Thesis Proposal

Hierarchical Learning for Fine-Grained Visual Recognition Using Vision Language Models

Student: [Bahareh Shirkani]
Supervisor: [Mathias Zinnen]
Program: Master's Thesis Proposal
Date: []

1 Motivation and Problem Statement

Historical artworks present unique challenges for computer vision systems due to their stylized, abstract, and context-dependent visual characteristics. Fine-grained recognition tasks in artistic imagery—such as identifying specific object categories, gestures, or symbolic elements—require models that can handle both the localization of relevant entities and their subsequent classification within complex visual contexts. While recent advances in vision–language models (VLMs) have demonstrated remarkable capabilities on natural images, their application to domain-specific artistic content remains underexplored.

This thesis proposes a generalizable hierarchical two-stage framework for fine-grained and context-bound recognition in artworks. The approach addresses key challenges including: (1) the need for accurate localization of diverse super-categories (e.g., persons, flowers, fruits) in complex artistic compositions, (2) the multi-label nature of fine-grained classification tasks, and (3) the integration of scene context to improve classification performance. By evaluating this methodology across multiple datasets, we aim to establish a robust and transferable approach for visual recognition in heritage imagery and beyond.

2 Research Questions

- How effectively can a hierarchical two-stage approach (localization followed by classification) perform fine-grained recognition in artworks compared to one-stage direct detection methods?
- How do vision–language models compare to traditional classification methods for fine-grained recognition in artistic imagery?
- Does incorporating scene context improve classification performance across different recognition tasks and datasets?
- Does fine-tuning VLMs improve performance on multi-label classification compared to zero-shot inference?
- How well does the proposed methodology generalize across different datasets and recognition tasks?

3 Methodology

Stage 1: Localization & Segmentation

Use a segmentation model to localize and segment relevant super-categories in artworks. Depending on the dataset and task, these super-categories may include persons, flowers, fruits, animals, or other semantically meaningful entities. This stage provides accurate bounding boxes and segmentation masks for subsequent fine-grained classification.

Stage 2: Context-Aware Fine-Grained Classification (Main Contribution)

Evaluate vision–language models for multi-label fine-grained recognition. Our approach incorporates contextual information by providing the model with both entity crops and scene-level cues, enabling reasoning about fine-grained categories based on surrounding context. Candidate approaches include:

- Zero-shot classification using pre-trained vision–language models
- Fine-tuned vision–language models using parameter-efficient techniques
- Traditional classification baselines for comparison

Stage 3: Multi-Dataset Evaluation & Comparison

Evaluate the proposed framework across multiple datasets to demonstrate generalizability:

- SensoryArt dataset
- ODOR dataset
- Potentially OpenImages, LVIS, or other benchmarks

Evaluation metrics include multi-label F1, precision, recall, mAP, and qualitative error analysis. We will compare the two-stage hierarchical approach against one-stage direct detection approaches to assess trade-offs in detection accuracy and classification performance.

4 Expected Contribution

- A generalizable hierarchical two-stage framework for fine-grained recognition in artworks, consisting of segmentation-based localization and VLM-based classification.
- A context-aware prompting strategy that leverages scene information for improved fine-grained classification.
- A comprehensive comparison of vision–language models versus traditional classifiers for multi-label recognition in artistic imagery.
- A systematic comparison between two-stage hierarchical and one-stage direct recognition approaches.
- Multi-dataset evaluation demonstrating the generalizability of the proposed methodology across SensoryArt, ODOR, and general-purpose benchmarks.
- Insights into how modern VLMs interpret historical, stylized visual content and recommendations for future applications in computational humanities.

5 Timeline (6 Months)

Phase	Period
Literature review & multi-dataset preparation	Month 1
SAM2 segmentation pipeline setup for multiple super-categories	Month 2
VLM & baseline classifier setup; initial evaluation	Month 3
Fine-tuning, context-aware prompting & multi-label learning	Month 4–5
Multi-dataset evaluation, comparison & ablation studies	Month 5–6
Thesis writing, revisions & presentation	Month 6

6 Expected Outcome

This thesis will establish a generalizable hierarchical framework for fine-grained recognition in artworks, identifying which combination of localization and classification approaches best handles multi-label recognition across diverse artistic imagery. The multi-dataset evaluation will demonstrate the transferability of the methodology, providing recommendations for future applications in computational humanities, digital art history, sensory studies, and AI-assisted heritage analysis. The findings will contribute both to the technical understanding of VLMs on domain-specific content and to practical tools for cultural heritage research.