学　位　論　文　内　容　の　要　旨

博士の専攻分野の名称　　博士（情報科学）　　氏名　甘　耀宗

学　位　論　文　題　名

A Study on Visual Recognition with Large Multimodal Models and Its Applications
（大規模マルチモーダルモデルを用いた視覚認識およびその応用に関する研究）

Deep learning has become a cornerstone of artificial intelligence (AI), driving remarkable progress across different tasks. Since the success of convolutional neural networks (CNNs) in large-scale image recognition, a series of CNN-based architectures, such as ResNet, EfficientNet, have significantly advanced computer vision by learning hierarchical and highly representative visual features. More recently, Transformer-based models have further revolutionized this field by introducing a new paradigm for global context modeling and attention-based representation learning. Meanwhile, in natural language processing (NLP), models such as BERT, the GPT-2 have demonstrated the effectiveness of large-scale pre-training in capturing structured knowledge from data. The convergence of these advances has led AI to evolve from unimodal perception to multimodal understanding, in which information from vision, language, and other modalities is jointly exploited for more comprehensive reasoning and decision-making.

Building on these developments, the emergence of large multimodal models (LMMs) marks a significant paradigm shift in AI research. Unlike traditional unimodal networks, LMMs such as Qwen, Gemini, and GPT are trained on massive datasets that combine heterogeneous modalities, enabling them to learn shared semantic representations across multimodal inputs, including texts and images. These models have shown impressive capabilities in vision–language alignment, cross-modal retrieval, multimodal reasoning, and generative tasks. Despite the remarkable progress achieved by LMMs, several unresolved challenges remain in applying them to visual recognition tasks under real-world conditions. First, the effectiveness of current LMMs still depends on a vast amount of paired multimodal data and extensive computational resources, which limits their scalability to data-scarce or domain-specific applications. Second, the alignment between different modalities is often imperfect, resulting in domain bias, ambiguous reasoning paths, and reduced interpretability when processing complex real-world visual scenes. Third, although pre-trained LMMs exhibit impressive visual recognition ability, their cross-domain generalization and fine-grained discrimination remain limited when facing distribution shifts or unseen visual categories. In summary, although LMMs have substantially advanced the field of visual recognition, key issues concerning data dependency and domain-specific applicability in real-world visual recognition scenes remain a challenge. Addressing these challenges is essential to realize more reliable and scalable multimodal visual understanding systems that can effectively operate in real-world recognition tasks such as intelligent transportation, remote sensing, and other complex visual environments.

The purpose of this thesis is to investigate how LMMs can be effectively adapted and extended for real-world visual recognition tasks. Special attention is given to applications such as traffic sign recog-

nition and remote sensing, with a particular focus on improving generalization, interpretability, and data efficiency under cross-domain conditions. To address the challenges mentioned earlier, this thesis introduces a progressive series of multimodal visual understanding methods. Each method is designed to mitigate one or more of these issues and collectively contribute to enhancing the adaptability and reasoning capability of LMMs. The studies can be summarized as follows. In the first part, we focus on feature-level alignment for zero-shot recognition. We investigate how structural relationships among visual categories can be leveraged to perform cross-domain recognition without labels. To this end, we construct a midlevel feature correspondence space that preserves structural similarity in the feature domain, enabling robust recognition of different categories and improving model generalization under data-scarce conditions. In the second part, we explore few-shot adaptation through in-context learning. Building upon the feature alignment concept, we propose an approach that utilizes multimodal contextual cues to guide LMMs in learning from a limited number of labeled examples. This method achieves effective knowledge transfer and enhances recognition accuracy across heterogeneous visual environments, demonstrating the feasibility of data-efficient multimodal learning. In the third part, we extend visual recognition to the reasoning level. We introduce a multi-step reasoning mechanism that combines visual and semantic information to perform fine-grained and interpretable recognition across domains. Inspired by human chain-of-thought reasoning, our approach enhances transparency and semantic understanding in zero-shot recognition tasks, allowing LMMs to perform stepwise inference under domain shifts. In the final part, we apply the developed methodologies to a large-scale multimodal fusion scenario in remote sensing. We integrate visual imagery with textual and geographic information to construct a multimodal fusion model for regional scene understanding. Through extensive experiments, we verify the scalability and practicality of the proposed multimodal recognition paradigm, demonstrating its potential to extend multimodal reasoning to geospatial and real-world applications.

The structure of the thesis is shown as follows. Chapter 1 describes the research background and purpose of this paper. Chapter 2 reviews the related work on deep learning, LMMs, and visual recognition applications in traffic sign and remote sensing. Chapter 3 investigates visual recognition under zero-shot conditions by analyzing structural relationships among different categories. Chapter 4 expands the investigation to few-shot learning scenarios. Chapter 5 explores multimodal reasoning for fine-grained and cross-domain visual recognition. Chapter 6 extends the proposed methodologies to large-scale multimodal applications in the remote sensing domain. Chapter 7 presents the conclusions of this thesis and the future directions are discussed.

In summary, this thesis systematically explores how LMMs can be effectively adapted for visual recognition through progressive advances in feature alignment, few-shot adaptation and multimodal reasoning. By addressing the key challenges of data dependency, imperfect cross-modal alignment, and limited cross-domain generalization in LMMs, the research aims to bridge the gap between general-purpose multimodal pre-training and domain-specific applications. This thesis not only advances the theoretical understanding of multimodal visual learning but also provides practical frameworks for reliable real-world recognition in intelligent transportation and remote sensing systems.