CV25_PromptViT-Percepta_Percepta

Mirkomil Mirzohidov

October 2025

1. Title & Team

Project Title: Prompt-Tuned ViTs for Explainable Fine-Grained Recognition
Team Name: Percepta
Course: Computer Vision (Fall 2025)
Instructor: Dr. I. Atadjanov & Dr. B. Kiani

Team Members

Name	**Student ID**	**Email**	**Role**
Mirkomil Mirzohidov	221408	221408@centralasian.uz	Model architecture & explainability
Muhammad Saidahmetov	220838	220838@centralasian.uz	Experiments & evaluation metrics
Asilbek Tashpulatov	221443	221443@centralasian.uz	Dataset preparation & report writing

GitHub Repository: github.com/mirkomil06/PromptViT-Percepta

2. Abstract (~180 words)

This project aims to develop an explainable fine-grained image recognition framework using **Prompt-Tuned Vision Transformers (ViTs)**. Fine-grained recognition—distinguishing between visually similar categories such as bird species, car models, or flower types—remains challenging due to subtle inter-class differences and limited labeled data. Conventional fine-tuning of ViTs provides high accuracy but requires substantial compute and offers little interpretability.

Our approach explores **lightweight prompt-tuning**, where learnable prompt tokens adapt pretrained ViTs to specific domains with minimal parameter updates. We further integrate **Prompt-CAM** and attention-based visualization to interpret model focus regions, improving transparency and user trust.

The project will compare baseline fine-tuning and prompt-tuning on three benchmark datasets: **CUB-200-2011**, **Stanford Cars**, and **Oxford Flowers-102**, evaluating both performance (accuracy, F1) and explainability (Pointing-Game score). Expected outcomes include an open-source prototype demonstrating that prompt-tuned ViTs can achieve competitive accuracy with significantly fewer tunable parameters while producing interpretable attention heatmaps.

3. Problem & Motivation

Fine-grained recognition tasks play a crucial role in real-world applications such as biodiversity monitoring, vehicle identification, and precision agriculture. However, these tasks require models to capture **subtle visual distinctions** between highly similar categories, often with **limited labeled data**.

Traditional deep CNNs or fully fine-tuned transformers are computationally expensive and typically act as **black boxes**, offering little insight into decision mechanisms. As explainability becomes increasingly important in computer vision, there is a need for methods that balance **accuracy, efficiency, and interpretability**.

Prompt-tuning provides a promising solution by introducing small learnable vectors—**prompts**—that condition the transformer without retraining all parameters. This significantly reduces resource demands and enhances adaptability to new datasets. Combined with **visual explanation techniques** like Prompt-CAM, such models can deliver both performance and transparency.

The measurable goal of this project is to achieve $\geq 90\%$ **of full fine-tuning accuracy** while enabling **human-interpretable attention visualizations**.

4. Related Work

Research on transformer-based vision models has expanded rapidly. Dosovitskiy et al. (2020) introduced **Vision Transformers (ViT)**, proving that self-attention mechanisms can outperform CNNs when pretrained on large datasets. Jia et al. (2022) proposed **Visual Prompt Tuning (VPT)**, demonstrating that prompts can adapt pretrained transformers efficiently for downstream vision tasks.

Zhou et al. (2016) pioneered **Class Activation Mapping (CAM)** for convolutional models, providing spatial visual explanations. Later, Chefer et al. (2021) extended interpretability to transformers, showing that attention-based explanations can localize decision-relevant regions.

Compared with these baselines, our project combines the **parameter-efficiency of prompt-tuning** with **interpretability from Prompt-CAM**, applied specifically to **fine-grained recognition**, a domain where both precision and transparency are essential.

