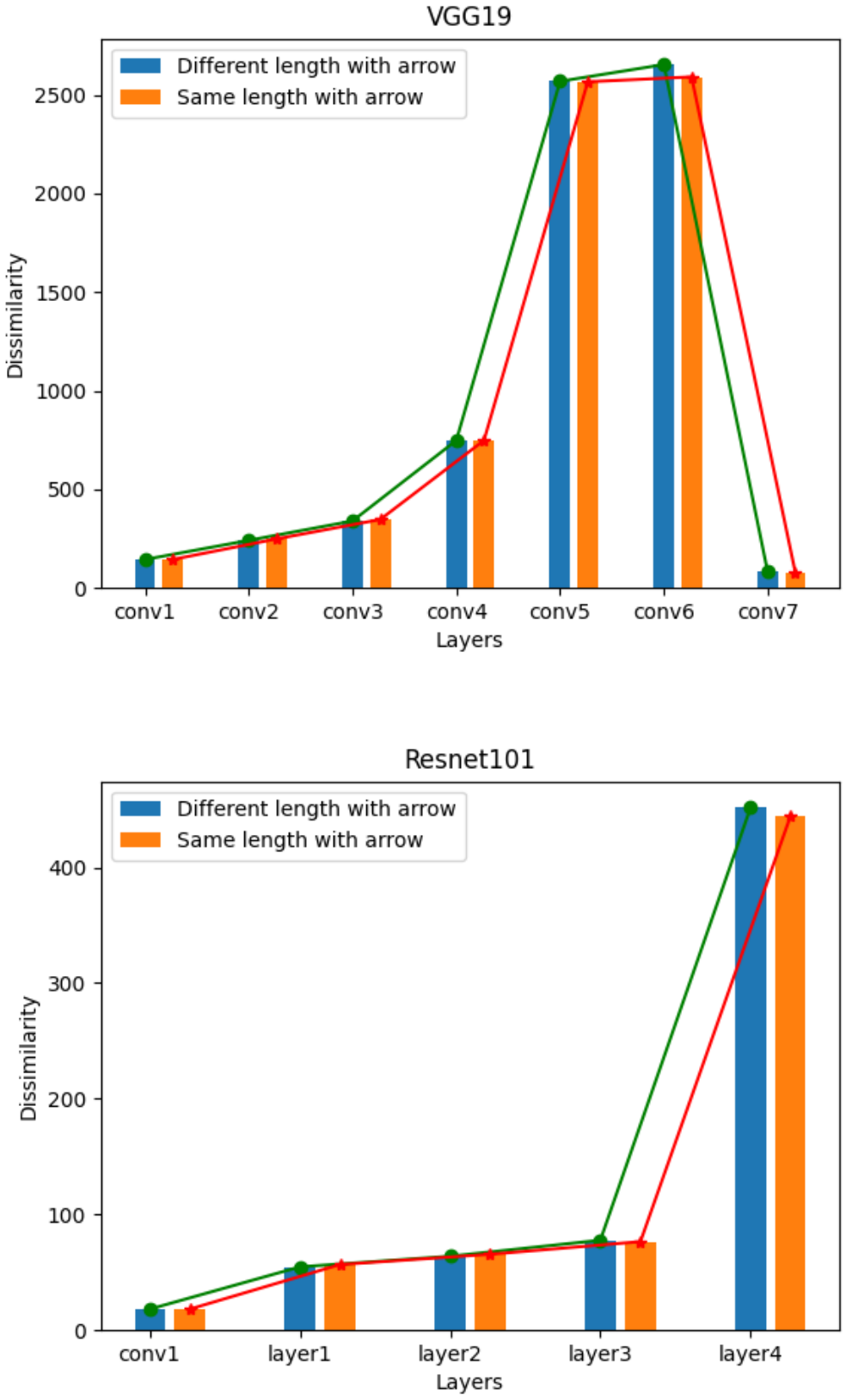


Figure 3.4: The illusion response about model depth of Vgg19 and ResNet101

3.2 Step1 :Müller-Lyer Illusion on More DNNs Models

Although the two classic DNN models exhibited human-like visual illusion responses, relying on these two models is insufficient to represent the brain-like characteristics and universality of visual illusions in DNNs. Therefore, we expanded our testing to include other models such as DenseNet201 [61], AlexNet [16], EfficientNet-b3 [62], ResNeXt101 [63], Vision Transformer [64], and Swin Transformer [65] for further testing.

- DenseNet201: This is a densely connected convolutional network characterized by obtaining feature information from all previous layers at each layer. This structure helps solve the problem of vanishing gradients in deep networks and enhances the reusability of features, keeping good performance with fewer parameters.

- AlexNet: It was the first deep convolutional network to successfully use ReLU as the activation function, accelerating the training process with GPUs, and introduced local response normalization and dropout to improve training efficiency and generalization ability.

- EfficientNet-b3: EfficientNet is a series of models where the b3 version achieves efficient performance optimization through compound scaling (simultaneously expanding the network's width, depth, and resolution). These models achieve exceptional accuracy and efficiency with lower computational costs.

- ResNeXt101: As an extension of ResNet, ResNeXt introduces the concept of grouped convolution, increasing the model's diversity by varying the pathways of convolution kernels. ResNeXt101 offers a way to enhance network capabilities through parameter reuse, improving the model's ability to handle complex data.

- Vision Transformer (ViT): ViT is a Transformer model that applies the self-attention mechanism, originally designed for natural language processing but later adapted for image classification tasks. ViT processes images by dividing them into multiple patches and treating them as a sequence, demonstrating competitive abilities in visual tasks against traditional convolutional networks.

- Swin Transformer: Swin Transformer is a hierarchical Transformer whose design allows it to scale more effectively to images of various sizes and handle more complex visual tasks. It achieves a balance between computational efficiency and model performance by using a moving window self-attention mechanism.

3.2.1 The RDM on Eight Models

Similar to previous work, the Euclidean distance of the feature vectors of lines with different arrow directions in the perception group and the control group was used to observe the performance of visual illusions. From the figure 3.5, differences can be seen between the models. Vgg19, AlexNet, and ResNet101 all show high similarity along the diagonal, while DenseNet201 shows high similarity near the diagonal area. The other models exhibit irregular distributions, especially models with Transformer architecture. The RDM of the control group shows an upward shift of high similarity within Vgg19, AlexNet, and ResNet101, which is also consistent with previous performances. The similar change also can be seen in DenseNet201. However, the other four models still exhibit chaotic distributions, revealing their lack of visual illusion performance.

3.2.2 The L2 Distance Changes on Eight Models

Then we also displayed the internal visual illusion performances of these eight models (Fig. 3.6). The trend of changes in the two image groups across all eight models was similar, further proving that the models' performances were not accidental. Generally, models that exhibited illusion responses showed decreasing representational dissimilarity with increasing network depth. However, DenseNet201 showed a trend of decreasing and then increasing to high representational dissimilarity. ViT showed relatively low representational dissimilarity but exhibited irregular representational similarity in the RDM. This may be related to ViT and Swin using a global modeling approach with self-attention mechanisms to process images. This processing is not confined to local receptive fields and spatial hierarchies and may differ from the biological visual system's processing methods. Similar behaviors were observed in ResNeXt101 32×8d and EfficientNet b3, which might because they focus more on global features and lack effective feature capture.

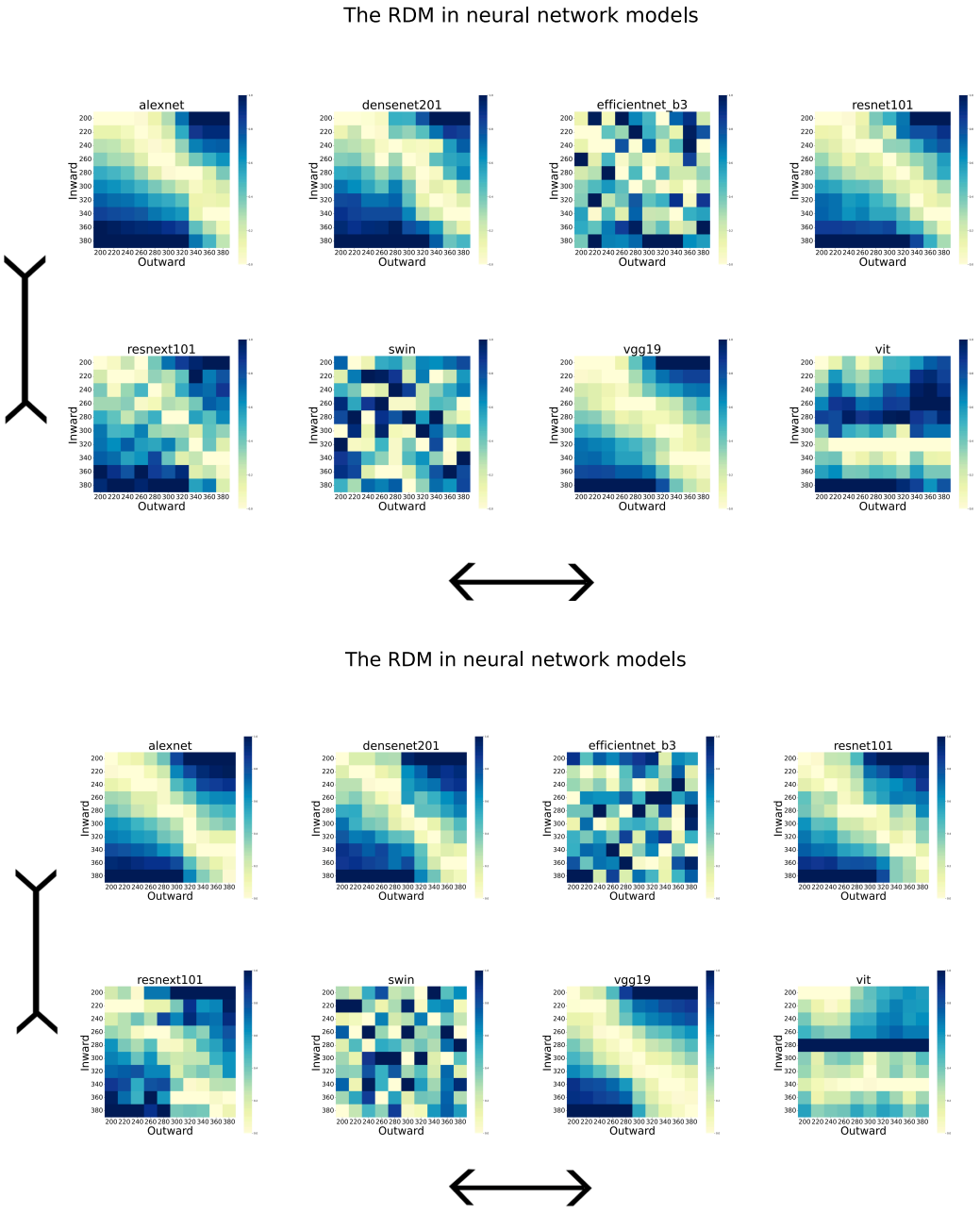


Figure 3.5: The RDMs of eight models on arrow outward and inward

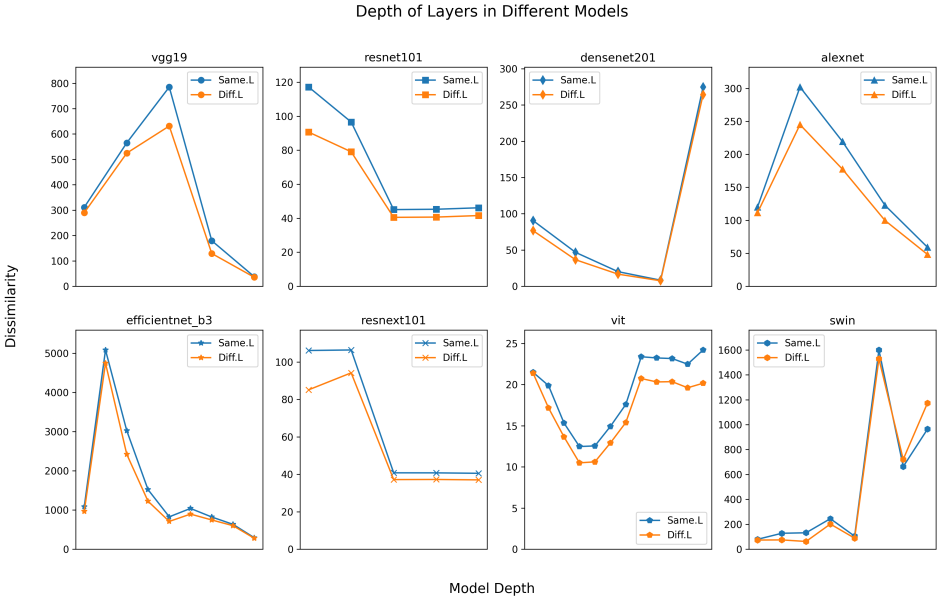


Figure 3.6: The L2 distance trend of two group on eight models

3.2.3 Grad-CAM Visualization on Eight Models

Based on the differences in model representations, we also used Grad-CAM [66] analysis to examine the feature preferences of different models to visually interpret the occurrence of illusion responses. As shown in the Fig. 3.7, we found that each model focuses on different features when handling illusion tasks. Transformer-based models and EfficientNet-b3 paid more attention to the arrows themselves, while VGG19 focused on the length of the lines. The illusion response seems to involve attention to the arrows and, to some extent, the lines between the arrows. These preferences might lead to illusion judgments or cause neural networks to make biased length judgments.

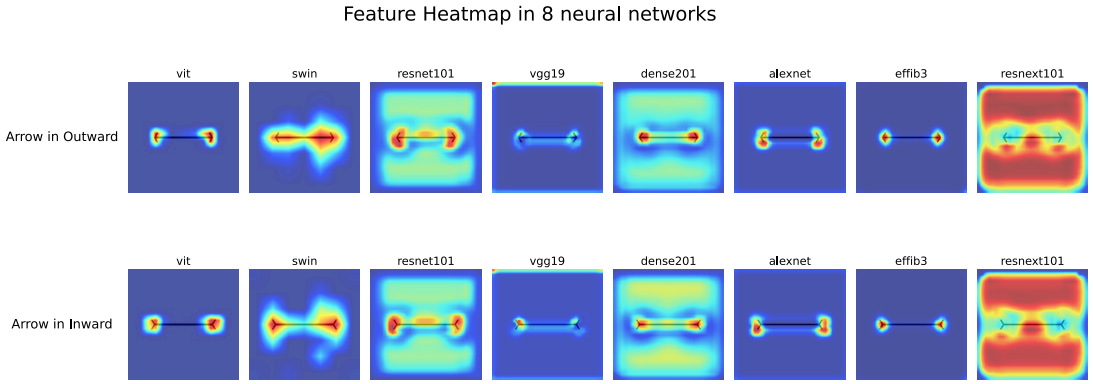


Figure 3.7: The heatmap of feature attention on eight models

3.3 Step 2: Five Optical Illusions

After exploring optical illusions in multiple DNN models, we further expanded the types of optical illusions to supplement the investigation and explanation of whether DNNs truly exhibit human-like optical illusions.

3.3.1 The Human Experiment on Five Optical Illusions

In this experiment, we used five classic optical illusions: color assimilation, the Hermann grid illusion, the Müller-Lyer illusion, the Zöllner illusion, and the Poggendorff illusion (Fig. 3.8). The specific experimental steps and settings are as follows:

- Color assimilation illusion: This illusion shows how visual perception is influenced by surrounding colors. In this illusion, the color of rectangles against backgrounds of different striped colors shows three different depths of color. We used twelve colors based on the RGB color wheel, with backgrounds consisting of alternating black, gray, and white stripes. The rectangles corresponded to three different depths, totaling twelve color combinations (36 stimuli). Each round displayed a group of colored rectangles and corresponding three different depths against black, gray, and white striped backgrounds, labeled as ASD on the display (Fig. 3.8 A). Participants observed the similarity between the color of the rectangle from ASD (corresponding to depth1 to 3) and the original color on the far right, ranking them by perceptual color similarity. After completing, they pressed the space bar to proceed to the next group.

- Hermann grid illusion: It is a pattern of grids and white dots creates flashing dots at grid intersections, revealing characteristics of the visual system in processing light contrast and edge detection. We set up five size ranges (6 10) and five grayscale levels from white to black ($\mu = 1 \sim 5$), with grid colors in red, green, blue, and gray, totaling 100 stimuli (Fig. 3.8 B). The experimental steps involved randomly displaying 100 stimulus images, with ASD corresponding to "Clear", "Weak", and "None" options. Participants selected based on their perception of the flashing points and proceeded directly to the next round.

- Müller-Lyer illusion: It includes two lines with different arrow shapes, illustrating how contextual cues affect our perception of length; the straight lines with different arrows make the lengths appear altered. We set lines ranging from 200 to 380 pixels in length, each paired with inward and outward arrows, totaling 20 stimuli (Fig. 3.8 C). The experimental steps involved randomly displaying lines from 200 to 380 pixels in length at the top of the screen, with lines with random arrow directions shown at the bottom. Participants adjusted the length of the arrowed lines using the WASD keys. AD for major adjustments ($\pm$10 pixels) and WS for minor adjustments ($\pm$1 pixel), and after completion, they pressed the space bar to proceed to the next round.

- Zöllner illusion: This kind of illusion shows how a straight line appears misaligned when partially occluded by rectangles, highlighting the limitations of the visual system in interpreting line directions and parallel relationships. We set seven rectangle width sizes ($120 \sim 240$, in 20 pixel increments) and five angles between the line and rectangle (15°, 30°, 45°, 60°, 75°), totaling 35 stimuli (Fig. 3.8 D). The experimental steps involved randomly displaying 35 stimuli in the center of the screen. Each round, a line parallel to the rectangle's top line and at the same angle appeared randomly below the rectangle. Participants adjusted the position of the lower line until it visually aligned with the upper line. AD for major adjustments and WS for minor adjustments, and after adjusting, they pressed the space bar to proceed to the next round.

- Poggendorff illusion: The lines at specific arrow positions create the sensation of a change in line angle, demonstrating how our visual system processes spatial positions and line alignment. We set six angles (30°, 40°, 45°, 55°, 60°, 75°), three

arrow spacings (15, 25, 35), and lines in three regions of the arrows (1/3, 1/2, 2/3), totaling 54 stimuli (Fig. 3.8 E). The experimental steps involved randomly displaying 54 stimuli at the top of the screen, with two parallel horizontal lines below. Participants adjusted the angle of the lower line based on the stimulus above. The default adjusted the first line, pressing Enter to switch line control. AD for major adjustments (±0.1°) and WS for minor adjustments (±0.02°). Adjustments were optional, and they pressed the space bar to proceed to the next round.

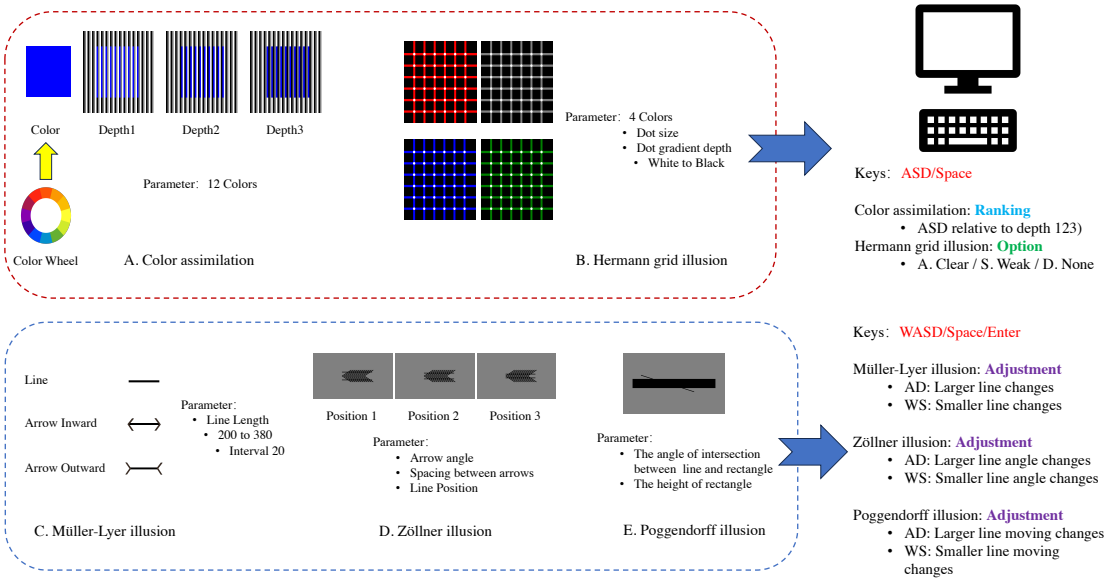


Figure 3.8: Five optical illusions and their human subject experiment configuration

The experiments were conducted in a dark room using an HP P244 monitor (23.8 inches, refresh rate: 60Hz, resolution: 1920×1080) for stimulus presentation. Participants stabilized their head position using a chin rest (Tobii Pro AB) to ensure that all participants viewed from the same angle and distance. The screen was 65 centimeters away from the participants. Before the formal experiment, all participants underwent practice and testing, and were informed to only observe and make adjustments based on their visual perception. The experiments were conducted using Matlab based on Psychotoolbox. Depending on the setting of total research, we collect participants' perceptual data as perceptual images for various visual tests comparing with optical illusion images.

3.3.2 Models and Processes

Before testing, we selected some models with single-path sequential feedforward architectures and extensive spatial integration features, including the Inception series [67] (Inception_v1, Inception_v3), AlexNet, and the VGG series (VGG16, VGG19). Additionally, to provide a more comprehensive perspective in our assessment, we also considered the ResNet and DenseNet series, which rank high on the Brain-Score, such as ResNet50, ResNet101, ResNet152, ResNet152_v2, DenseNet169, and DenseNet201.These models are loaded and used via Pytorch's torchvision and timm packages. We continue to use pretrained methods to load the models, and Table 3.1 displays the pretrained weights and total parameters of each model.

Table 3.1: The main configuration and parameters of various DNNs Models

Model	SOURCE	Package	Parameters (Millions)
AlexNet	IMAGENET1K	Torchvision	61.10
Vgg16	IMAGENET1K	Torchvision	138.36
Vgg19	IMAGENET1K	Torchvision	143.67
ResNetv2_50	IMAGENET1K	Timm	25.55
ResNetv2_101	IMAGENET1K	Timm	44.54
ResNet152	IMAGENET1K	Torchvision	60.19
ResNext101	IMAGENET1K	Torchvision	88.79
Inception_v3	IMAGENET1K	Timm	23.83
Inception_v4	IMAGENET1K	Timm	42.68
DenNet121	IMAGENET1K	Torchvision	7.98
DenNet169	IMAGENET1K	Torchvision	14.15
DenNet201	IMAGENET1K	Torchvision	20.01

The experiment also used Euclidean distance to construct a Representational Dissimilarity Matrix (RDM), extracting the feature vectors from the last layer before the classification layer and calculating the Euclidean distances (L2 distance) between different feature vectors to construct the matrix. Considering the experimental differences of the five visual illusions, apart from color assimilation and the Hermann grid illusion, the remaining three visual illusions constructed unconventional RDMs, namely by extracting feature vectors of stimulus images and adjusted perceptual images. The horizontal and vertical axes correspond to the stimulus images and the perceptual images adjusted by human subjects respectively.

The construction of specific RDMs is as follows:

- Color Assimilation: A 48×48 RDM composed of 12 colors.
- Hermann Grid Illusion: A 25×25 RDM for each color with different parameters.

- Müller-Lyer Illusion: A 10×10 RDM, based on perceptual data of arrow direction.
- Zöllner Illusion: A 54×54 RDM, based on stimulus images and adjusted perceptual images.
- Poggendorff Illusion: A 35×35 RDM, also based on stimulus images and adjusted perceptual images.

To further understand the internal mechanisms of the network, especially in cases where it exhibits human-like perception, we used Class Activation Mapping (CAM) visualization techniques, including Grad-CAM [66] and Grad-CAM++ [68], to interpret the internal decision-making processes of DNNs in processing visual information.

3.3.3 Color Assimilation

In this study, we investigated subjects' ranking perceptions of three different color depths (Fig. 3.9A). Through average frequency, we found that the second depth (depth2) of all colors received the highest ranking, displaying the highest visual similarity to the original colors (Fig. 3.9A). Totally, two main ranking were observed: depth (231) and depth (213), showing consistency in perception of these specific color depths.

After testing the color similarity rankings of these 12 models, as seen in Fig. 3.9B, Vgg16 and ResNetv2_101 exhibited the highest frequencies in the rankings for color depth at depths (231) and (213). Furthermore, we analyzed the depth variations of 12 colors across 12 models, extracted the feature vectors from the last convolutional layer, and calculated the L2 distance to assess similarity (Fig. 3.9C). All models almost show a trend of color ranking change in the last module, but from the ranking distribution, the shallower module does not have ranking on "231" and "213" until the depth6. It conflicts with the finding that DNNs early layer has the highest color sensitivity and decreases when depth improves [69].

Based on the RDM heatmap of 12 colors across all models (Fig. 3.10), there are significant differences among the models. The two models that most closely approximate human color perception rankings, ResNetv2_101 and Vgg16, do not exhibit consistent similarity distributions, whereas other models display varying degrees of similarity distributions. The RDM results further suggest that the models are not particularly sensitive to color or have a weaker understanding of color.

Regarding the models' responses to different color depths, we found that the ranking of green color frequently appeared in "231" and "213" in all models. Thus, we utilized CAM visualization method on green color stimulus (Fig. 3.11). From the figure, most networks focused within the color blocks, but Vgg16 and Vgg19 had fewer focal points. This suggests that DNNs are relatively sensitive to physical attributes but lack an understanding of color.

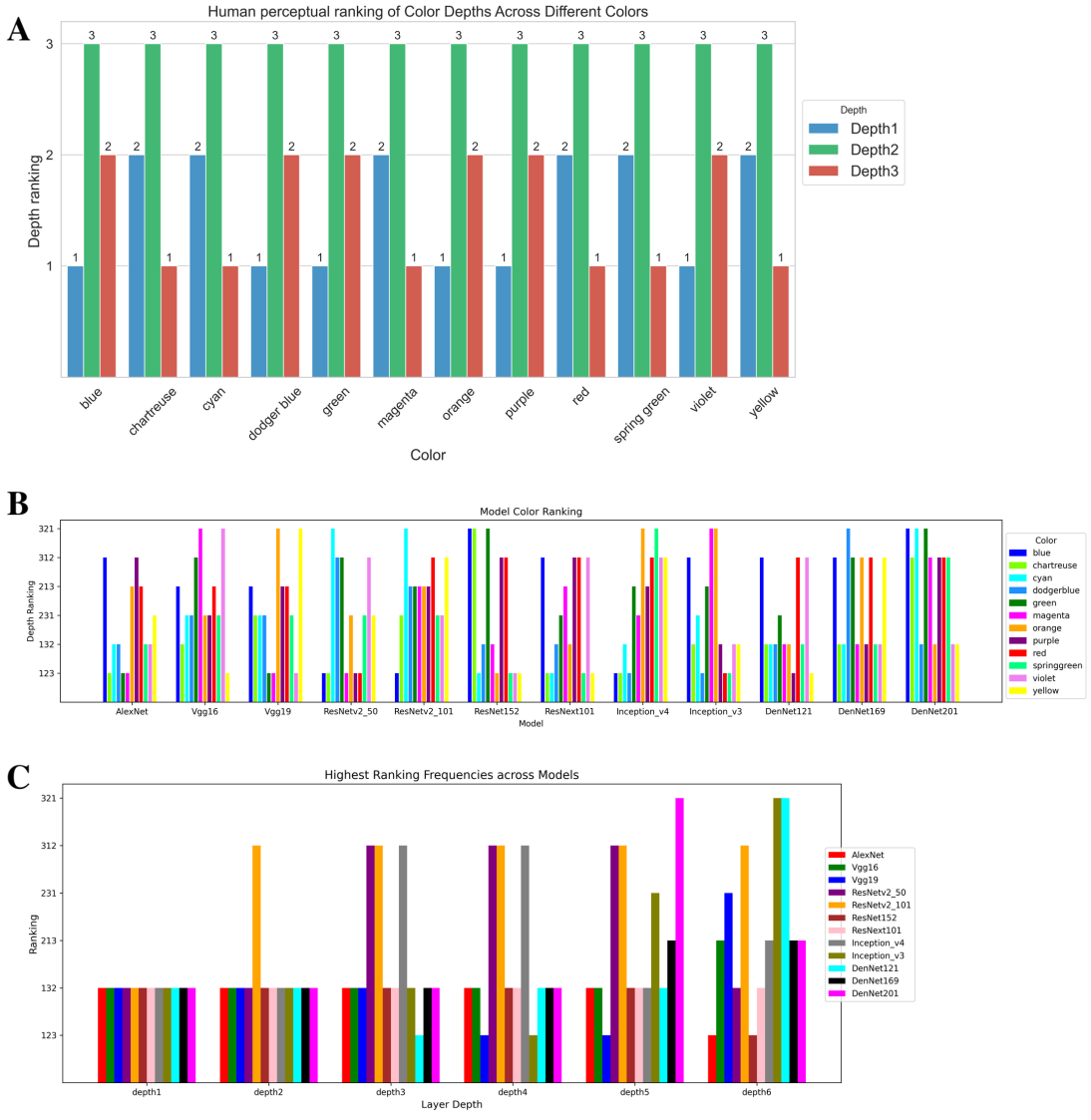


Figure 3.9: Human color ranking and models color test on 12 colors. A: Human subjects' color ranking on three depth of each color. B: The color depth ranking from 12 DNNs models. C: The color depth ranking on different model depth.

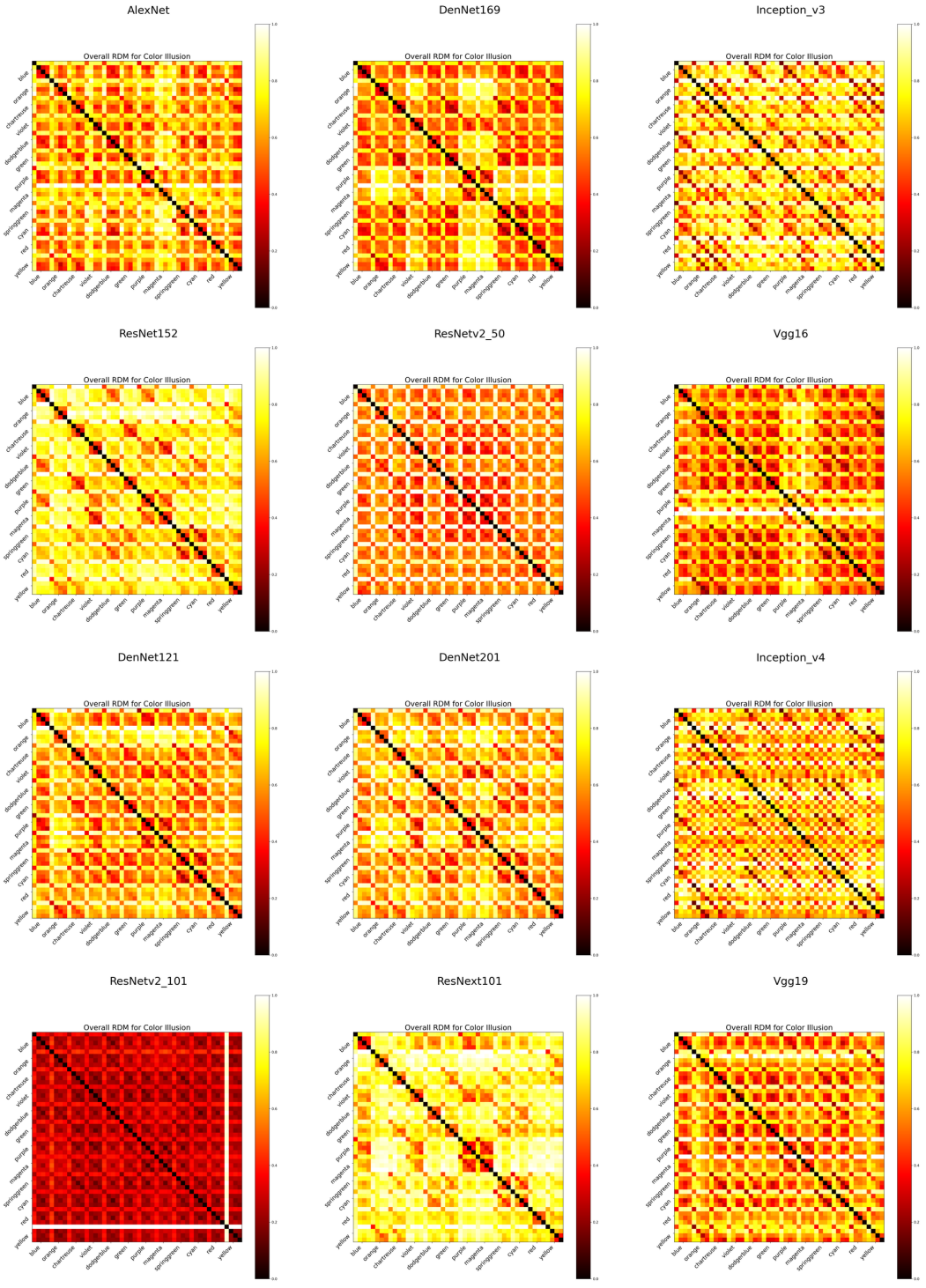


Figure 3.10: The RDMs of 12 colors within 12 models

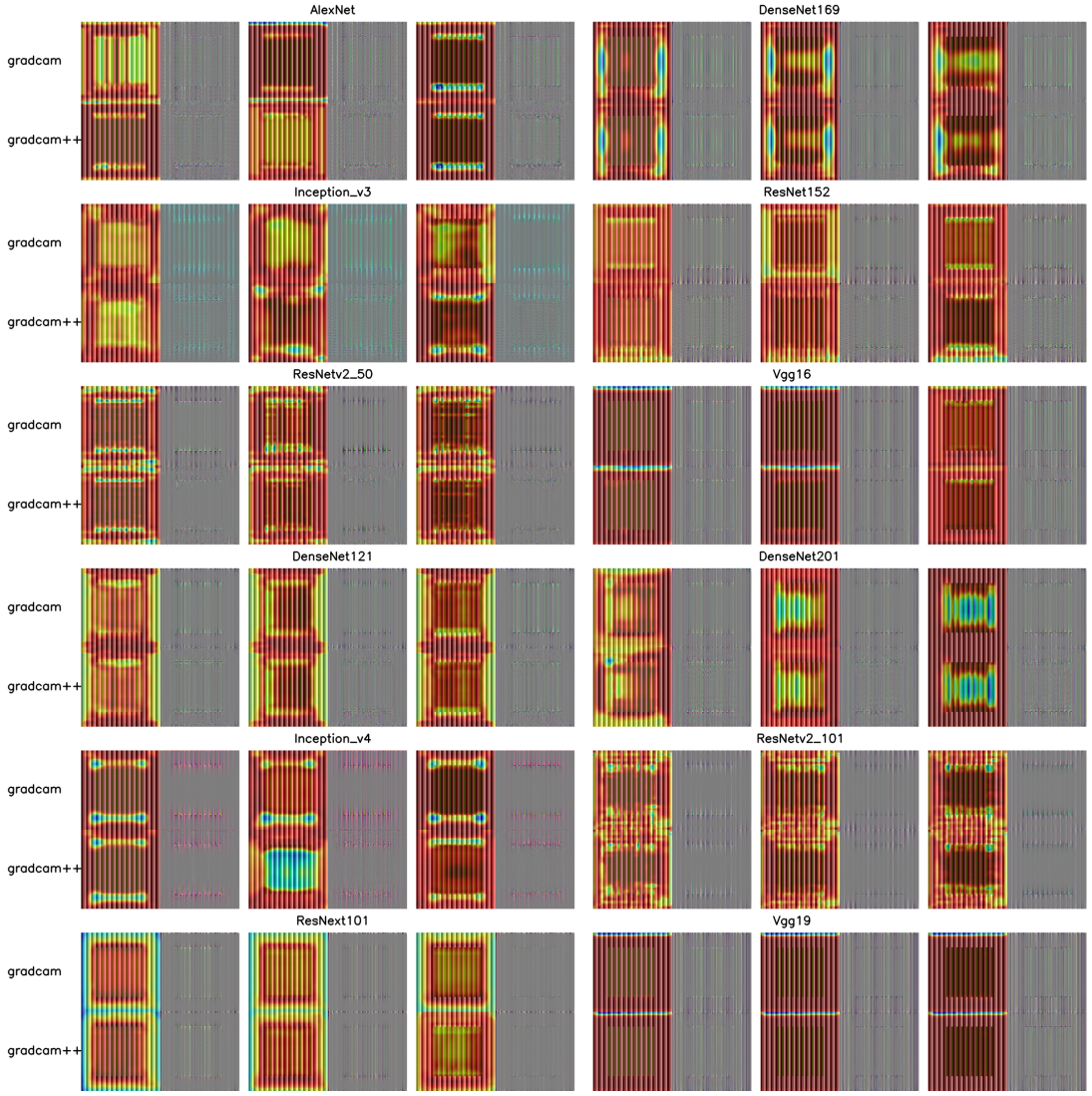


Figure 3.11: The CAM heatmap of green color on 12 models

3.3.4 Hermann Grid Illusion

We asked participants to evaluate the perception of flashing dots in the grid illusion images, classifying them into three levels: clear, weak, and none (see Fig. 3.12A). The results indicated that as the gradient depth increased from $\mu(1)$ to $\mu(3)$, the intensity of the perceived flashing dots decreased, especially in the green grids.

We tested 12 DNN models, evaluating their responses to grid images in four colors (Fig. 3.12B). We selected four colors ($\mu(1)$, dot_size=6) as the baseline condition and calculated the L2 distance of feature vectors for other gradient depths and dot size combinations in the same color. As shown in Figure 7B, with the increase in gradient depth from white to black, the 12 models initially showed a rising trend in similarity, but then almost all models exhibited a declining trend in the gray grids. Typically, as the gradient depth increases, similarity should monotonically decrease, i.e., the L2 distance should increase. However, the actual results showed a complex trend, and this declining trend suggests that models exhibit judgments similar to flashing dots under certain conditions, where the declining slope of the curve might reflect the strength of the flashing dots.

Also, Figure 3.12B shows that the flashing dots in the gray grids were the most pronounced, with green and red grids also showing similar but less frequent effects, and blue grids almost showing no flashing dot effects. Moreover, Inception_v4 exhibited a trend of decreasing and then increasing in the green grids, which might reflect a weak perception of flashing dots similar to "Clear." Similarly, DenseNet121 at dot_sizes of 7 and 8 and ResNetv2_101 in the gray grids also showed this trend. Under different dot sizes, the performance of flashing dots generally showed a positive correlation, i.e., the larger the dot size, the weaker the flashing dot effects. Overall, the 12 models varied in their perception of flashing dots across the four colored grids, with most models showing weak performance (low slopes). The three networks in the DenseNet series showed more pronounced flashing dot effects in gray, red, and green grids.

We further analyzed the performance of different colored grids in RDMs (see Fig. 3.13A). The RDMs of the four colors were overall similar, but there were significant differences between different gradient depths. Especially in gray grids, the deeper $\mu(3)$ and $\mu(4)$ showed higher similarity.

When analyzing the models with GradCAM, we found that different models focused on different features in the grid illusion images (Fig. 3.13B). The Vgg series mainly

focused on each dot, while the DenseNet series and the Inception series paid more attention to the grid edges, showing a weak feature preference for points in the middle areas.

Finally, we evaluated the performance of flashing dots at different network depths (see Fig. 3.14). The results showed that the response of the models was not monotonous with increasing network depth, possibly reflecting the complexity of human-like perception of flashing dots.

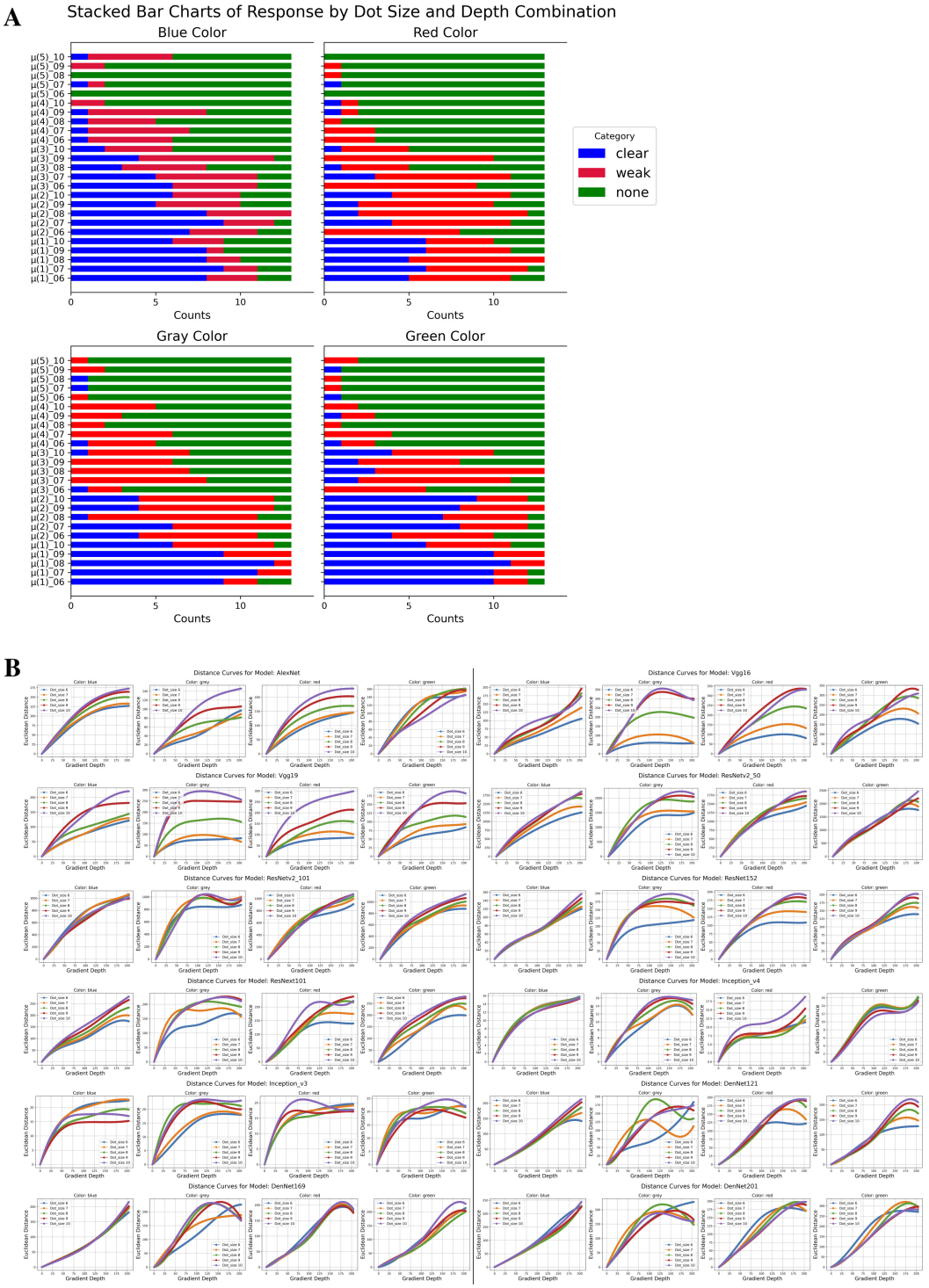


Figure 3.12: The response of grid illusion from human subjects and DNNs illusion test. A: The distribution of four colors gird illusion from subjects. B: The DNNs test on four colors gird illusion.

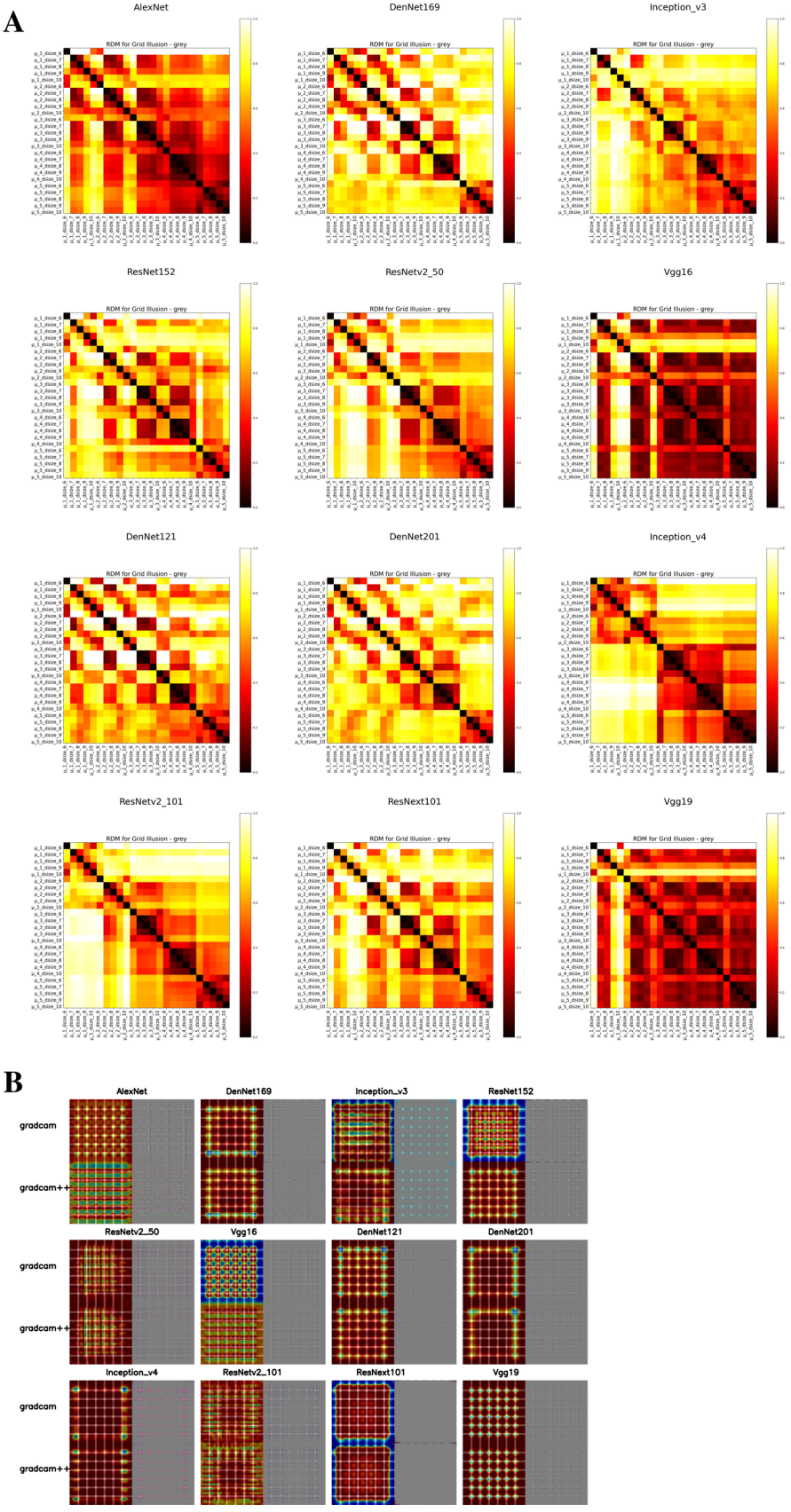


Figure 3.13: The visualization heatmap of Hermann Grid Illusion on 12 DNNs. A: RDMs of four color on DNNs. B: The feature attention focus from 12 DNNs.