Table 1: Comparison of Related Approaches

Method	**Architecture**	**Explainability**	**Fine-Tuning Cost**	**Typical Accuracy**
CNN + CAM (Zhou et al., 2016)	CNN	Partial	High	Medium
Full ViT Fine-Tuning (Dosovitskiy et al., 2020)	ViT	Limited	Very High	High
Visual Prompt Tuning (Jia et al., 2022)	ViT + Prompts	None	Low	High
Proposed (Prompt-Tuned + Prompt-CAM)	ViT + Prompts	✓ Full	Low	High

5. Data & Resources

We will use three open-source datasets suitable for fine-grained classification:

- **CUB-200-2011 (Birds)** — 200 species, 11,788 images.
- **Stanford Cars** — 16,185 images, 196 categories.
- **Oxford Flowers-102** — 102 flower species, 8,189 images.

All datasets are publicly available for educational use.

Compute Environment: Google Colab GPU / Kaggle TPU; development via **Visual Studio Code (VS Code)**.

Frameworks: Python 3.10, PyTorch, Hugging Face Transformers, Matplotlib/Seaborn.

Ethical note: no personal or sensitive data are included.

6. Method

6.1 Baseline

We begin with a **pretrained ViT-B/16** model fine-tuned on CUB-200-2011 to establish baseline accuracy and F1 scores.

6.2 Prompt-Tuning

We introduce **learnable prompt tokens** into the transformer's input embeddings while freezing backbone parameters. This reduces training cost by $> 90\%$.

6.3 Explainability Module

Using **Prompt-CAM** (adapted from Chefer et al., 2021), we compute attention-weighted relevance maps that highlight decision regions.

6.4 Ablations

We will vary prompt length, number of tuned layers, and dataset size to measure effects on accuracy and interpretability.

A simplified pipeline:

```
Input Image → ViT Backbone + Prompt Tokens → Classification Head → Prompt-CAM Heatmap
```

7. Experiments & Metrics

Experiments will evaluate:

Category	**Metric**	**Goal**
Classification	Accuracy, F1-score	$\geq 90\%$ of full fine-tuning baseline
Explainability	Pointing-Game Score, Localization Accuracy	$\geq 70\%$ correct focus regions
Efficiency	Tunable Parameter Ratio	$< 10\%$ of full fine-tuning
Visualization	Qualitative Prompt-CAM maps	Human-interpretable heatmaps

Each model will be trained on 70/20/10 splits and validated on unseen classes to test generalization.

8. Risks & Mitigations

Risk	**Impact**	**Mitigation**
Limited GPU compute	Medium	Use smaller ViT-Tiny models and lower batch sizes.
Overfitting on small datasets	High	Apply data augmentation and early stopping.
Poor interpretability	Medium	Experiment with Prompt-CAM parameter tuning.
Implementation bugs	Medium	Unit test modules and use pretrained checkpoints.

9. Timeline & Roles

Week	Key Milestone	Owner
1	Team setup + repo initialization	All
2	Literature review + dataset access	Asilbek
3	Baseline ViT implementation	Mirkomil
4	Prompt-tuning development	Muhammad
5	Explainability integration	Mirkomil
6	Evaluation + comparison	All
7	Proposal + slides writing	Asilbek
8	Final presentation	All

ROADMAP.md: link

10. Expected Outcomes

- Trained **Prompt-Tuned ViT** achieving near-baseline accuracy with reduced parameters.
- **Explainability visualizations** (Prompt-CAM heatmaps).
- Comparative analysis vs. full fine-tuning.
- **Deliverables:** Code repository, report, trained weights, and presentation slides.

11. Ethics & Compliance

All datasets are open-access and used under research licenses. No private, medical, or biometric data will be used. Bias analysis will be noted if class imbalance affects results. Project outputs are open-sourced for academic reproducibility.

12. References

1. Chefer H., Gur S., Wolf L. *Transformer Interpretability Beyond Attention Visualization.* Proc. IEEE/CVF CVPR, 2021, pp. 782–791.
   [PDF Link]
2. Dosovitskiy A. *An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale.* arXiv:2010.11929, 2020.
   [PDF Link]
3. Jia M. et al. *Visual Prompt Tuning.* ECCV 2022, pp. 709–727.
   [PDF Link]
4. Zhou B. et al. *Learning Deep Features for Discriminative Localization.* Proc. IEEE CVPR, 2016, pp. 2921–2929.
   [PDF Link]