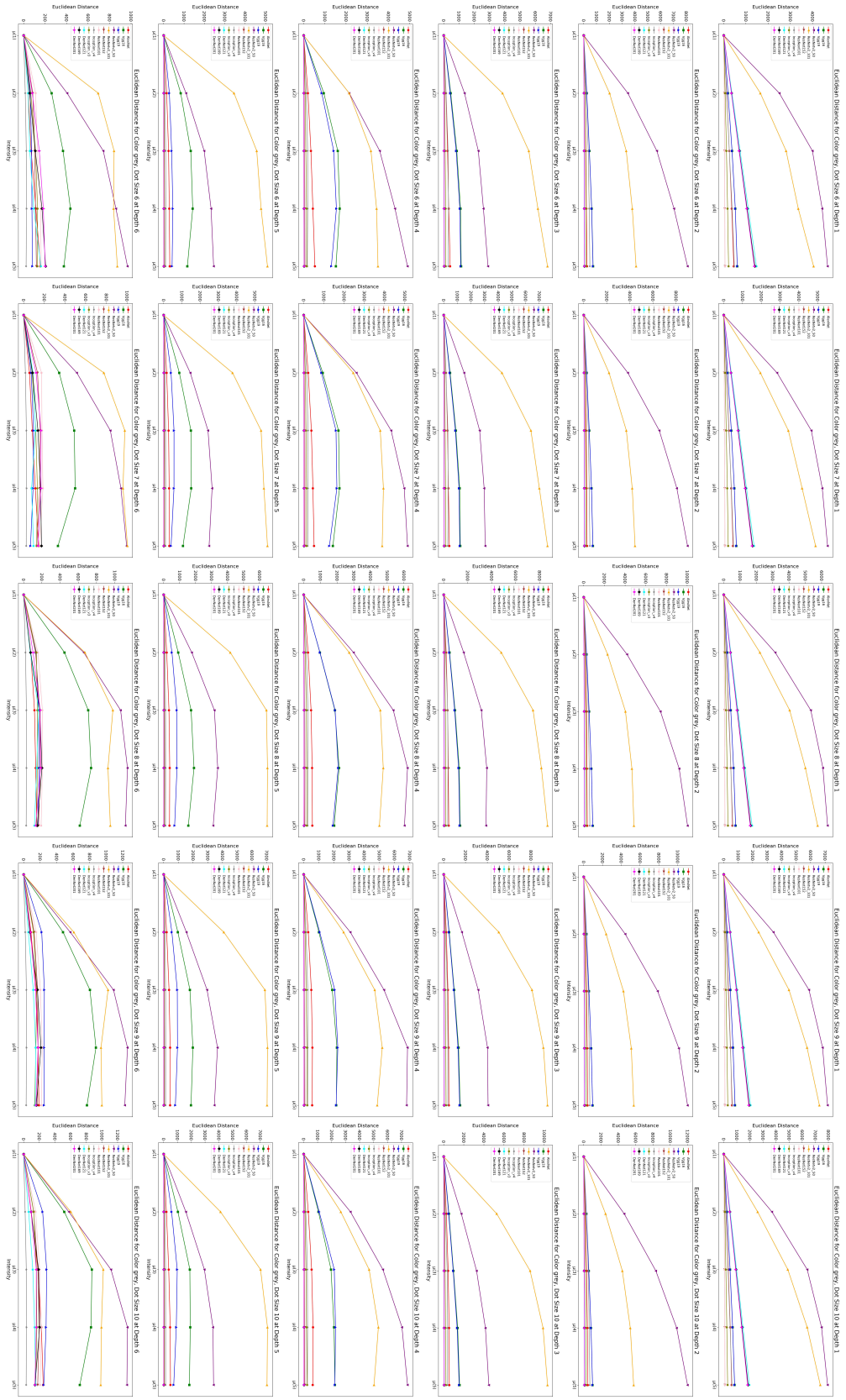


Figure 3.14: The illusion trend on different model depth.

3.3.5 Müller-Lyer Illusion

As for the Müller-Lyer Illusion, we averaged the data on participants' perceived lengths and compared it with the standard actual lengths to quantify the degree of visual illusion. The degree of visual illusion refers to the difference between the perceived and actual lengths. As shown in Fig. 3.15A, under different length baselines, the inward-pointing arrows (light green bar graph) and outward-pointing arrows (pink bar graph) produced different changes in length. The visual illusion for outward-pointing arrows tended to present as positive values, while inward-pointing arrows showed negative values in illusion strength. This means that lines with outward-pointing arrows appeared shorter, while those with inward-pointing arrows appeared longer.

We then calculated the Euclidean distance between the feature vectors of perceived lines with inward and outward-pointing arrows after adjustments and constructed an RDM (Fig. 3.16A). As seen in the figure, DenseNet169/201, Vgg19, and ResNetv2_50 showed high similarity at the diagonal, indicating that these networks highly similarly represented the perceived lines with inward and outward arrows, reflecting model performances similar to human visual illusions. In Fig. 3.16B, the control group showed a diagonal upward shift after adding arrows in opposite directions to lines of the same length, except for ResNetv2_50. This further validated the human-like perception shown by the models at the diagonal in Figure 3.16A.

In the visualized CAM methods, models with visual illusion manifestations showed feature focus on lines and arrows in both GradCAM and GradCAM++, while models without visual illusion manifestations mostly focused only on the arrows or lines (Fig. 3.15B). This difference may lead to the models' visual illusion performance in the Müller-Lyer illusion. Also, from the Fig. 3.15C, we can see that two group showed highly similar trend which indicate the illusion performance exist in DNNs.

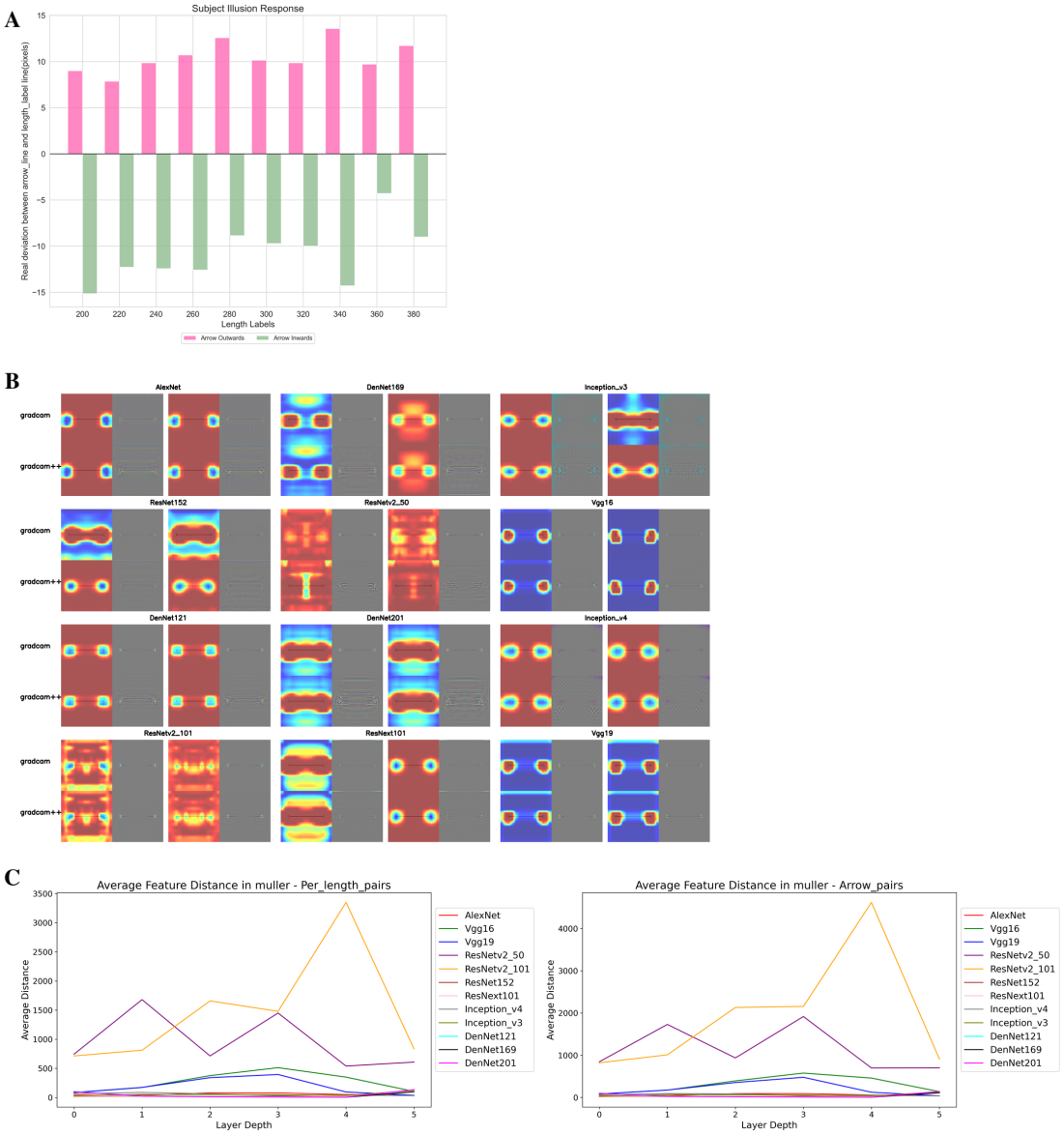


Figure 3.15: The perceptual length of Müller-Lyer Illusion and illusion test of DNNs. A: Human subject perceptual length between arrow outward and inward. B: The heatmap of feature attention focus on 12 DNNs. C: The L2 distance of different model depth on two groups.

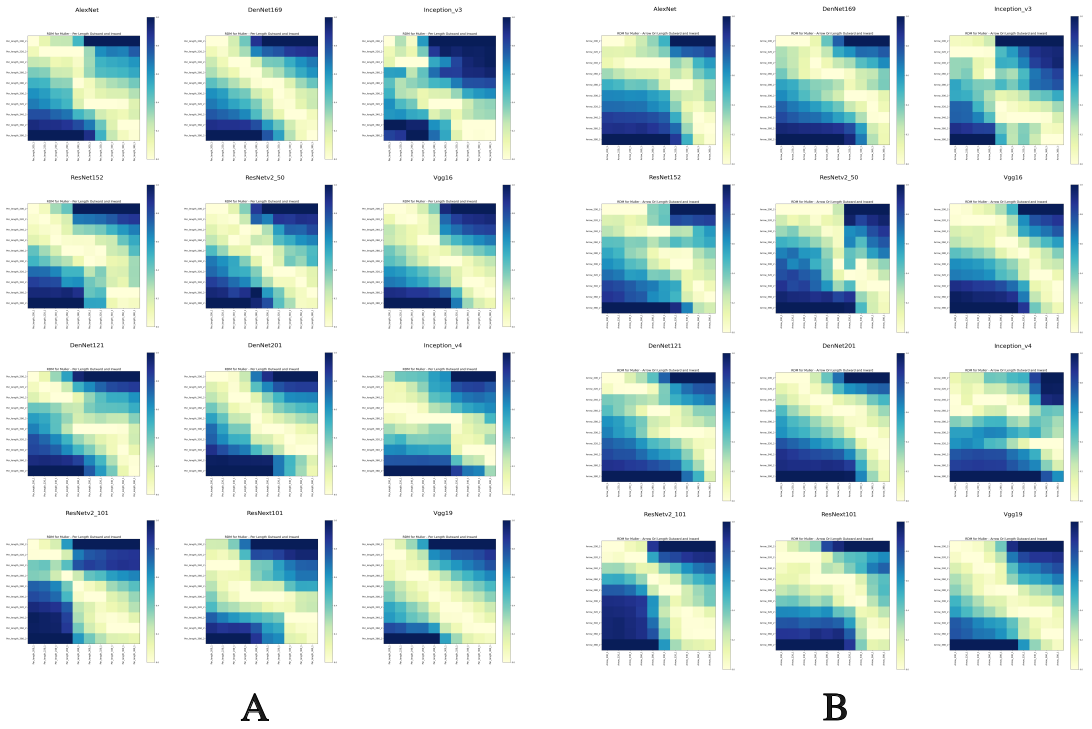


Figure 3.16: The RDMs on Müller-Lyer Illusion within perceptual gourp and control group in the DNNs.

3.3.6 Poggendorff Illusion

In the experiments on the Poggendorff illusion, we recorded the subject's data on adjusting the position of lines under varying angles and widths of rectangular overlays, and calculated the average positional deviation. Figure 3.17A shows that the larger the angle between the line and the rectangular overlay, the lower the deviation, and as the width of the rectangular overlay increases, the actual distance deviation fluctuates around a RectHeight of 120. Also, Figure 3.17B show that all models possibly has illusion response expect for ResNet_v2 structure.

We considered the average deviation of the positions as the perceptual image adjusted by the users, and constructed an RDM with the actual line images (Figure 3.17C). All models exhibited high similarity along the diagonal, with several models showing multiple parallel lines of similarity near the diagonal. Combined with the heatmaps visualized using the CAM method (Fig. 3.18), models focusing on line features in the RDM showed several clear parallel lines besides the diagonal, while models with fewer feature focuses on the lines showed similarity across many areas. Although the high similarity along the diagonal in all models indicates human-like visual illusion judgments,

these differences reflect whether the models truly focus on and understand the images and the manifestation of visual illusions.

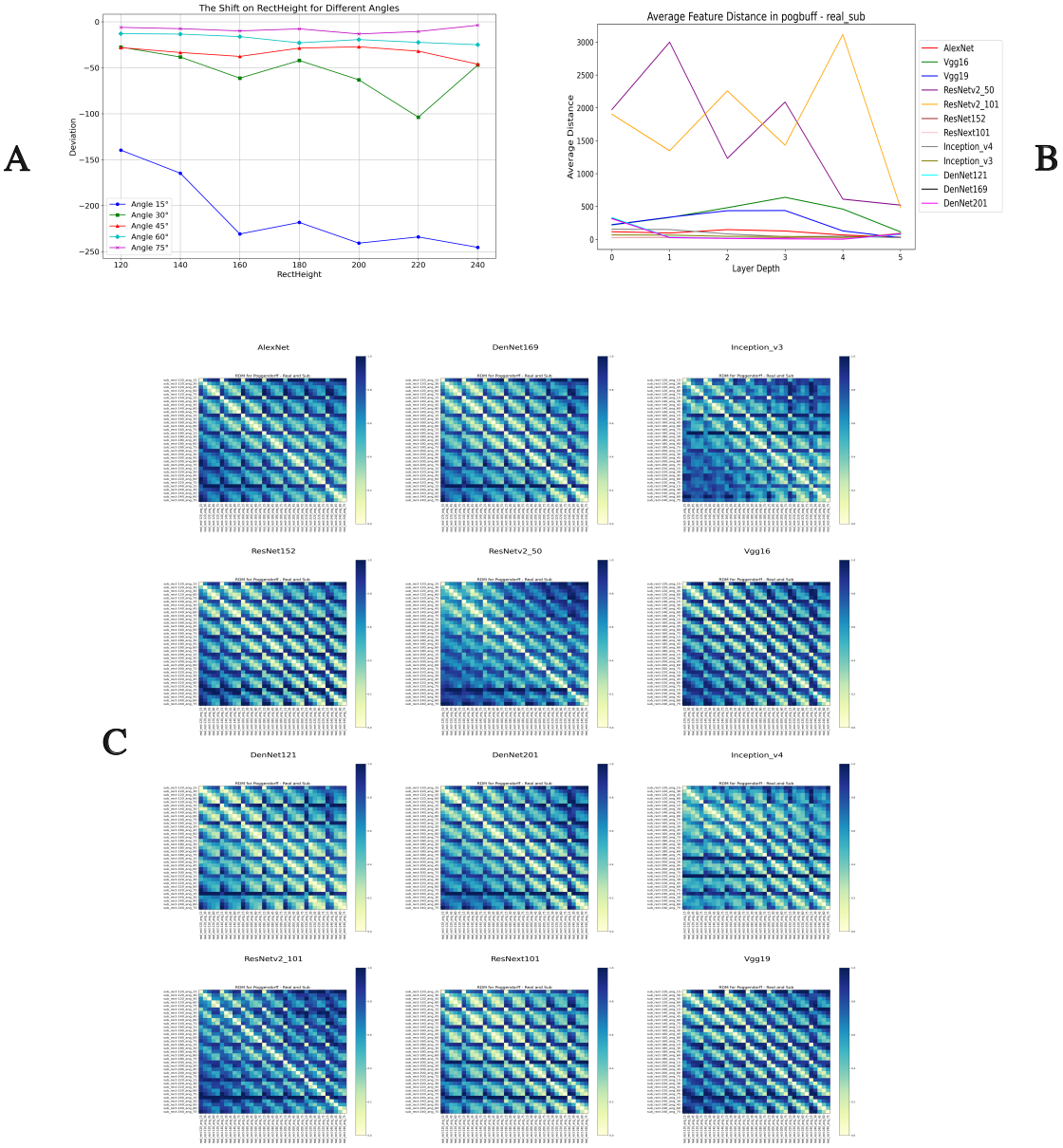


Figure 3.17: Human subject perceptual illusion and DNNs testing. A: The visual bias of Poggendorff illusion. B: Different model depth's L2 distance of illusion and perceptual stimulus. C: The RDMs of 12 DNNs

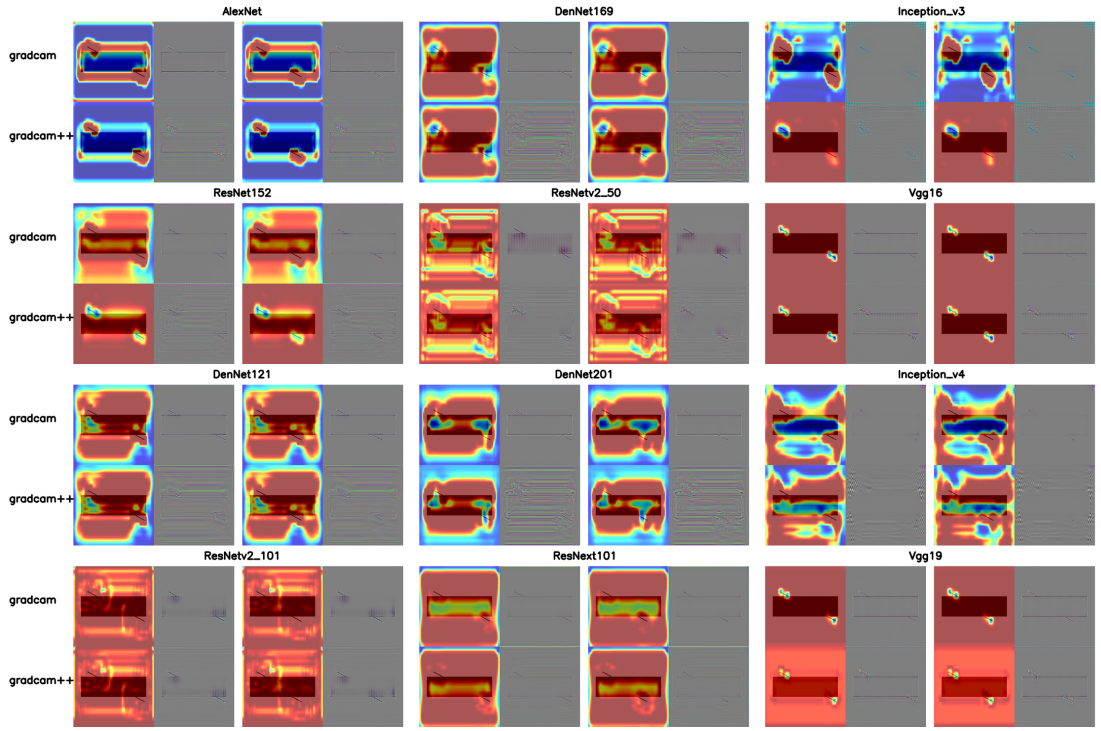


Figure 3.18: The heatmap of feature focus on Poggendorff Illusion from 12 DNNs.

3.3.7 Zöllner Illusion

Participants made angle adjustments after observing the Zöllner illusion to reflect the perceived tilt angle, indicating the strength of the visual illusion. Figure 3.19A shows the average perceived angles under different arrow spacings and angles. The perceived angles were larger at arrow angles of 30 and 44 degrees, while the other angles showed very small perceived angles, considered as no illusion. Generally, the line positions were relatively larger when centered by the arrows.

We prepared images of the averaged perceived angles for comparison with the stimulus images in constructing the RDM (Fig. 3.19C). The 12 models also showed high similarity along the diagonal. In the visualization with CAM (Fig. 3.20), nearly all models showed a high degree of feature focus on the overall combination of arrows and lines, without a greater feature bias towards the lines. In the Zöllner illusion, it does not indicate that DNNs exhibit visual illusions. Under different network depths, the main trends were similar to those in the Zöllner illusion (Fig. 3.19B, Fig. 3.17B), suggesting that DNNs are sensitive to the physical distances of lines rather than angles.

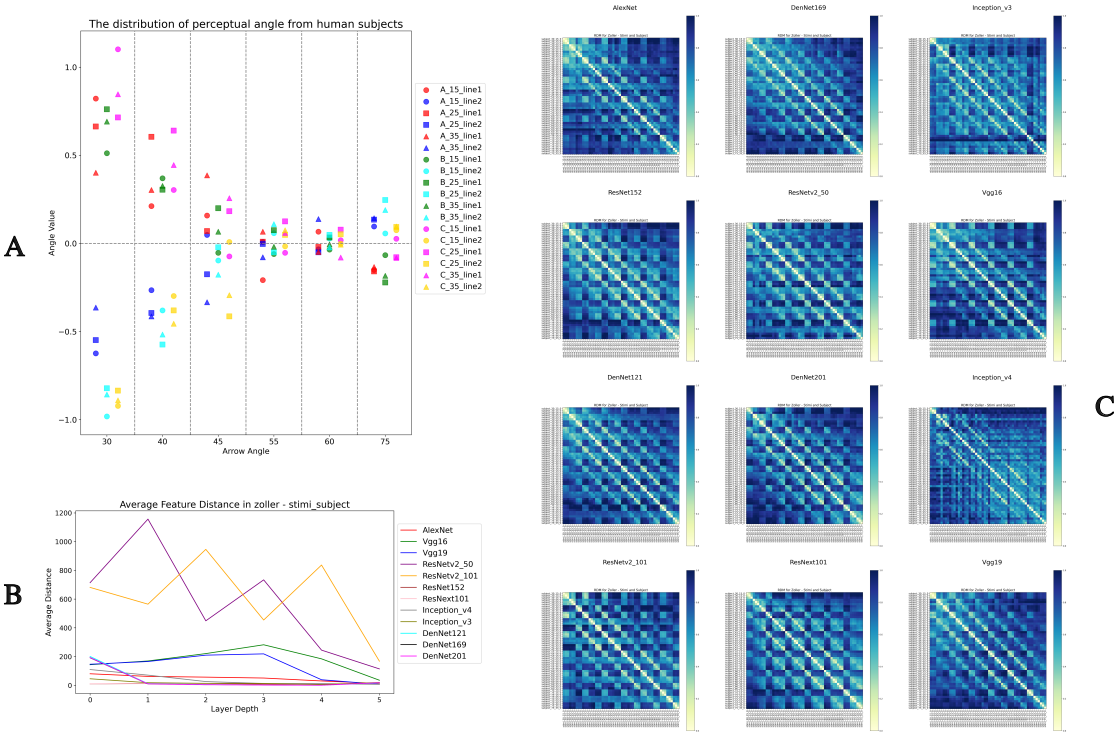


Figure 3.19: The huamn subject perceptual angles and model test. A:The average perceptual angle on Zöllner illusion. B: The L2 distance of different model depth. C: The RDMs of 12 DNNs.

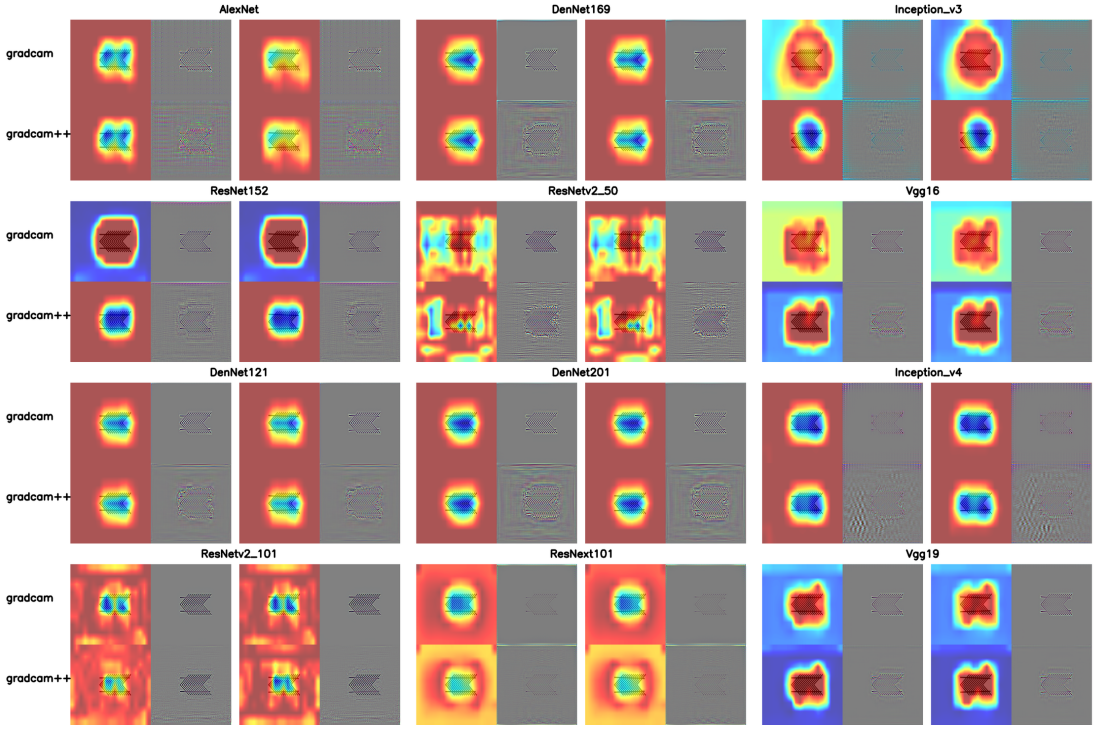


Figure 3.20: The CAM visulization of 12 DNNs on Zöllner illusion.

Chapter 4

Illusion Performance in DNNs by Specific Designed Dataset

This chapter primarily explores the performance of deep neural networks (DNNs) on a specific visual illusion dataset without pre-trained models. Additionally, it delves into the mechanisms underlying visual illusions and their potential brain-like guidance [54].

4.1 Oblique Illusion and Human Subject Experiment

Considering that pre-trained models cannot fully grasp physical attributes, particularly angles, such as inclination, we explored the performance of DNNs using a more complex and parameter-rich visual illusion, the Skye's Oblique Grating Illusion (Fig. 4.1). This illusion, a variant of the café wall illusion [70], consists of multiple parallel and horizontal bars combined with black-and-white alternating cubes. The cubes form patterns that create either a clockwise or counterclockwise appearance.

We designed an experiment to collect participants' responses to perceived inclination angles. During the experiment, participants need to observed 144 different visual stimuli combinations and adjusted the angle of colorless bar to match their perceived angle of stimulus bar above. Simply, the display shows two types of bars set at 0°(Fig. 4.2A). Participants used a keyboard to adjust the angle of the black bar below to align with their perception of the illusion stimulus above.

The experiment included options for large and small angle adjustments (±0.5°, ±0.1°). Before the main experiment, participants completed a practice session with 30 trials to familiarize themselves with the process. After the practice, participants could

start the formal experiment by pressing any key. They moved to the next trial by clicking a mouse after adjusting four angles and took a five-minute break between each of the four sets of 36 trials.

The visual illusion stimuli consisted of four bars of the same color and alternating black-and-white diamonds on a background of black and blue stripes (see Fig. 4.1B). The bars were 64 pixels long, and the black-and-white diamond edges ranged randomly from 5 to 10 pixels in width. The bars' colors were randomly generated from an RGB color wheel, totaling 12 colors (Fig. 4.2B). The positions of the black and white areas in the diamonds were also randomly switched, resulting in 144 combinations.

All participants had taken a color blindness test before the experiment to ensure they could correctly identify colors.

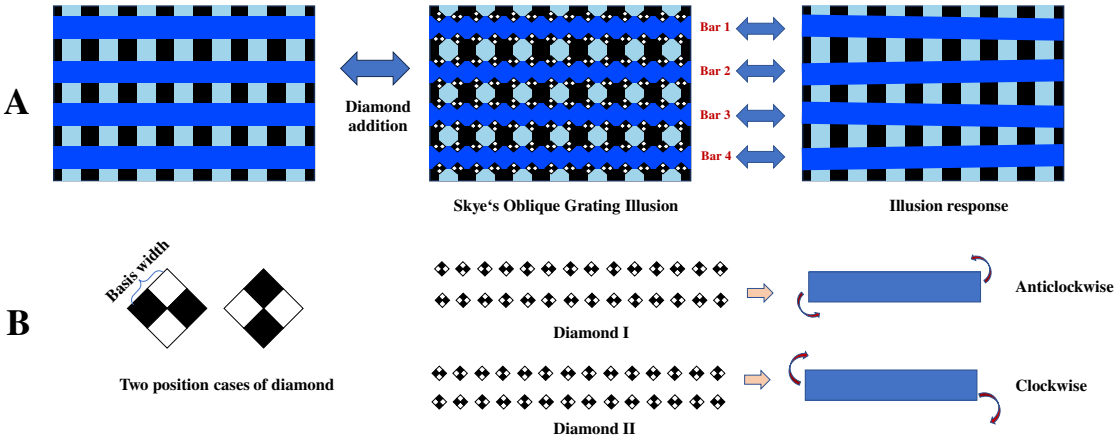


Figure 4.1: The main components of the Skye's Oblique Grating Illusion.

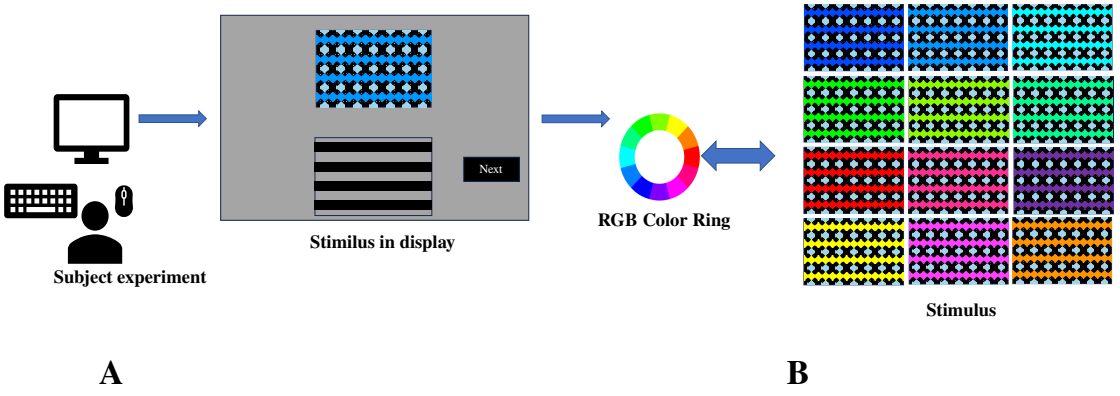


Figure 4.2: The preparation of the Skye's Oblique Grating Illusion.

4.2 Training Approach and Testing

4.2.1 Model Selection and Training Strategy

According to Brain-Score [44, 56], DNNs correspond to different regions of the visual cortex (Fig 2.8), simulating the hierarchical structure of the brain. But Nonaka et al.,(2021) [71] found that some DNNs show a negative correlation between image recognition performance and brain hierarchy scores, while simpler traditional models exhibit high similarity. Therefore, we selected four top-performing and four traditional DNN models from Brain-score.

Then we categorized participants' adjusted angle data into eight illusion strength levels. Based on the interval of 0.1 degree, these eight categories were further divided into two main categories: With-illusion (C2) and No-illusion (C1) (Fig. 4.3). We generated training datasets (3,000 images per category) based on the distribution of illusion strength shifts. The models were evaluated using k-fold cross-validation with the 144 stimulus images as the test set. Training was conducted using PyTorch with a learning rate decay strategy. The training focused on a binary classification task with the labels C2 and C1 to ensure the models accurately understood the concept of "inclination" before testing them on illusion images.

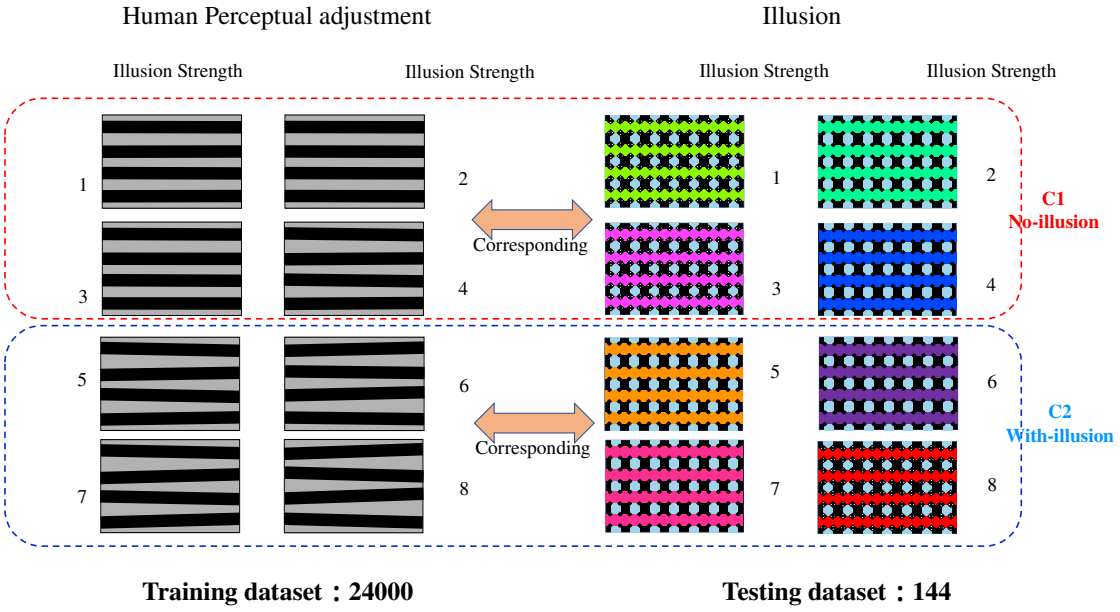


Figure 4.3: The eight illusion strength of optical illusion and dataset preparation.

4.2.2 Permutation Test

To validate the model testing's rationality, we employed permutation tests, which create "null distributions" by shuffling data and determining whether a test statistic is significantly different from random noise. In machine learning, significance tests can prove the reliability and validity of test results. We trained and predicted using the original data-labels, calculated the correlation (r-value) and p-value between predicted and original labels, and noted the accuracy as the r-value. Then, we shuffled the original labels to create per_labels, reassigned them to the data, and repeated the training and prediction. The correlation between per_predict_labels and labels was recalculated. Each model underwent 1,000 significance tests. We conducted similar tests during the testing phase to ensure accuracy.

4.2.3 The Visualization on RDMs

To evaluate and compare DNNs' representations under different stimuli, we constructed representational dissimilarity matrices (RDMs) to compare the differences between illusion and non-illusion images in the networks. RDMs still used Euclidean distance (L2 distance) as the measure of similarity. Here we chose the best illusion performance model as RDM visualization. Through ResNet101's various depth modules, we calculated the feature vectors of the stimulus and perception data. Each tensor was sized 280x160x1, generating 8x8 RDMs. We also extracted features from the module before the fully connected layers of the network for a comprehensive comparison of all eight models.

4.2.4 The Visualization on Grad-CAM

To visualize the regions responsible for the networks' responses to visual illusions, we utilized Grad-CAM after model training and testing. Grad-CAM creates heatmaps to explain the DNN's classification decisions by highlighting image pixels contributing to a specific class. We compared the differences between images with and without inclination illusions using the most responsive network (ResNet101).

4.3 Results

4.3.1 Perceived Angles

After obtaining the perceptual angles, we averaged the perception data and took the absolute values to reflect the illusion strength uniformly (Fig. 4.4). The x-axis in the figure corresponds to the 12 RGB colors, with each color group showing different cube widths, totaling 72 groups. Eight colors exhibited significant illusion deviations. Participants' feedback indicated that the other weaker colors were adjusted without bias. The perceived angle deviation increased and then decreased with the diamond edge width, with a width of 10 showing minimal deviation, similar to the colors that close to 0°.

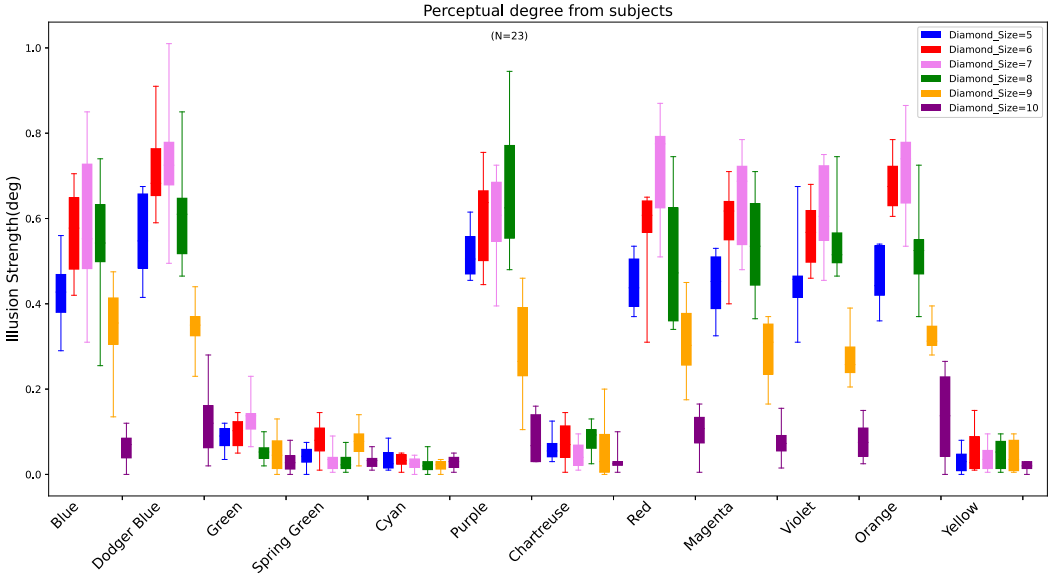


Figure 4.4: The human subjects' perceptual degree on 12 colors.

4.3.2 Model Performance

Before testing the eight models with visual illusion stimuli, we conducted 1,000 permutation tests. Fig. 4.5 shows the permutation test performance distribution for each model, displayed as light blue histograms and probability density curves in a 2x4 subplot layout. The actual test scores of the models are represented by orange dashed lines, and the 95% percentile of the permutation test by red dashed lines. Except for Vgg19, the actual test scores of the other seven models were significantly higher than their 95% permutation test percentiles. This is evident in the histograms where the orange dashed lines are typically to the right of the red dashed lines ($p = 0.05$), indicating that these models'

success in visual illusion image tests is not random and reflects an inherent "visual illusion" mechanism. Conversely, Vgg19's actual test score aligns with its 95% permutation test percentile, suggesting it lacks the capability to correctly recognize visual illusions.

As shown in Figure 4.6, the classification accuracy of the models varied significantly, with no clear correlation between the number of parameters and visual illusion classification performance. Among all models, ResNet101 performed the best, achieving a classification accuracy of 90.28%. In addition to accuracy, this model also excelled in recall and F1 scores, further confirming its advantage in visual illusion classification tasks. Although ResNet101 ranked mediocre or lower in Brain Score, its strong response to visual illusions highlights the complexity of neural and computational mechanisms in visual perception. In contrast, Vgg19, considered a good model for human visual perception, performed poorly in visual illusions, with an accuracy of only 61.81%. Other models, such as EfficientNet-B1, EfficientNet-B6, ResNeXt101_32×8d, ResNet152, DenseNet201, and PNASNet_5_Large, showed varying classification accuracies ranging from 74.31% to 88.20%. DenseNet201 performed slightly better in recall with a score of 0.99, compared to ResNet101's 0.97, and also had fewer parameters.

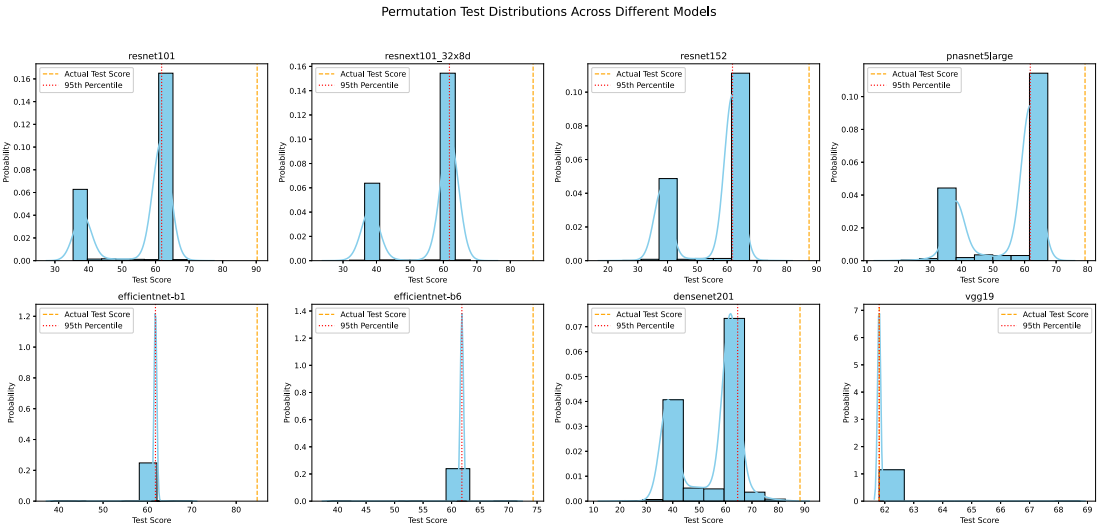


Figure 4.5: The permutation test result of 12 DNNs.

Additionally, Figure 4.7 shows the performance of the eight models across twelve colors and eight strength levels. From a color recognition perspective, most models performed exceptionally well on certain colors, such as "green," "spring green," "cyan," and "yellow," with many achieving 100% accuracy. However, performance was poorer on colors like "blue," "magenta," and "purple," indicating that some models are more

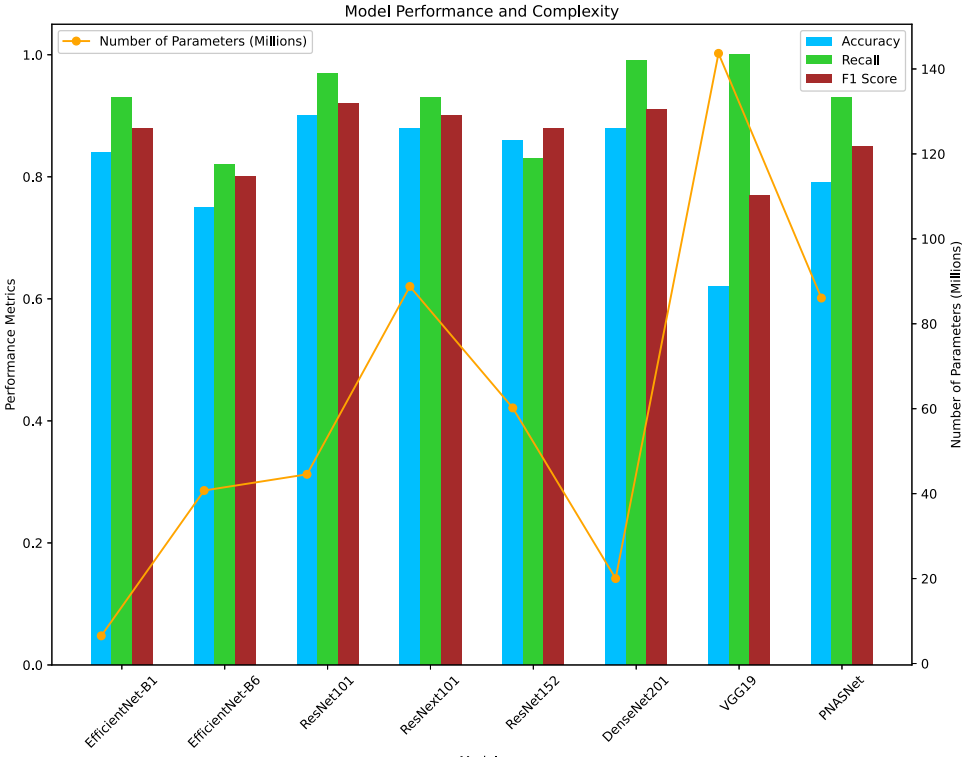


Figure 4.6: The models testing of C1 and C2.

adept at specific color ranges. Notably, EfficientNet-B1 and ResNet101 showed higher accuracy across most colors, reflecting their potential advantage in handling natural tones, while EfficientNet-B6 and Vgg19 showed lower accuracy, particularly on "orange" and "purple," indicating a sensitivity gap for certain hues. Almost all models showed low accuracy on "royal blue," suggesting a common recognition challenge for this color.

In terms of strength recognition, most models were more accurate at medium strength (e.g., 4 or 5), while accuracy generally dropped at extreme strength (e.g., 1 or 8), indicating challenges in handling subtle or very pronounced strength changes. ResNet152 and DenseNet201 performed well across most strength levels, especially between medium to high strength, while ResNet101 also performed well at medium strength (e.g., 3 to 6), demonstrating balanced capabilities in handling medium-level visual variations. On the other hand, Vgg19 and PNASLarge performed poorly at extreme strength, reflecting a lack of sensitivity to subtle or very strong visual effects. EfficientNet-B6 also performed poorly at low strength, indicating limitations in handling fine visual changes.

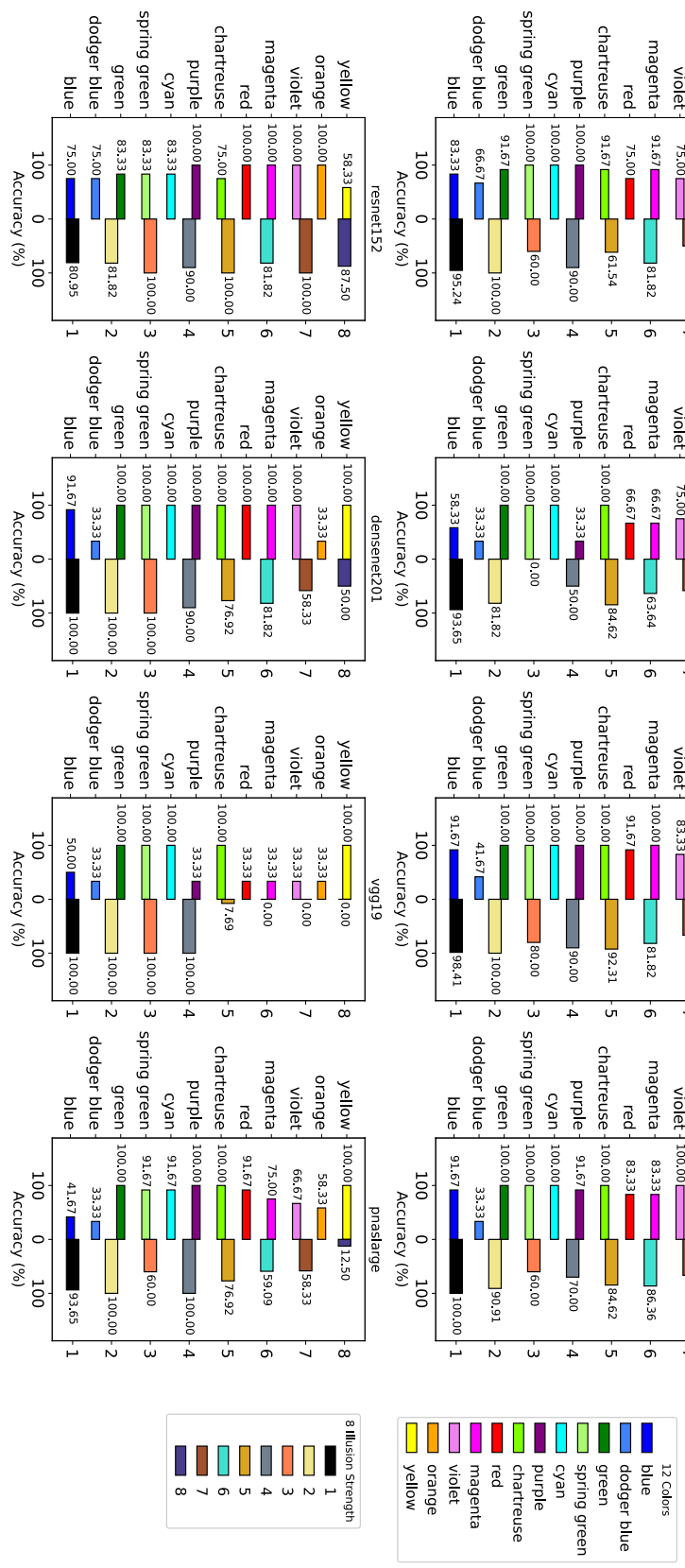


Figure 4.7: The result of color and illusion strength on DNNs.

4.3.3 Visualization Interpretations and Differences

To further explore how visual illusions manifest in deep neural networks (DNNs), we employed various visualization techniques to analyze DNN responses to illusion stimuli. Using Grad-CAM, we tested images with and without illusions separately, extracting and visualizing features to distinguish the DNNs' classification bases.

For No-illusion stimuli, DNNs exhibited multiple feature trends related to the colors of the bars. Different colors of non-illusion stimuli showed four distinct feature preferences. In contrast, With-illusion stimuli exhibited overall bending trends in their features, as shown in Figures 4.8and 4.9. These trends sharply contrasted with the features of non-illusion stimuli.

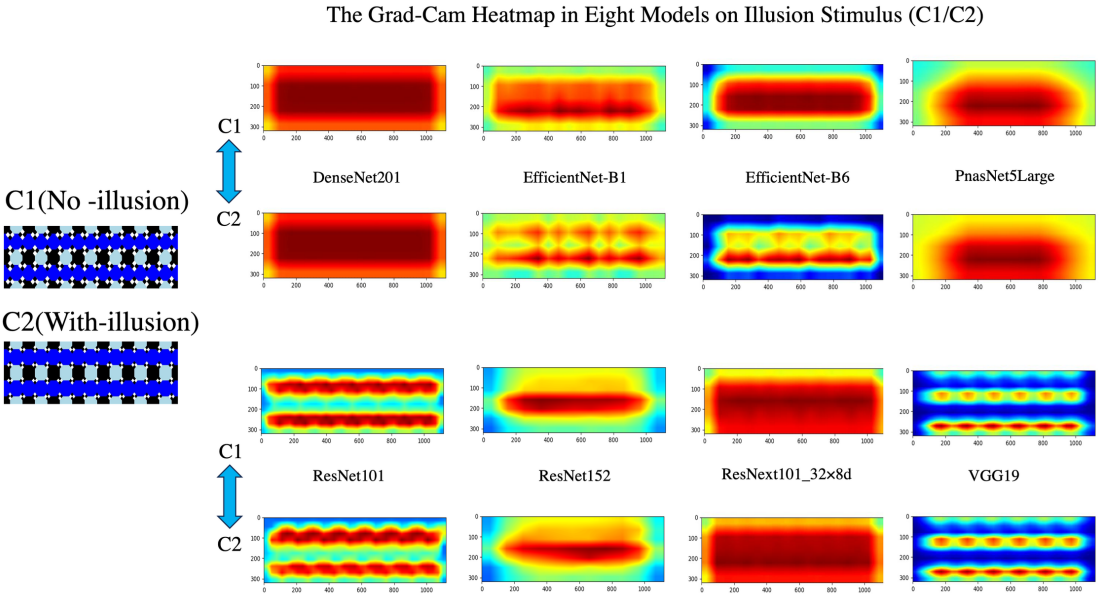


Figure 4.8: The feature attention focus of C1 and C2 on DNNs.

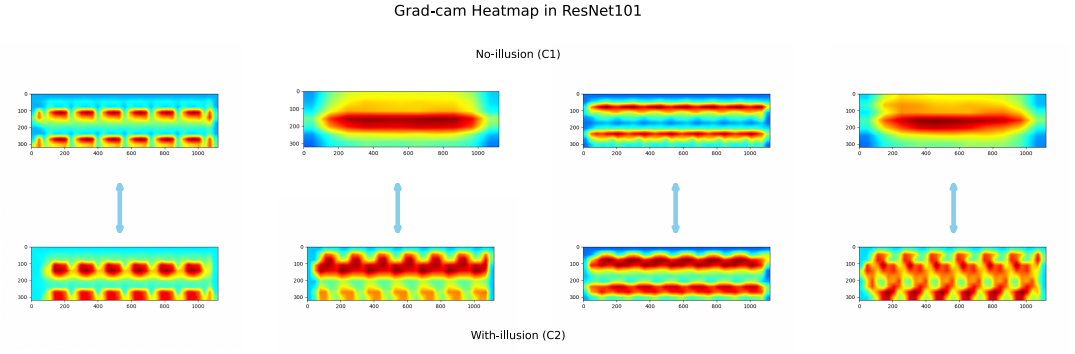


Figure 4.9: The feature attention focus of ResNet101.

In the classification of "With-illusion" (C2) and "No-illusion" (C1), we calculated the L2 distances for correctly and incorrectly recognized instances to further reveal the performance differences among various deep learning models in visual illusion recognition tasks (Fig. 4.10). By combining overall model accuracy with the average L2 distance per layer, we identified the strengths and weaknesses of each model in extracting and classifying illusion features. By combining the average L2 distances and overall classification accuracy of different models, we found significant differences in how models processed visual illusion features. EfficientNet-B1 and EfficientNet-B6 showed greater L2 distance variation in shallow layers but improved in deeper layers. ResNet101 and ResNext101 performed well in processing visual illusion images, with high accuracy, despite showing similarity in some layers for misclassified and correctly classified samples. In contrast, Vgg19 had lower accuracy and higher L2 distances for misclassified samples across all layers, indicating its inadequacy in recognizing illusion features. ResNet152, DenseNet201, and PNASNet showed a balance between accuracy and L2 distance but still had trend for improvement.

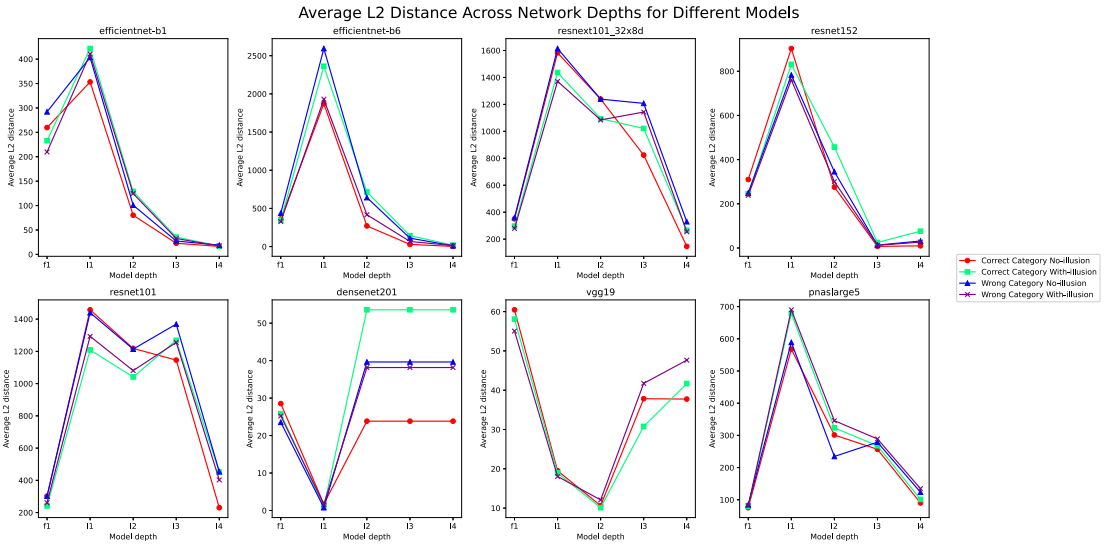


Figure 4.10: The average L2 distance on DNNs.

We focused on the ResNet101 model, which performed exceptionally well in recognizing visual illusions. Figure 4.11 shows the RDM of ResNet101 at different network depths, based on the L2 distance between human-perceived adjusted images and visual illusion images. The strength of the color represents the degree of similarity, with lighter colors indicating higher similarity.

In terms of representational similarity (Fig. 4.11), ResNet101 showed high similarity in shallow layers, which gradually decreased with depth. Similar patterns were observed in untrained networks, though the changes more rapidly.

Figure 4.11: The RDMs of pretrained and self-data trained ResNet101.

Interestingly, based on the RDMs of ResNet101 across different network depths, we explored the effect of the initial layers on visual illusions by training only the initial module of ResNet101 separately. As shown in Figure 4.12, the RDM heatmap of the separately trained initial module of ResNet101 is very similar to that of the full architecture of ResNet101, particularly in the high similarity distribution along the diagonal. Given the mapping relationship between DNNs and the ventral stream, the visual illusion performance of the initial layers may suggest the importance of V1 in responding to visual illusions. This also implies that the high visual illusion performance of V1 potentially influences the IT cognitive layer through other mechanisms.

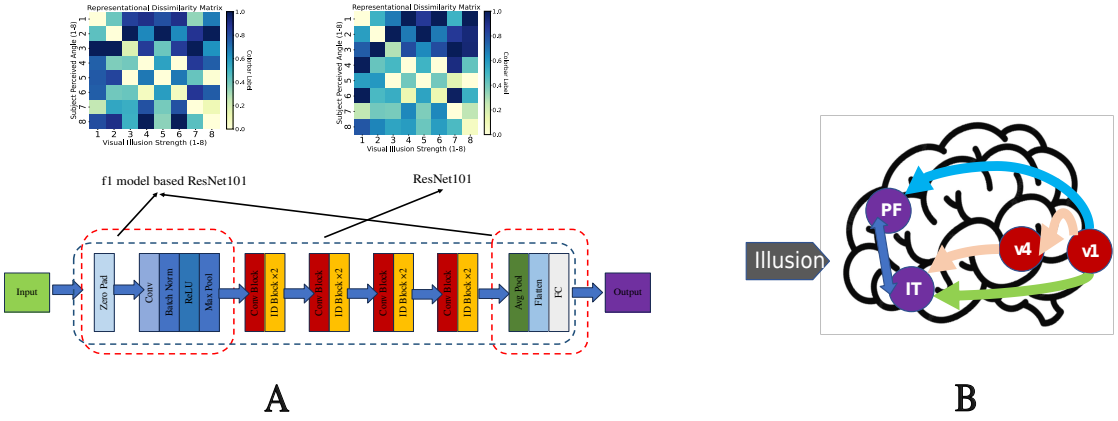


Figure 4.12: The RDM of the initial module of ResNet101 and mapping realtionship.

Chapter 5

DNNs: Spatiotemporal vs Static

In this chapter, we continue to delve into whether video classification models and predictive coding models exhibit more brain-like performance in visual illusions. Specifically, we explore if DNNs can more realistically simulate the human visual mechanism in processing visual illusions by combining self-supervised learning and training strategies that consider spatiotemporal features.

5.1 Main idea

Although previous research focused on feedforward DNN architectures to discuss the universality and potential issues of DNNs in visual illusions, it is crucial to fully consider DNNs in all aspects, especially those with spatiotemporal characteristics. This involves considering the dynamic and complex context-dependent aspects of visual processing.

Recent studies indicate that considering spatiotemporal characteristics and adopting self-supervised learning methods can approximate the human brain's approach in some visual tasks. For example, visual transformers and convolutional neural networks demonstrate a hierarchical similarity to the human visual cortex when dealing with dynamic stimuli [72]. Multimodal and temporal networks perform better in interpreting neural activities of the visual cortex [73]. Recurrent generative networks trained to predict future video frames can capture complex perceptual motion illusions, highlighting the importance of recursion in the neural encoding of dynamic visual scenes [74].

Thus, we use the Müller-Lyer illusion as a case study for model training and testing on spatiotemporal characteristics. The human perceptual lengths from previous Müller-Lyer illusion experiments serve as the perceptual data for this experiment, specifically,

the average perceived lengths with arrows pointing outward and inward in Figure 3.2 as the perceptual group images for testing. Besides, we also keep the control group setting, which includes lines of the same length but with differently oriented arrows. After completing the relevant model training, the testing phase involves comparing and analyzing static classic DNN models with video models.

5.2 Video Models and Training Strategies

5.2.1 Video Dataset

During the preparation of the Müller-Lyer line video dataset, based on the attributes of lines and arrows, we set up two types of datasets (as shown in Figure 5.1(c) and Table 5.1): Type A and Type B. Type A includes videos with varying line lengths and fixed arrow attributes, totaling 2160 videos; Type B includes videos with varying arrow attributes and fixed line lengths, totaling 5616 videos, mainly involving changes in arrow color, angle, and length (Table 5.1).

The specific types of videos in Categories A and B are as follows:

- Line Length Variation (Type A): The line length is set from 110 to 210 pixels, at intervals of 20 pixels, with arrow attributes (such as arrow length, angle, and color) remaining constant. This allows for studying the impact of line length variation on illusions.

- Arrow Length Variation (Type B): The arrow length is set from 10 to 60 pixels, at intervals of 10 pixels, with line length remaining fixed. This helps explore the impact of arrow length on the Müller-Lyer illusion.

- Arrow Angle Variation (Type B): Different arrow angles are set (such as 60°, 80°, 100°, and 120°), with line length remaining constant. These changes help understand the impact of different arrow angles on illusions.

- Arrow Color Variation (Type B): Different arrow colors are selected (such as red, green, blue, yellow, magenta, and cyan), with line length remaining constant. These changes help study the impact of arrow color on line length perception.

As shown in Table 5.1, the dataset parameters are set considering the models' limitations in understanding lines and avoiding catastrophic forgetting. Videos from

Types A and B are used for model training siperately. For Type A (line length variation), we use the pretrained weight on Kinetics-400 video dataset for training, where the line length varies over time, with labels representing the average of the starting and ending lengths. For Type B (arrow attribute variation), same as Type A training setting, but with arrow attributes (such as length, angle, color) varying over time, with labels for the fixed line length.

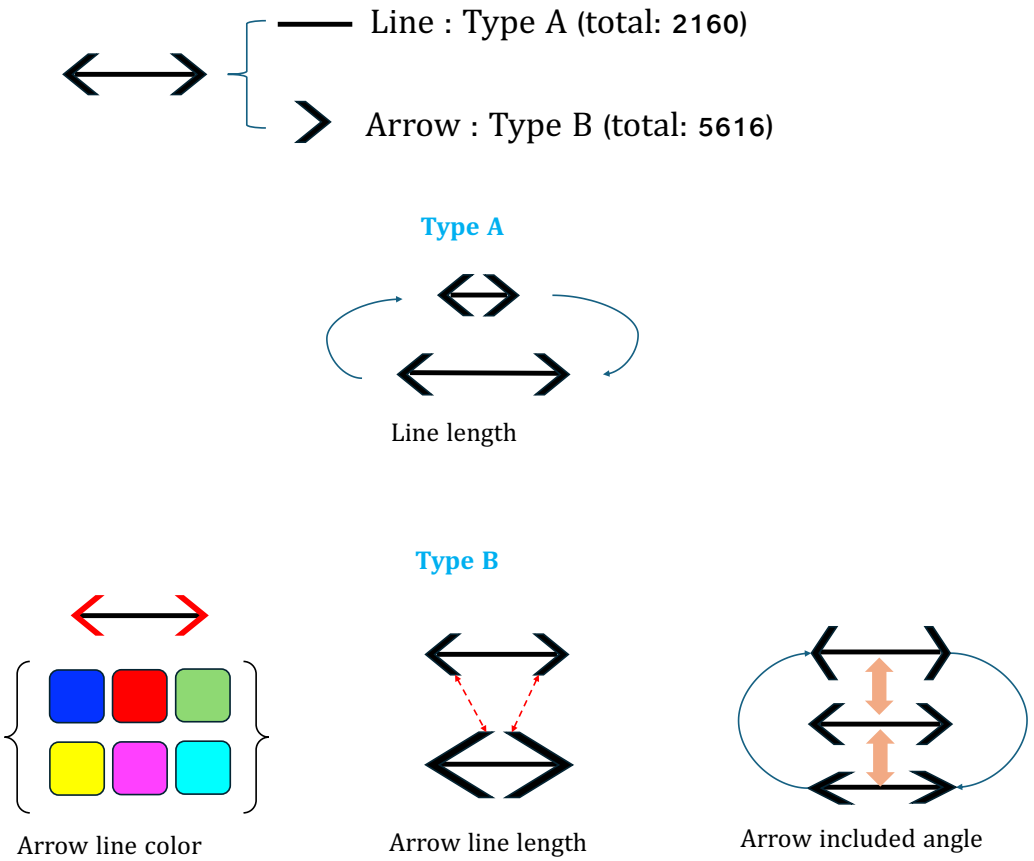


Figure 5.1: The video dataset of Müller-Lyer illusion (Type A and B).

Table 5.1: Müller-Lyer Illusion Video Dataset Overview

Attribute	Type A	Type B
Number of Videos	2160	5616
Variable Attributes	Line Length	Arrow Attributes
Fixed Attributes	Arrow Attributes	Line Length
Length Details	110-210 pixels	10-60 pixels
Arrow Details	Arrow colors: 6; Angles: 60°, 80°, 100°, 120°; Arrow line length: 10 to 60	Line length:110 to 210

5.2.2 Video Models

In this study, we have tested video models on Müller-Lyer illusion. Specifically, we have employed four types of video classification models: R3D-18 [75], MViT-V1-B [76], S3D [77], and Swin3D-T [78], along with the predictive coding model PredNet [20], to test the performance of these five models in handling the Müller-Lyer illusion (as shown in Figure 5.2). The details are as follows:

- R3D-18: A 3D convolutional neural network that extends the ResNet-18 architecture to video data, effectively capturing short-term spatiotemporal features.
- MViT-V1-B: A multi-scale visual transformer that utilizes a hierarchical attention mechanism to capture long-range dependencies and multi-scale features, making it suitable for complex video understanding tasks.
- S3D: Integrates separable convolution, maintaining the ability to learn spatial and temporal features while reducing computational complexity, suitable for efficient video classification.
- Swin3D-T: Extends the Swin Transformer architecture to video tasks, using shifting windows for attention allocation to enable scalable video analysis and capture fine-grained spatiotemporal patterns.
- PredNet: A predictive coding model that processes video data through recurrent prediction and error calculation, excelling in predicting the next video frame and capturing dynamic changes.

The specific configurations of the five models are shown in Table 5.2.

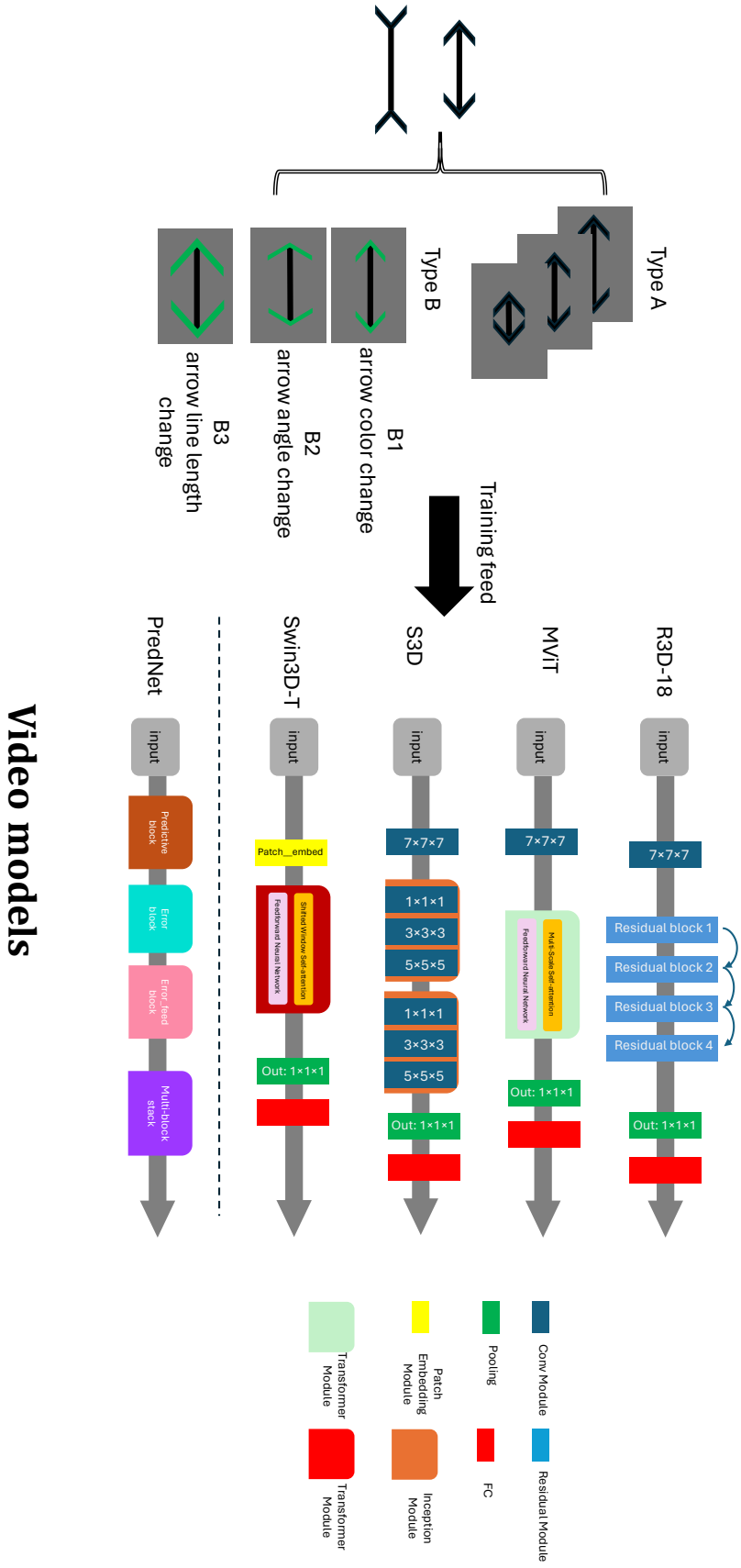


Figure 5.2: The main architecture of five video models.

In our research, the five models can be understood as:

$$\hat{y} = f_\theta(\mathbf{X}) \tag{5.1}$$

where $\mathbf{X}$ represents the input sequence of video frames, $\hat{y}$ is the model's predictive output for the target (e.g., line length or visual illusion effect), and f_θ is the nonlinear mapping function defined by model parameters θ. This mapping covers the entire processing chain from video frames to prediction output, including feature extraction, spatiotemporal analysis, and the final decision layer.

All models are implemented using the PyTorch framework, with the four video classification models implemented using the torchvision library. Training is conducted on NVIDIA A1000, with an initial learning rate of 0.001. We adjust hyperparameters such as learning rate, batch size, and training epochs through cross-validation to achieve optimal performance. Given the unique architecture and tasks of PredNet, we trained PredNet separately and used it only for comparative testing. By using these different models, we aim to utilize video models as brain-like models to comprehensively explore the visual illusion performance of DNNs combining spatiotemporal characteristics and self-supervised learning strategies, thereby further investigating the biological similarity of neural networks in visual mechanisms, especially the universality of visual illusions in DNNs.

Table 5.2: Video Model Configuration Parameters

Model	**Pretrained Dataset**	**Number of Parameters**	**Package**
R3D-18	Kinetics-400	33.2M	Torchvision
MViT-V1-B	Kinetics-400	36.5M	Torchvision
S3D	Kinetics-400	18.0M	Torchvision
Swin3D-T	Kinetics-400	27.6M	Torchvision
PredNet	Kinetics-400	6.9M	Pytorch

5.2.3 Teacher-Student Self-Supervised Learning

In this section, we detail the two-stage training method adopted to optimize the performance of video classification models on the Müller-Lyer illusion video dataset. As shown in Figure 5.3, the first phase includes supervised training of the teacher model, aiming to enable the model to accurately recognize line lengths. The second phase involves teacher-student architecture-based self-supervised learning [79], where the student

model self-optimizes under the guidance of the teacher model without labels, thereby learning and understanding the characteristics of lines. Furthermore, considering that neural networks based on spatiotemporal characteristics are closer to the way the human brain processes information, we explore whether deep neural networks still exhibit visual illusions in self-supervised learning.

The teacher model is trained using a supervised learning method, aiming to minimize the mean squared error (MSE) between the predicted line lengths and the actual lengths. We chose four video classification models—R3D-18, MViT-V1-B, S3D, and Swin3D-T—and replaced their last layer to output a scalar value, the predicted line length. During training with the preprocessed Müller-Lyer video dataset, the dataset is divided into 80% training set and 20% validation set. The training process uses the Adam optimizer, with an initial learning rate set at 0.001, combined with a StepLR learning rate scheduler, reducing the learning rate by tenfold every 30 training cycles. The models are fine-tuned after pre-training on the Kinetics 400 dataset to obtain richer and more general feature expressions, thus achieving better performance in specific tasks.

In the self-supervised learning phase, the student model is trained using the same architecture without preloaded pretrained weights. We designed a compound loss function that includes regression loss and feature consistency loss, aiming to enable the student model to not only estimate line lengths but also maintain feature-level consistency with the teacher model in a label-free environment. The specific expression of the compound loss function is:

$$L = \alpha \cdot \text{MSE}(y_{\text{pred, teacher}}, y_{\text{pred, student}}) + (1-\alpha) \cdot (1 - \text{CosineSimilarity}(f_{\text{teacher}}, f_{\text{student}})) \tag{5.2}$$

Where $y_{\text{pred, teacher}}$ and $y_{\text{pred, student}}$ represent the predicted lengths by the teacher and student models for the same video frame, respectively; MSE is used to measure the error between their predictions; CosineSimilarity evaluates the similarity in feature level between the teacher and student models; and α is a weighting factor used to adjust the proportion of the two types of losses.

Through the training strategy described above, we are able to effectively train video classification models, enhancing their performance on the Müller-Lyer illusion video dataset, while also improving the models' understanding of line lengths and their visual illusions. This teacher-student training method not only increases the models'

prediction accuracy but also enhances their understanding of complex visual illusions.

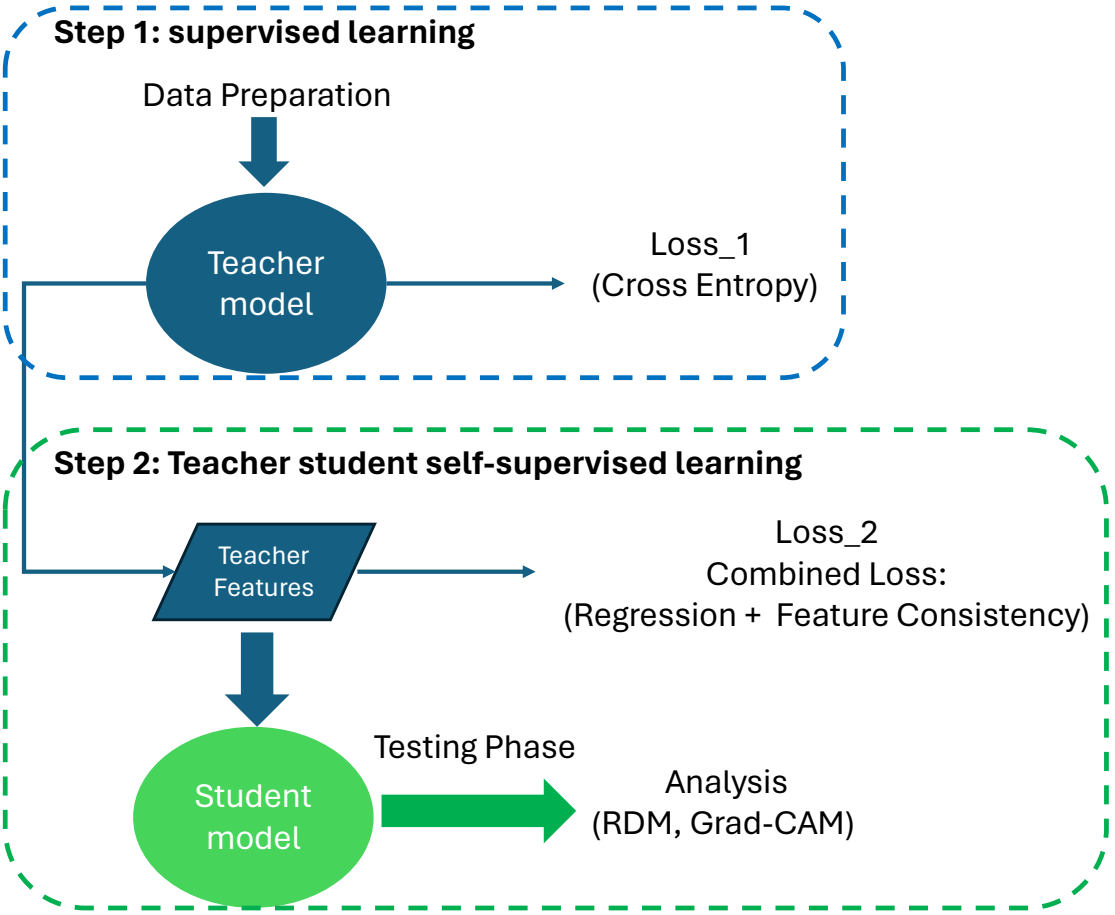


Figure 5.3: The main training strategy through two steps.

5.2.4 Representation Similarity Analysis (RSA)

Representation Similarity Analysis (RSA) is used to compare feature representations extracted by models under different conditions, quantifying the behavioral differences of models when processing the Müller-Lyer illusion. By computing the Representation Dissimilarity Matrix (RDM) from feature vectors, we can observe the models' responses to different visual inputs and analyze whether they exhibit human-like visual illusions.

After the student model completes its training, we extract feature vectors from the model's final module, which correspond to different perceived lengths (200 to 380 pixels) with arrows pointing outward and inward. We then calculate the Euclidean distance between these feature vectors to construct a 10x10 RDM, visualized through a heatmap. The presence of visual illusions is assessed by the distribution of the heatmap, with the primary formula as follows:

$$\mathrm{RDM}_{i,j} = \sqrt{\sum_{k} (X^{(i)}_{\mathrm{out},k} - X^{(j)}_{\mathrm{in},k})^2} \tag{5.3}$$

where $X^{(i)}_{\mathrm{out},k}$ represents the k-th feature value for the i-th perceived length under the condition of arrows pointing outward, and $X^{(j)}_{\mathrm{in},k}$ represents the k-th feature value for the j-th perceived length under the condition of arrows pointing inward. Euclidean distance measures the dissimilarity between two feature vectors.

The RDM calculated allows us to generate a heatmap that visualizes the similarity differences of feature vectors under different input conditions. The horizontal and vertical axes of the heatmap represent different perceived lengths, and the color depth indicates the degree of dissimilarity between feature vectors—darker colors indicate higher dissimilarity, lighter colors indicate lower dissimilarity. This visualization method helps explore and reveal the model's focus on features of the Müller-Lyer illusion lines, thus understanding its internal mechanisms.

5.2.5 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) reveals the focal points of attention of models when processing the Müller-Lyer illusion. By calculating the weighted sum of gradients of the target category against the convolutional layer feature maps, Grad-CAM generates a heatmap to visualize the areas of focus during the model's decision-making process. The main calculation steps are as follows:

$$\alpha^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}} \tag{5.4}$$

$$L^c_{\mathrm{Grad-CAM}} = \mathrm{ReLU}\left(\sum_k \alpha^c_k A^k \right) \tag{5.5}$$

where $A^k \in \mathbb{R}^{u \times v}$ is the k-th feature map output by the convolutional layer, u and v being the width and height of the feature map; α^c_k represents the weight of category c for the k-th feature map; $L^c_{\mathrm{Grad-CAM}}$ is the generated Grad-CAM heatmap. The final heatmap is overlaid on the original image to intuitively show the model's area of focus.

It should be noted that Grad-CAM is not applicable to PredNet, as PredNet is a recurrent neural network used for video prediction, producing the next frame of the video rather than a classification result. PredNet relies on pixel-level prediction errors for

training, not category-based gradients, making it impossible for Grad-CAM to calculate meaningful gradient information. Therefore, we only generate feature heatmaps for the four video classification models.

5.2.6 Total Research Method

This study proposes an integrated framework by combining self-supervised learning and a teacher-student architecture to address the Müller-Lyer illusion, exploring the capacity of deep neural networks (DNNs) to simulate human visual mechanisms. Through a composite training process, the study learns to understand line lengths from video data and uses Representation Similarity Analysis (RSA) and Gradient-weighted Class Activation Mapping (Grad-CAM) to assess the models' performance in visual illusions and explain the decision-making process.

Step One: Composite Training Process The composite training process combines self-supervised learning and feature analysis. The target optimization function (loss function) is defined as:

$$L_{supervised} = \text{SupervisedLoss}(y_{\text{true}}, y_{\text{pred, teacher}}) \tag{5.6}$$

$$L_{self} = \lambda \cdot \text{Self\_Loss}(y_{\text{pred, teacher}}, y_{\text{pred, student}}) + \gamma \cdot \text{FeatureConsistency}(f_{\text{teacher}}, f_{\text{student}}) \tag{5.7}$$

$$L = L_{supervised} + L_{self} \tag{5.8}$$

Where:

- SupervisedLoss is the mean squared error, used to measure the difference between the teacher model prediction and the true label.
- Self_Loss the self-supervised learning loss of the student model using the teacher model output as the pseudo label.
- FeatureConsistency represents the feature consistency between the teacher and student models, usually measured by cosine similarity.

- λ and γ are hyperparameters used to adjust the weights of different parts in the loss function.

Step Two: Model Behavior Assessment By using RSA and Grad-CAM, we assess and visualize model behavior, generating heatmaps and dissimilarity matrices that provide intuitive feedback and deep insights into how the models handle visual illusions.

Specifically, the comprehensive sensitivity or responsiveness of the models to visual illusions, S_{DNN}, is evaluated through the following formula:

$$S_{DNN} = \text{Analyze}(\text{Grad-CAM}(I), \text{RDM}(I_{\text{perception\_out}}, I_{\text{perception\_in}})) \tag{5.9}$$

where:

- Grad-CAM(I) represents the heatmap generated by applying Grad-CAM to the input image I, highlighting the areas of focus during the DNN's decision-making process.
- RDM($I_{\text{perception\_out}}, I_{\text{perception\_in}}$) calculates the representational difference between lines with perceived lengths where the arrows point outward $I_{\text{perception\_out}}$ and inward $I_{\text{perception\_in}}$, measured using Euclidean distance.
- Analyze combines the visual attention heatmap generated by Grad-CAM and the representational dissimilarity matrix calculated by RDM for analysis.

This comprehensive approach enables us to effectively train video classification models, which perform excellently when processing the Müller-Lyer illusion video dataset, and further explore the universality and spatiotemporal characteristics of visual illusions to determine whether DNNs exhibit brain-like properties.

5.3 Results

5.3.1 Model Understanding of Line Length and Basis for Visual Illusions

Compared to previous experiments, merely testing the L2 distance between the perception group and the control group to verify the presence of visual illusions is insufficient. We need to ensure that the model can indeed recognize and understand line lengths, and then use the test situation of the perception group to corroborate the occurrence

of visual illusions. Figure 5.4 shows the video model's ability to understand line length and its performance of visual illusions.

As shown in Figure 5.4A, we first calculate the Euclidean distance of feature vectors for lines with the same arrow orientation and the same labeled length in both the perception and control groups. This is to test whether the model can accurately understand line lengths, as the actual lengths of lines under the same labeled length in the two groups are inconsistent. The length of the bars reflects the model's ability to understand line lengths, especially under two types of video datasets. On this basis, calculating the Euclidean distance of feature vectors for lines with different arrow orientations under the same labeled length within each group visually demonstrates the model's visual illusions (Figure 5.4B). Simply put, the greater the L2 distance values in Figures 5.4A and the smaller the L2 values for the perception group in Figure 5.4B, the higher the authenticity of the model's visual illusions.

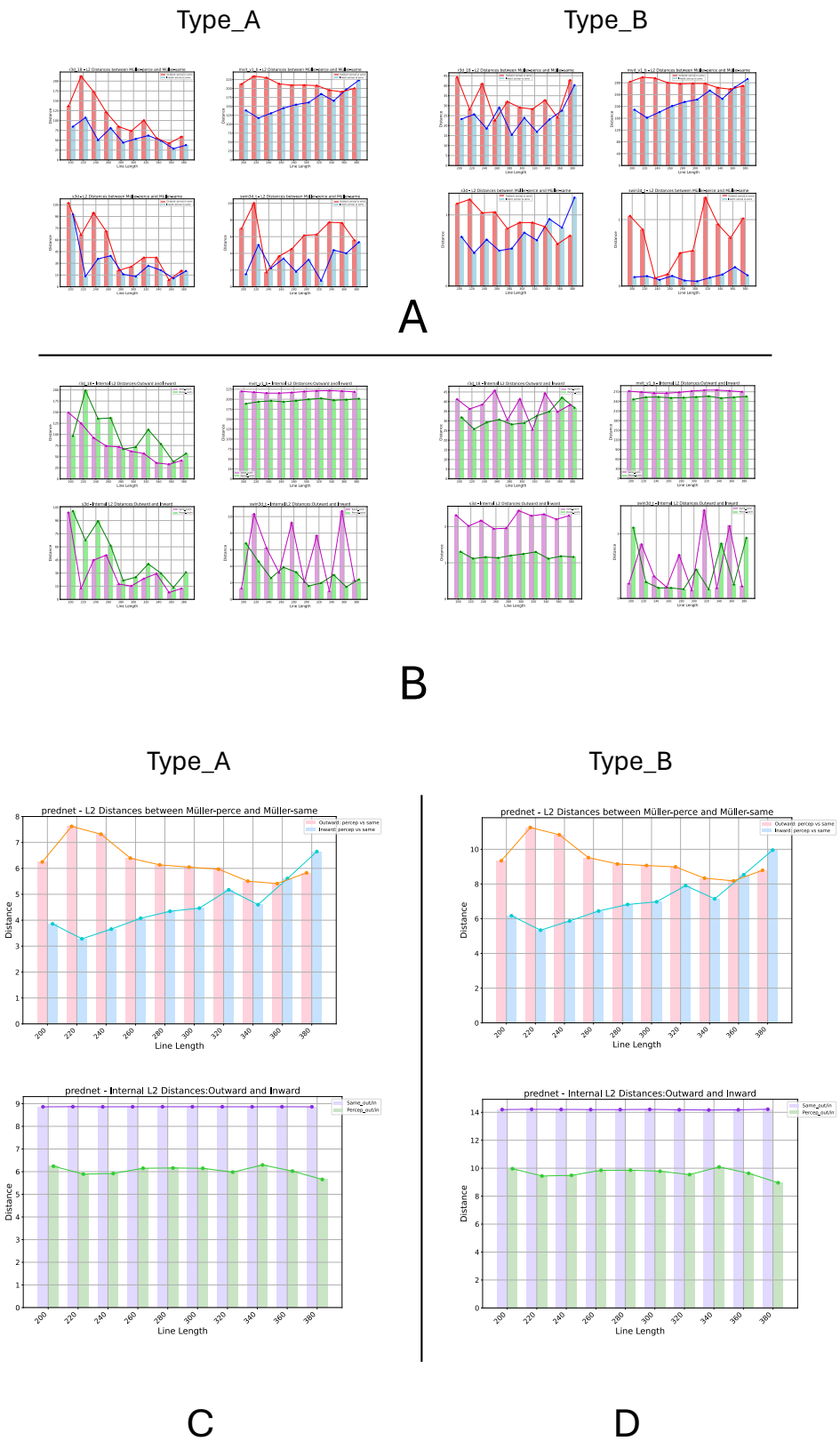


Figure 5.4: L2 distance between same group and different group. A: The L2 distance of 10 length label on same arrow orientation from perception group and control group. B: The L2 distance of 10 length label on different arrow orientations within two groups.

It is evident that under training with type A and type B datasets, the four models exhibited different behaviors. From the left side, Figure 5.4(A) Type_A, MViT-V1-B shows the greatest dissimilarity among the four models, indicating a significant difference in the lengths of Müller-Lyer lines between the perception and control groups. Although all four models show varying distributions of dissimilarity, S3D and R3D-18 exhibit a trend of decreasing dissimilarity with increasing labeled line lengths. Similarly, on the right side in Type_B, MViT-V1-B still shows the greatest difference, but the other three models show higher dissimilarity compared to Type_A, especially S3D and R3D-18, with high dissimilarity. Compared to training with line length variations (Type_A), the four models under training with arrow attribute variations (Type_B) are more sensitive to line lengths. However, Swin3D-T performs worse in Type_A when arrows point inward. In Figure 5.4(B), MViT-V1-B shows high dissimilarity in both types, suggesting a lower likelihood of visual illusion manifestation. Interestingly, Swin3D-T shows very small L2 distances within the perception group (Percep_out/in), which might indicate a similar display of visual illusions. In contrast, the other two models, S3D and R3D-18, show lower L2 distances in Type_B than in Type_A, and the internal L2 distances of the control group (Same_out/in) are greater, potentially indicating a greater possibility of visual illusions under arrow variation training in Type_B. In addition, from the PredNet situation (Fig. 5.4(C,D)), Type_A and Type_B show similar distributions and trends in calculations within the same arrow orientation type and within groups with arrows pointing inward and outward, with not much difference.

5.3.2 Temporal vs. Static Characteristics on RDMs

To deepen our understanding of the display of visual illusions, we extracted feature vectors from the last layer before the decision module to build a Representational Dissimilarity Matrix (RDM), which visualizes whether the model exhibits visual illusions. We calculated the Euclidean distances (L2) for the inward and outward arrow orientations within both the perception and control groups. Figures 5.5A and B display the RDM heat maps under two training sets, while Figure 2.2C shows the RDM heat maps of six pre-trained brain-like DNN models. In the heat maps, the darker the color, the greater the similarity.

Figure 5.5 show that within the perception group, R3D-18, MViT-V1-B, and S3D

exhibit high similarity on the diagonal in both Type_A and Type_B RDMs. This indicates that under the same length labels, the models perceive the line lengths of arrows pointing outward and inward as being the same. Conversely, the RDM for the control group shifts slightly upward near the diagonal, suggesting that under the same length labels, lines with arrows pointing outward appear longer than those with arrows pointing inward. This means these models exhibit visual illusions similar to human perception. PredNet also displays a similar pattern. However, the similarity distribution in Swin3D-T's RDM is irregular, possibly indicating that this model does not effectively understand the Müller-Lyer illusion and does not exhibit similar visual illusions.

Interestingly, some models used for static image classification tasks also exhibit similar phenomena (Figure 5.5C). AlexNet, VGG19, and ResNet101 show higher similarity at the diagonal, while the similarity in the control group shifts upward. The performance of the other three models is less clear or irregular, which similar the Swin3D-T. Compare these two different characteristic DNNs' RDMs, video model presented more clear high similarity on diagonal than statics DNNs.

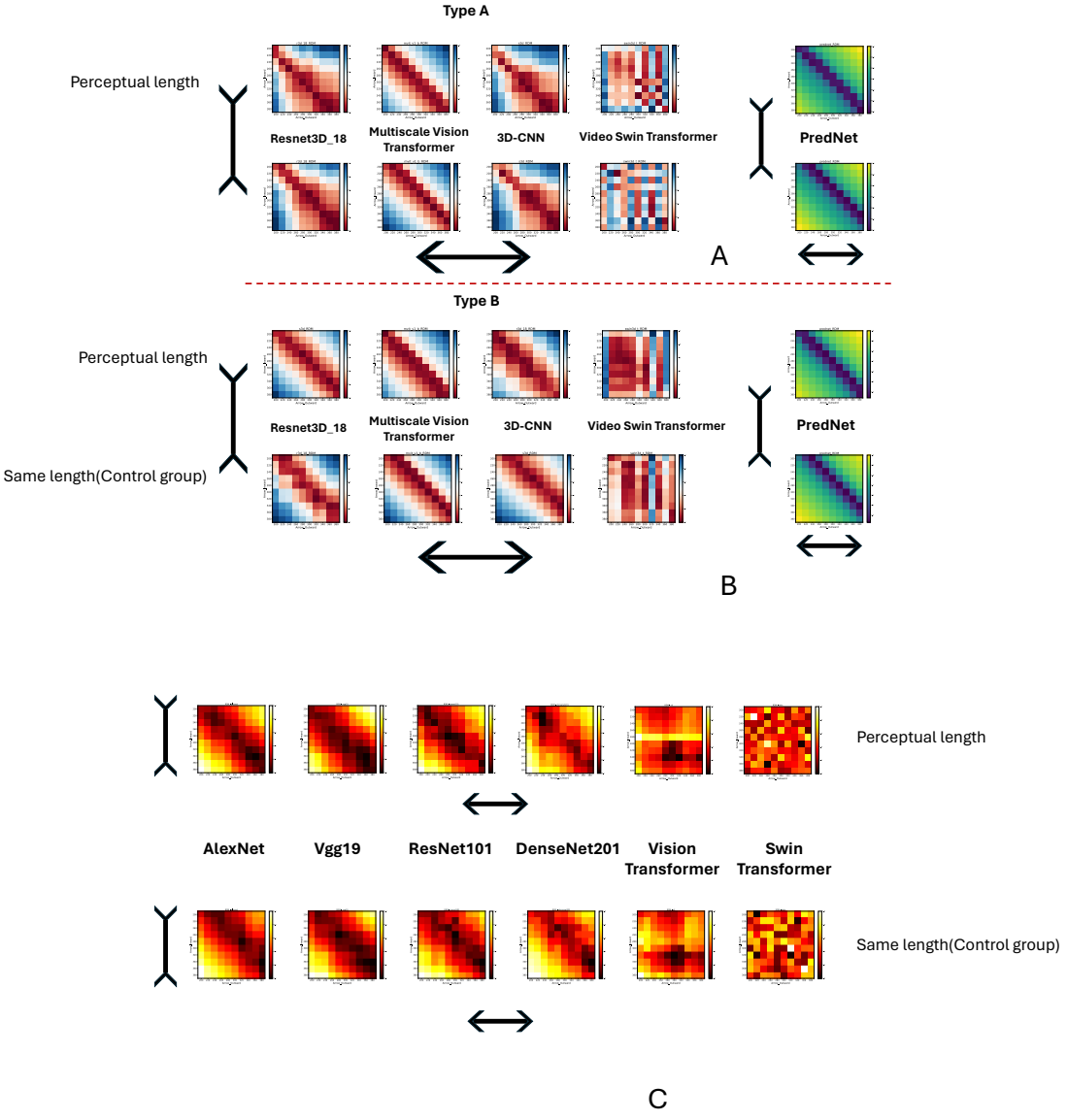


Figure 5.5: RDMs of perception group and control group on video model and statics model.

5.3.3 Temporal vs. Static Characteristics on Grad-CAM

To more visually illustrate the models' feature attention on Müller-Lyer illusion , we used Grad-CAM, a post-hoc explanation method. We visualized the feature heat maps for four models under two types of datasets (Figure 5.6). The upper figure A displays the Grad-CAM heat maps for four video classification models, and the lower figure B shows the Grad-CAM heat maps for six static vision models.

In the Type_A dataset, R3D-18 and S3D primarily focus on the overall features of the line, showing consistent behavior. The other two models with transformer architectures, MViT-V1-B, focus on the overall features of the line and arrows, while Swin3D-T focuses on the entire picture, although it also pays attention to the line itself. In the Type_B dataset, the focus areas of R3D-18 and S3D differ, with R3D-18 focusing more on the line itself and the left area, while S3D focuses on the line and the area below. The focus points of MViT-V1-B and Swin3D-T are essentially consistent with those in Type_A.

Furthermore, the models exhibiting visual illusions, such as AlexNet, VGG19, and ResNet101, focus on the arrows to varying degrees. DenseNet201's feature focus is similar to R3D-18 in Type_A. The Vision Transformer only focuses on the line itself, paying less attention to other areas. The Swin Transformer focuses on a small area below the line, not showing a significant feature tendency towards the line or arrows.

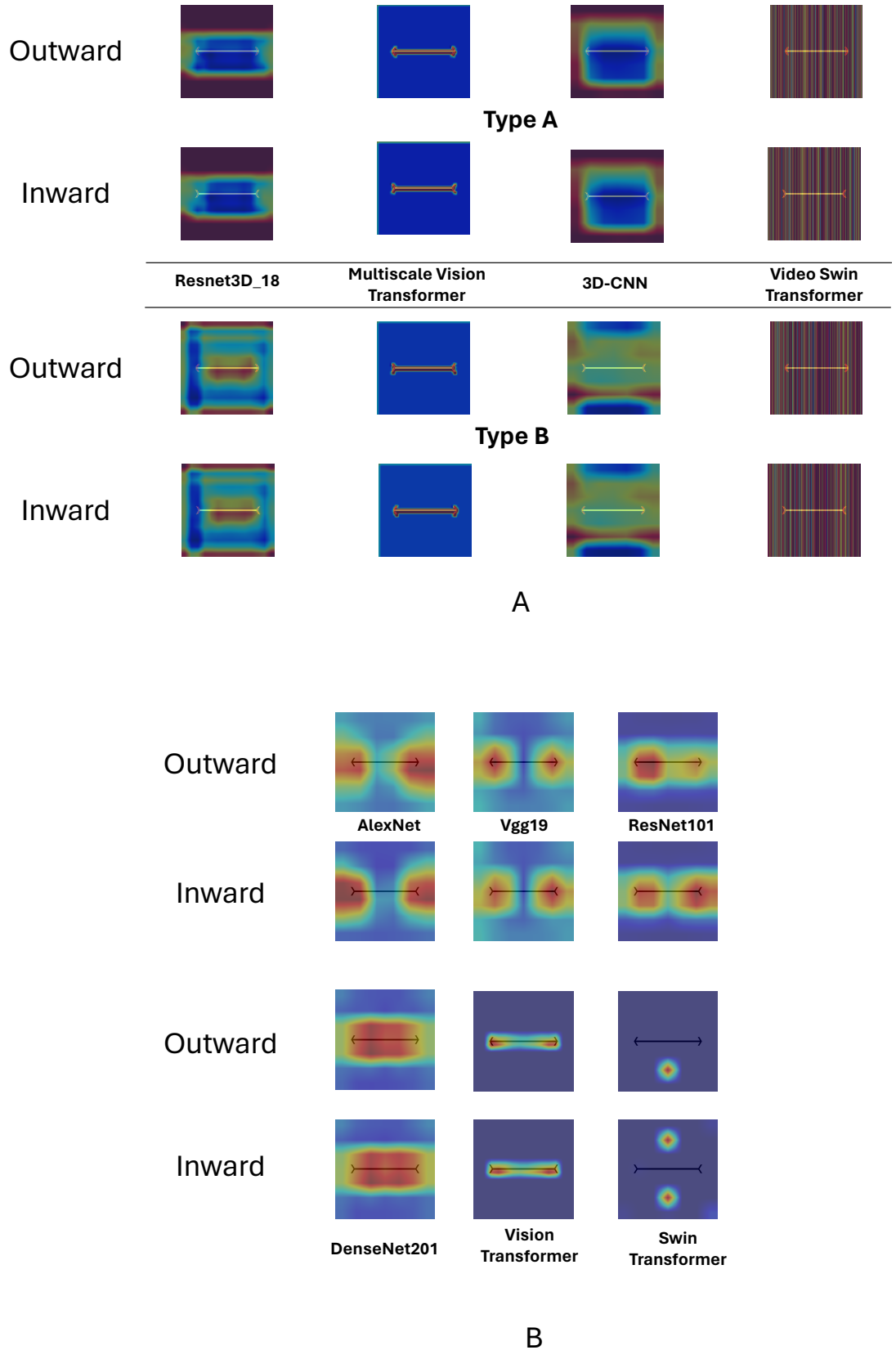


Figure 5.6: The heatmaps of feature focus from four video classification models and statics DNNs

Chapter 6

Mapping relationship between DNNs and visual pathway through fMRI and optical illusion

This chapter mainly explores the relationship between deep neural networks and regions along the visual pathway through visual illusion and fMRI study.

6.1 Experimental Introduction and Procedures

Previously, we found that the manifestation of visual illusions in the primary modules of DNNs may suggest that V1 play an importran role on response of optical illusion. Based on this, we have designed an fMRI experiment focusing on Skye's Oblique Grating illusion to further investigate the correlations in the mapping relationship between DNNs and the ventral stream. As shown in Figure 6.1, we designed three steps for the experiment:

1. Using the previous average perceived angle, which was categorized into eight illusion strength (Figgure 4.3) .

2. Six DNN models were selected for visual illusion testing, with preparations for visualization using RSA and CAM methods.

3. Two types of visual illusion stimuli groups were designed for task-based fMRI experiments, focusing on both collective and individual differences.

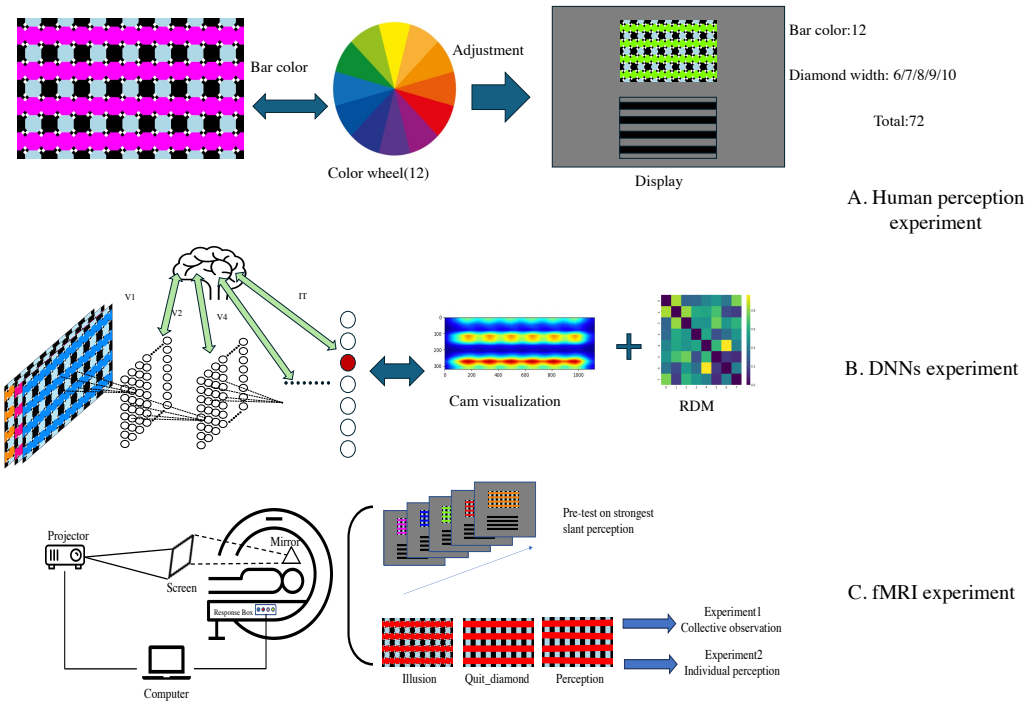


Figure 6.1: The main three step on exploring the mapping relationship between DNNs and ventral pathway.

6.1.1 Visual Illusions and DNNs

Based on the prior distribution of perceived lengths as shown in Figure 4.3, eight levels of visual illusion strength were set for the stimulus images. Considering the performance of the pre-trained models, which was similar to that after training on separate datasets, pre-trained models were used for testing visual illusions. Models selected for the study included AlexNet, Vgg19, Inception-V3, ResNet101, ResNext101_32x8d, and DenseNet169. Then RDM, GradCAM heatmaps, and fMRI correlation results were used to analyze the potential mapping of DNNs and ventral pathway.

6.1.2 fMRI Experiments

To further investigate the similarities and differences between human brains and deep neural networks (DNNs) under visual illusion conditions, we employed functional magnetic resonance imaging (fMRI). Our focus was on regions of interest (ROIs) within V1, V2, V4, and the inferotemporal cortex (IT), guided by the mapping relationships between DNNs and the ventral pathway. The experimental design included three types of image stimuli: illusion images, non-illusion images (images without illusion elements), and perceptual matching images, as depicted in Figure 6.1 C. The non-illusion and

perceptual matching images corresponded to the illusion images, providing a basis for comparison.

- Experimental stimuli: Participants viewed image stimuli inside an MRI scanner through a ProPixx DLP LED projector, which projected images onto an internal screen (size: 706mm wide x 397mm high) with a resolution of 1920x1080 and a refresh rate of 60Hz. Participants viewed the screen through mirrors mounted on the head coil from a distance of 132cm. The experiment included a pre-test and two experimental phases, "collective observation" and "individual perception." During "collective observation," three types of images were randomly presented across eight blocks in a mixed design, with 24 images per block; in "individual perception," block design was used, with the strongest illusion intensity determined by pre-testing, and each stimulus type divided into four blocks, each with eight images. Each image was displayed for 1.5 seconds, followed by a rest of 0.5 seconds, with a rest period of 8 seconds between blocks to minimize cross-interference.

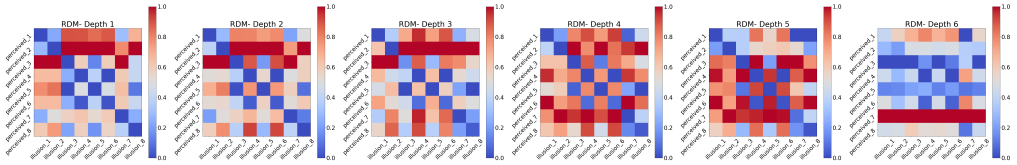
- Data collection: The study was conducted at the Brain Communication Research Center of Kochi University of Technology using a 3T Siemens Prisma MRI scanner and a 24-channel head coil to collect structural and functional data. Structural images were acquired using a T1-weighted MPRAGE sequence, and functional images were collected through a 2D EPI sequence, ensuring image quality and analytical accuracy.

- Preprocessing and analysis: Data preprocessing and analysis were performed using fmriprep [80] and freesurfer [81], ensuring high-quality data. A general linear model (GLM) was initialized using the nilearn [82] library, with appropriate settings for repetition time (TR), slice timing reference, and Gaussian filtering. Masks adapted to the resolution of the BOLD images were created for each ROI [83, 84], and beta values were calculated for each condition, reflecting the activity level of the respective areas. The Mann-Whitney U test [85] was used to assess activity differences under different conditions, with significant results highlighted in charts. Bar charts of average beta values for each ROI under different conditions were plotted for each participant, and data from all participants were combined to provide an overview of brain responses under different task conditions.

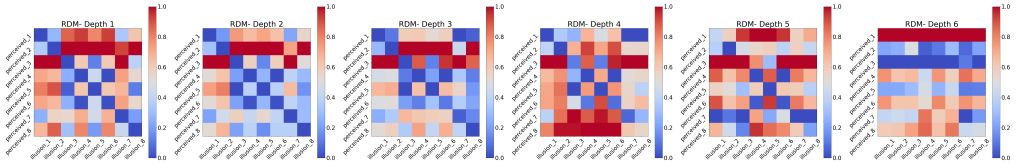
6.2 Result

6.2.1 Heatmaps of Visualization

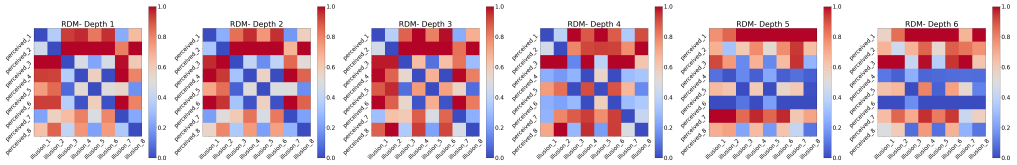
From the RDM heatmaps (Fig. 6.2), all six models displayed a high similarity distribution along the diagonals within their primary modules. As the depth of the models increased, the similarity gradually decreased and eventually disappeared. This suggests that visual illusions in the primary modules might influence the final decision-making process, causing the models to still make biased judgments. Additionally, as shown in Figure 6.3, the features focused on by the models are evident in the two CAM methods. VGG19 and ResNet101 both focused on features of the lines and also included the areas between the lines. The other models showed varying degrees of focus on different areas. Combined with earlier findings, focusing on the areas of the lines themselves further demonstrates the potential for models to exhibit visual illusions. Interestingly, the pretrained VGG19, unlike when trained solely on skewed datasets, also displayed visual illusions, suggesting that training VGG19 on more general and expansive datasets still have the potential to mimic brain-like processing. Furthermore, it hint that DNNs easily influenced by dataset.



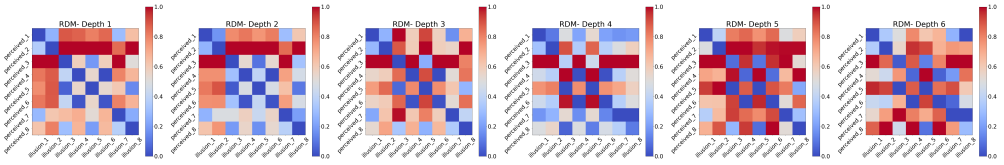
AlexNet



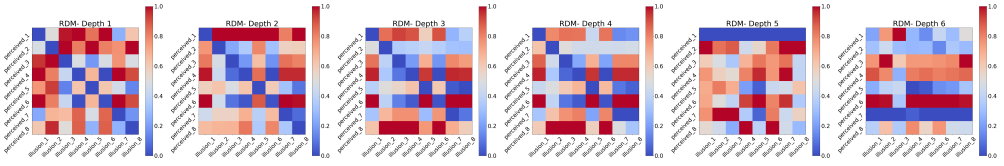
Vgg19



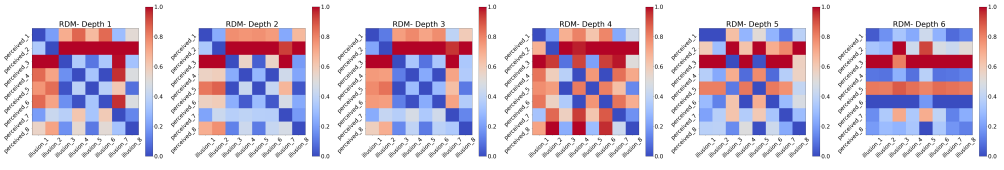
ResNet101



ResNext101_32×8d



Inception-v3



DenseNet169

Figure 6.2: The RDMs of six on C1 and C2.

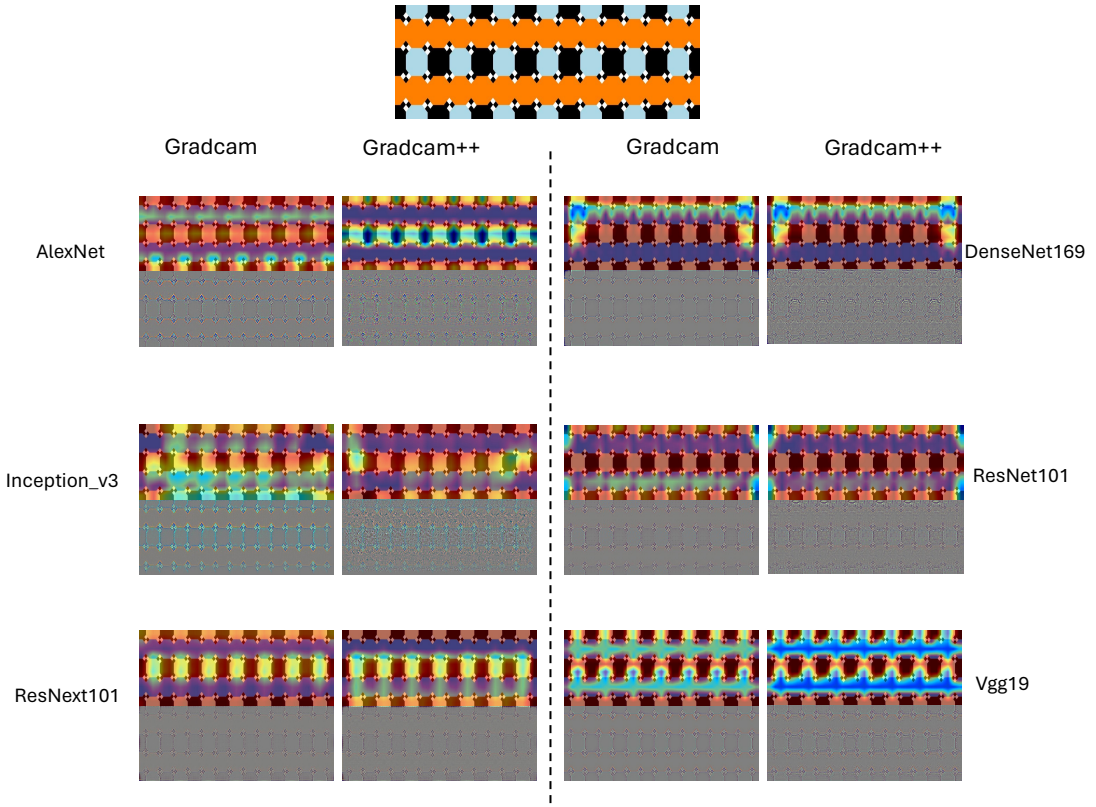


Figure 6.3: The attention heatmaps of six DNNs on illusion.

6.2.2 ROIs Response

In the "collective observation" experiments, we systematically analyzed the responses of specific brain regions of interest (ROIs) such as V1, V2, V4, and IT under conditions of illusion, non-illusion, and subjective perception. The activation in the V2 region was most prominent under illusion conditions, with an average beta value of 0.4658, highlighting its crucial role in deciphering visual illusions. Conversely, the strongest response under non-illusion conditions was observed in the V1 region, with an average beta value of 0.2586. This condition revealed significant differences in activation levels between V1 and V4, showing different modes of processing visual information. The subjective perception condition showed relatively lower activation across all ROIs, especially a significant decrease in activity in the IT region (average Beta: 0.0606), reflecting its minor role in processing specific visual stimuli.

In a deeper study of individual perception, significant differences were observed in participants' responses to illusions. For example, participant 01 had the strongest response in the V2 region to illusions, with a beta value of 0.8075, while participant 03 showed a reduction in activity in the V1 region under subjective perception conditions

(Beta: -0.0783), possibly indicating different perceptual processing mechanisms. These individual differences highlight the variability of the visual system among individuals and personalized responses to visual stimuli.

By comparing "collective observation" and "individual perception," we found that significant activation in the V2 region under illusion conditions supports its important role in decoding complex visual information. Meanwhile, enhanced activation in the V1 region under non-illusion conditions reflects its dominant role in processing basic visual elements, particularly when the illusion effects are weakened or removed. Additionally, the generally low activation across all ROIs under subjective perception conditions, particularly in the IT region, reveals adaptive adjustments of the visual system when processing these specific visual stimuli. These findings, compared with the response patterns of deep neural networks (DNNs), show some commonalities in visual processing strategies between the human brain and DNNs, especially in the activities of primary visual processing areas.

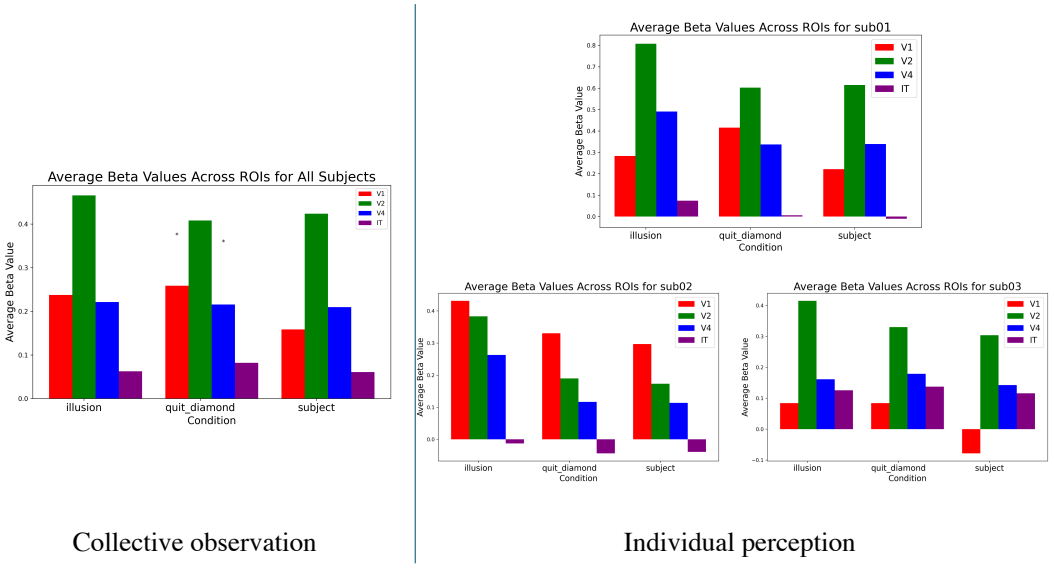


Figure 6.4: Average response distribution under specific ROIs (V1/V2/V4/IT).

Chapter 7

Discussion

This study tests the performance of visual illusions by setting up various types of DNNs and introduces visualization methods for interpreting the internal mechanisms of visual illusions: visual interpretations combining representational similarity analysis with post-hoc interpretability of class activation maps. This elaborates on the potential and limitations of DNNs as visual model through the testing of visual illusions. Particularly, the results from our multiple experimental steps show that while DNNs exhibit excellent performance in visual tasks, they possess unique advantages in brain-like features. However, there remains a significant gap between them and actual visual mechanisms. Certain regularities in visual illusions provide guidance for brain-like modeling, yet DNNs themselves have many limitations and shortcomings, necessitating extensive foundational visual research and support from advancements in neuroscience. Next, we will discuss specific research questions based on the study findings.

7.1 How Do DNNs Respond to Different Types of Visual Illusions?

The response of DNNs to different types of visual illusions is diverse, as reflected in their sensitivity and decoding methods. For example, in our Chapter 3 testing on five types of illusions—focusing on color, brightness contrast, length, angle, and perception—it is discussed. Regarding color sensitivity, among 12 DNN models, only two align relatively well with human perception of color depth rankings (Fig. 3.9B). There is no regular pattern in the ranking distribution among the models, but an increase in network depth (Fig. 3.9C) leads to changes in color ranking, indicating a change in color sensitivity,

though this change is only apparent in the last module. In terms of feature focus visualization (Fig. 3.10), DNNs also show significant differences, with ResNext101 recognizing the entire color rectangle and focusing on the whole area, while other models focus on partial areas. This differs from our understanding of vision; DNNs cannot comprehend color and its resulting shapes, affecting attention differences in color depth rankings. Moreover, ResNext101 does not exhibit a performance closer to the depth ranking of human subjects.

It seems that DNNs clearly respond more to intuitive physical properties, especially in illusions like the Müller-Lyer illusion, which involves a combination of straight lines and arrows. From related studies, including those on static models under pre-training and video models trained with a teacher-student self-supervised strategy, performances similar to human perception have been observed. From the perceptual group's RDM (Fig. 3.5,3.16,5.5), high representational similarity along the diagonals indicates that the model's judgment of line lengths is close to human. In the GradCAM responses (Fig. 3.7,3.15,5.6), it is generally found that similar to humans, there is a focus on the arrows at both ends of the line, which are more prone to visual illusions, including the areas near the arrows. Similar is the case with the Poggendorff illusion, another perceptual visual illusion. However, once the composition of the illusion becomes more complex or more integrated, it is difficult for the model to demonstrate evidence of visual illusions from RDM. From the feature focus heatmap of the Zöllner illusion in Figure 3.20, although high diagonal similarity in RDM indicates potential visual illusion performance, the heatmap shows all models considering the whole stimulus features, indicating that DNNs still lack adequate response to angles or are easily "attracted" by the overall shape.

7.2 What Are the Similarities and Differences Between Human and DNN Perceptions of Visual Illusions?

By using representational similarity analysis (RDM) and class activation maps (GradCAM) visualization techniques, we can initially explore the internal mechanisms of how DNNs handle visual illusions. RDM analysis reveals that different levels of feature extraction have varying sensitivities to visual illusions, with lower levels typically responding to simple visual information, while higher levels involve more complex integration of visual information. GradCAM shows the image areas focused on by DNNs when

making decisions, revealing key features that might be considered during the processing of illusions. These methods help us understand the internal workings of DNNs when dealing with visual illusions.

However, the difference lies in the internal mechanisms of DNNs in processing these illusions, which often rely on training data and optimization of network parameters, rather than the biological logic and neuronal activity in the human visual system. We found that under settings such as pre-training, the vast majority of models show obvious flaws, or a preference for more intuitive physical features. For this, specific datasets (Fig. 4.3) were used in Chapter 4 to train models to understand the concept of tilt. The test accuracy of DNNs was not particularly high, with the best judgment on tilt angle reaching about 90% with ResNet101. Differently, Vgg19, which responded to several visual illusions, did not react to the tilt illusion. This is inconsistent with findings that closely mimic human visual processing, yet Vgg19 and ResNet101 both focused on the stripes themselves, aligning with our observations of this illusion.

7.3 How Do Different DNN Architectures Compare in Their Ability to Simulate Visual Illusions?

Different DNN architectures show varying abilities to simulate visual illusions. Generally, convolutional neural networks (CNNs) are more effective in handling geometric illusions in images due to their strong spatial capabilities. Recurrent neural networks (RNNs) might be better at handling illusions that require analysis over time, such as motion illusions. Choosing the appropriate network architecture is key as it determines the model's sensitivity and ability to process illusions. However, our experimental results do not fully support this.

From the perspective of static characteristics, increasing network complexity does not necessarily result in better performance on visual illusions. Notably, classic architectures like DenseNet201, exhibit performance similar to simpler ResNet and Vgg series. However, in terms of feature attention, they are relatively close to these models. As for the currently popular Transformer architectures in DNNs, none of the model types demonstrated visual illusions, particularly irregular performances in RDM.

Looking at DNNs with spatio-temporal characteristics, architectures based on convolutional neural networks (CNNs) like R3D_18 and S3D have certain advantages in handling both static and dynamic visual information, while architectures based on

self-attention mechanisms like Swin3D_T perform better in managing complex spatio-temporal information. Moreover, models that integrate multiple perspectives, such as MVIT_V1_B, demonstrate higher robustness and precision in simulating visual illusions. Visualization results indicate that purely convolutional architectures like R3D_18 and S3D somewhat exhibit visual illusions. Interestingly, the distribution of high similarity in RDM compared to static models and other video models is clearer, suggesting that models based on brain predictive coding principles might be closer to some brain mechanism processes,

In terms of single architecture or static characteristics, DNNs in feature attention are closer to human attention traits, such as the arrows on the Müller-Lyer lines. In contrast, models with spatio-temporal characteristics still focus on the whole line and arrows, although they also share similarities with some static characteristic DNNs, such as ResNext101 (Fig. 3.7, 5.6). Overall, more advanced models with better visual task performance do not broadly exhibit a visual illusion mechanism response similar to humans, although inspired by brain attention mechanisms. This also indicates the current gap between neural networks and visual mechanisms.

7.4 What Are the Computational Principles Underlying the DNNs' Ability to Simulate Visual Illusions?

The computational principles behind DNNs' ability to simulate visual illusions involve multiple aspects, including feature extraction, hierarchical information processing, and the model's training strategies. In terms of feature extraction, DNNs capture spatial and temporal characteristics of images through multiple layers of convolution, allowing them to somewhat mimic human visual perception. Hierarchical information processing involves lower network layers capturing simple edge and texture features, while higher layers integrate these features into complex visual representations. Additionally, training strategies, including data augmentation and multitask learning, significantly impact the ability of DNNs to simulate visual illusions. These computational principles provide a theoretical basis for understanding and optimizing the performance of DNNs in simulating visual illusions based on their learning capabilities and adaptability. Networks minimize prediction errors on specific tasks by adjusting their weights and parameters,

inadvertently learning to simulate illusions. Moreover, this capability may also be related to the network's depth, type of layers, activation functions, and regularization strategies during training.

Based on the BH-Score, architectures that are simpler and closer to brain-like. Looking at two classic models, VGG19 and ResNet101, both perform well in six visual illusions, and their feature attention is also quite close to human vision. Combined with the similarity between VGG19 and ResNet and human visual processing mechanisms, this may suggest that visual illusions typically involve a more singular early response in the visual system.

7.5 Can the Findings From DNN Simulations of Visual Illusions Inform the Development of More Advanced AI Systems?

In this study, we discuss how deep neural networks (DNNs) simulating visual illusions can inform the development of more advanced AI systems. Initially, through studies in neuroscience and behavioral data, we find that DNNs exhibit behaviors in processing visual illusions similar to humans. For instance, research on the Müller-Lyer illusion shows that the decision-making distribution in DNNs mirrors that of humans. These findings suggest that DNNs can serve as effective tools for studying human visual illusions.

However, although DNNs demonstrate potential in simulating visual illusions, their performance and behaviors are significantly influenced by the model architecture and training data. Our study indicate that different DNN architectures vary in their sensitivity to visual illusions, which may relate to differences in their internal mechanisms and learned representations. For example, post-hoc interpretability analyses using Class Activation Mapping (CAM) and Representational Similarity Analysis (RSA) provide insights into the DNNs' decision-making processes, but these methods struggle to capture the underlying causes of these decisions.

Moreover, preliminary results from our fMRI experiments suggest that the genesis of visual illusions may occur at the level of the retina and early visual areas such as V1, aligning with the behaviors of DNNs' primary modules, implying a similarity with human visual pathways. This observation indicates that further research on DNNs could enhance our understanding of visual processing mechanisms. Thus, emphasizing

the study of these primary modules can help us gain deeper insights into how DNNs handle complex visual tasks, especially in recognizing and interpreting visual illusions.

To improve DNNs' capabilities and accuracy in simulating visual illusions, we need to explore new architectural improvements. This includes developing models that can more finely mimic human visual processing features, such as enhancing the network's spatiotemporal feature handling abilities, and optimizing the network's architecture to better reflect the flow of visual information processing. For example, introducing cross-layer connections might help the model better integrate and process information from various visual scenes, thus more precisely simulating complex visual illusions.

Furthermore, considering the unique demands of different visual illusions on visual information processing, model design should include specialized processing modules for specific types of illusions. For instance, designing different network modules for line illusions and color illusions could enable DNNs to more effectively learn and simulate these illusion characteristics. This modular design approach not only improves the adaptability and flexibility of the model but may also enhance its robustness in handling unknown or complex visual information.

Nevertheless, there are limitations to the capabilities of DNNs in simulating visual illusions. Inconsistencies in model performance and dependencies on specific instances suggest that future research should consider a more diverse range of visual illusion types and more complex datasets during model design [86]. Additionally, ongoing neuroscience research and comparisons between DNNs and human visual processing methods are crucial; they will help develop more precise models and enhance the simulation of human visual perception.

Finally, multimodal models that integrate visual and linguistic information exhibit capabilities closer to human cognitive levels. Future research should consider these models to develop more advanced, human-like AI systems. These advancements will propel us forward in understanding and simulating complex visual illusions, while also continuing to inspire ideas for brain-like modeling.

Chapter 8

Conclusion

In this study, we compared and tested the brain-like characteristics of DNNs by integrating six classic visual illusions and proposed a comprehensive interpretive visualization method to elucidate the underlying principles of these illusions. This approach explores potential issues with DNNs as models for human visual learning, similarities to the visual system, and directions for improvement. Based on the proposed comprehensive interpretive visualization method, the study's specific approach is divided into four steps: first, verifying and testing the pre-trained DNNs' performance on visual illusions, followed by comparisons based on training with specific visual illusion datasets. Next, differences based on the models' architectures are examined in detail, and finally, the findings are used to explore potential brain-like characteristics through fMRI experiments.

In Chapter 3, several top-ranking DNNs on Brain-Score were selected to test the Müller-Lyer illusion. The differences among the models in terms of feature attention distribution were significant. Advanced models with excellent performance in visual tasks, such as the Transformer-based VIT and Swin-T, did not exhibit visual illusions. In contrast, classic networks with single architectures like AlexNet and ResNet101 showed the illusion of line length changes. This phenomenon emphasizes the differences in brain-like characteristics of DNNs, where high performance in visual tasks does not equate to brain-like characteristics. Although some advanced DNNs perform well in visual tasks, they may lack the ability to handle certain human visual illusions, whereas some simpler traditional networks may more closely resemble human visual system characteristics in certain aspects. Additionally, DNNs tend to focus more on overall features and cannot understand more complex physical concepts, highlighting the gap between them and human visual processing mechanisms.

In Chapter 4, further training on multiple models with specific datasets showed significant differences among the models. Notably, VGG19 almost did not exhibit any visual illusions during this training. The training with specific visual illusion datasets mainly aimed to develop DNNs' understanding of single physical attributes, followed by related physical attribute visual illusion tests on these trained models, such as the tilt illusion. The results showed that the performance of DNNs in visual illusions is indeed influenced by the training datasets. Among them, ResNet101 performed the best in the tests. Although VGG19's feature attention distribution was similar to ResNet101, it did not exhibit any visual illusions in the tests. Additionally, ResNet101's representational dissimilarity matrices (RDMs) indicated the highest representation similarity in its early modules, suggesting the importance of visual illusion responses in early visual regions (such as the V1 area).

Combining the temporal and static characteristics of the models, Chapter 5 explored the visual illusion performance of four video classification models and one predictive coding model. A new training strategy, teacher-student self-supervised learning, was proposed to fully simulate human-like learning methods to enhance the brain-like characteristics of DNNs. The results showed that the models exhibited visual illusion responses in terms of representational similarity, particularly similar to the distribution shown by previous static models. However, in GradCAM analysis, static models like AlexNet, VGG19, and ResNet101 focused more on the arrows themselves in the feature attention heatmaps, similar to the human visual system, which is strongly influenced by the direction of the arrows when perceiving visual illusions. In contrast, the video models only focused on the combination of arrows and lines as a whole. This significant difference in attention indicates that although video models have advantages in global and spatiotemporal analysis, they may be less precise than static models in capturing key visual cues directly related to visual illusions.

In Chapter 6, fMRI-based experiments further explored the correlation between visual regions and visual illusions, based on the visual illusion data from Chapter 4. The results showed that the response regions were closer to the early regions of the ventral pathway, such as V1/V2, similar to the RDM distributions at different network depths in Chapter 4. This result suggests a potential relationship between DNNs and the ventral pathway in the human visual system, highlighting the importance of shallow modules in brain-like modeling of DNNs.

Totally, our experimental results indicate that the universality of visual illusions in DNNs is not particularly apparent, especially across different architectures and task characteristics, where model performance varies significantly. However, from the RDM and feature attention heatmaps, models with single architectures like VGG19 and ResNet101 showed visual illusion phenomena closer to human perception. Greater depth and complexity in model architecture do not necessarily lead to better visual illusion mechanisms. The mechanisms of visual illusions still need to fully consider the limitations of models, particularly in understanding complex physical attributes. This suggests that future visual illusion research needs to cautiously consider the limitations of DNNs in handling complex cognitive tasks.

Moreover, our findings emphasize the differences in visual illusion performance between pre-trained and self-trained models, requiring comprehensive consideration of these differences' impact on research results. Additionally, from the perspective of changes in model depth, shallow modules typically outperform deeper modules in visual illusion performance. Finally, to enhance the brain-like characteristics of DNNs, future work needs to further set specific visual illusion datasets and design models with specific architectures, particularly focusing on the feature information of shallow modules for brain-like modeling. Through such optimizations, we can better simulate the human visual system, thereby promoting the development and application of artificial intelligence technology.

Bibliography

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

[2] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[4] Nikolaus Kriegeskorte and Pamela K Douglas. Cognitive computational neuroscience. *Nature neuroscience*, 21(9):1148–1160, 2018.

[5] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010.

[6] Shimon Ullman. *High-level vision: Object recognition and visual cognition*. MIT press, 2000.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[11] Demis Hassabis and Eleanor A Maguire. Deconstructing episodic memory with construction. *Trends in cognitive sciences*, 11(7):299–306, 2007.

[12] Richard L Gregory. Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1358): 1121–1127, 1997.

[13] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8827–8836, 2018.

[14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[15] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[17] Grace W Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, 33(10):2017–2031, 2021.

[18] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[19] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[20] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[21] James Outram Robinson. *The psychology of visual illusion*. Courier Corporation, 2013.

[22] Dejan Todorović. What are visual illusions? *Perception*, 49(11):1128–1199, 2020.

[23] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger Von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172, 2012.

[24] ML Seghier and P Vuilleumier. Functional neuroimaging findings on the human perception of illusory contours. *Neuroscience & Biobehavioral Reviews*, 30(5):595–612, 2006.

[25] Kalanit Grill-Spector and Rafael Malach. The human visual cortex. *Annu. Rev. Neurosci.*, 27:649–677, 2004.

[26] Scott O Murray, Daniel Kersten, Bruno A Olshausen, Paul Schrater, and David L Woods. Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences*, 99(23):15164–15169, 2002.

[27] David M Eagleman. Visual illusions and neurobiology. *Nature Reviews Neuroscience*, 2(12):920–926, 2001.

[28] Dale Purves and R Beau Lotto. *Why we see What we do: An Empirical theory of Vision.* Sinauer Associates, 2003.

[29] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1: 417–446, 2015.

[30] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[31] Leslie G Ungerleider and James V Haxby. 'what'and 'where'in the human brain. *Current opinion in neurobiology*, 4(2):157–165, 1994.

[32] A David Milner and Melvyn A Goodale. Two visual systems re-viewed. *Neuropsychologia*, 46(3):774–785, 2008.

[33] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports*, 6(1):27755, 2016.

[34] Umut Güçlü and Marcel AJ Van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[35] Marin Dujmović, Gaurav Malhotra, and Jeffrey S Bowers. What do adversarial images tell us about human vision? *Elife*, 9:e55978, 2020.

[36] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

[37] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

[38] David GT Barrett, Ari S Morcos, and Jakob H Macke. Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current opinion in neurobiology*, 55:55–64, 2019.

[39] Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896, 2016.

[40] E Charles Leek, Ales Leonardis, and Dietmar Heinke. Deep neural networks and image classification in biological vision. *Vision Research*, 197:108058, 2022.

[41] Yaoda Xu and Maryam Vaziri-Pashkam. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature communications*, 12(1):2065, 2021.

[42] Radoslaw M Cichy and Daniel Kaiser. Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4):305–317, 2019.

[43] Simon Farrell and Stephan Lewandowsky. *Computational modeling of cognition and behavior*. Cambridge University Press, 2018.

[44] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

[45] Eiji Watanabe, Akiyoshi Kitaoka, Kiwako Sakamoto, Masaki Yasugi, and Kenta Tanaka. Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in psychology*, 9:340023, 2018.

[46] Alexander Gomez-Villa, Adrian Martín, Javier Vazquez-Corral, Marcelo Bertalmío, and Jesús Malo. Color illusions also deceive cnns for low-level vision tasks: Analysis and implications. *Vision Research*, 176:156–174, 2020.

[47] Hongtao Zhang, Shinichi Yoshida, and Zhen Li. Decoding illusion perception: A comparative analysis of deep neural networks in the müller-lyer illusion. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1898–1903. IEEE, 2023.

[48] Hongtao Zhang and Shinichi Yoshida. Exploring deep neural networks in simulating human vision through five optical illusions. *Applied Sciences*, 14(8):3429, 2024.

[49] Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64(12):3197–3234, 2022.

[50] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

[51] Per-Erik Danielsson. Euclidean distance mapping. *Computer Graphics and image processing*, 14(3):227–248, 1980.

[52] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[53] Hongtao Zhang, Zhen Li, and Shinichi Yoshida. Müller-lyer illusion is replicated by higher layer of pre-trained deep neural network for object recognition. In *The 10th*

International Symposium on Computational Intelligence and Industrial Applications (ISCIIA2022). Beijing Institute of Technology, 2022.

[54] Hongtao Zhang, Shinichi Yoshida, and Zhen Li. Brain-like illusion produced by skye's oblique grating in deep neural networks. *PLOS ONE*, 19(2):1–24, 02 2024.

[55] Barbara Gillam. Illusions at century's end. *Perception and cognition at century's end*, pages 95–136, 1998.

[56] Martin Schrimpf, Jonas Kubilius, Michael J Lee, N Apurva Ratan Murty, Robert Ajemian, and James J DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 2020.

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[59] Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32, 2019.

[60] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[61] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[62] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[63] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[64] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[65] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[66] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[67] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[68] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[69] Martin Engilberge, Edo Collins, and Sabine Süsstrunk. Color representation in deep neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2786–2790. IEEE, 2017.

[70] Akiyoshi Kitaoka. Tilt illusions after oyama (1960): A review 1. *Japanese Psychological Research*, 49(1):7–19, 2007.

[71] Soma Nonaka, Kei Majima, Shuntaro C Aoki, and Yukiyasu Kamitani. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *IScience*, 24(9), 2021.

[72] Yajing Zheng, Shanshan Jia, Zhaofei Yu, Jian K Liu, and Tiejun Huang. Unraveling neural coding of dynamic natural visual scenes via convolutional recurrent neural networks. *Patterns*, 2(10), 2021.

[73] Qiongyi Zhou, Changde Du, and Huiguang He. Exploring the brain-like properties of deep neural networks: A neural encoding perspective. *Machine Intelligence Research*, 19(5):439–455, 2022.

[74] William Lotter, Gabriel Kreiman, and David Cox. A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature machine intelligence*, 2(4):210–219, 2020.

[75] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[76] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021.

[77] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

[78] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[79] Sajjad Abbasi, Mohsen Hajabdollahi, Nader Karimi, and Shadrokh Samavi. Modeling teacher-student techniques in deep neural networks for knowledge distillation. In *2020 International Conference on Machine Vision and Image Processing (MVIP)*, pages 1–6. IEEE, 2020.

[80] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

[81] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[82] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:14, 2014.

[83] PT Jean Talairach. Co-planar stereotaxic atlas of the human brain. *(No Title)*, 1988.

[84] Jack L Lancaster, Marty G Woldorff, Lawrence M Parsons, Mario Liotti, Catarina S Freitas, Lacy Rainey, Peter V Kochunov, Dan Nickerson, Shawn A Mikiten, and Peter T Fox. Automated talairach atlas labels for functional brain mapping. *Human brain mapping*, 10(3):120–131, 2000.

[85] Nadim Nachar. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, 4, 03 2008.

[86] Johannes Mehrer, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1):5725, 2020.

List of Publication

Publication of journal:

- Zhang, Hongtao, Shinichi Yoshida, and Zhen Li. "Brain-like illusion produced by Skye's Oblique Grating in deep neural networks." Plos one 19, no. 2 (2024): e0299083.
- Zhang, Hongtao, and Shinichi Yoshida. "Exploring Deep Neural Networks in Simulating Human Vision through Five Optical Illusions." Applied Sciences 14, no. 8 (2024): 3429.

Publication of conference:

- Zhang, Hongtao, Shinichi Yoshida, and Zhen Li. "Decoding Illusion Perception: A Comparative Analysis of Deep Neural Networks in the Müller-Lyer Illusion." In 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1898-1903. IEEE, 2023.
- Zhang, Hongtao, Zhen Li, and Shinichi Yoshida. "Muller-Lyer Illusion is Replicated by Higher Layer of Pre-trained Deep Neural Network for Object Recognition." In The 10th International Symposium on Computational Intelligence and Industrial Applications (ISCIIA2022), Beijing, 2022.