

Sonu Dileep

PhD Student in the Department of Computer Science at Colorado State University

Title: Fine-Grained Deep Visual Understanding

Summary (or project abstract): Despite the rapid progress of AI in visual tasks, current computer vision systems still struggle with fine-grained, pixel-level understanding across varied visual domains. Many state-of-the-art models achieve high performance on benchmark datasets, but fall short when applied to complex, real-world visual scenes—where objects may be small, partially occluded, amorphous, or captured under varied lighting and environmental conditions.

My central research goal is to push the limits of computer vision systems towards achieving fine-grained, human-like visual understanding at the pixel level. The emphasis is on building lightweight, data-efficient models that can handle high visual ambiguity, operate across spatial and temporal dimensions, and perform reliably in operational settings.

To explore and progress towards this goal, I've worked on a set of projects that challenge the current limitations of computer vision systems in different domains. These are not end goals in themselves, but milestones along the path toward achieving more general, precise, and adaptable vision systems.

Research Progress:

1. Understanding Amorphous Structures in Video: Smoke Opacity Estimation

One aspect of fine-grained vision is detecting non-rigid, dynamic structures such as smoke. This problem is especially challenging due to smoke's lack of clear shape, rapidly changing form, and similarity to other background elements like clouds.

To study this challenge, I developed a compact deep learning model to estimate Ringelmann smoke opacity levels from surveillance video. The goal was to approximate the precision of certified human observers using a repeatable, automated method. The model uses a two-stage architecture with segmentation pretraining and classification refinement and incorporates Adaptive Fourier Neural Operators (AFNO) to improve computational efficiency while capturing spatial context. It achieves 95.27% accuracy on an unseen test set.

This work reflects the broader challenge of understanding subtle motion-based cues in noisy video data and is part of my ongoing effort to build models that can reason about soft, ambiguous object classes.

2. Detecting Subtle Structure in Satellite Imagery: Infrastructure Segmentation

Another challenge in fine-grained understanding is detecting small, low-contrast, or partially occluded man-made structures in low-resolution visual data like satellite imagery. To explore this, I developed a system to detect oil and gas infrastructure from satellite images. This model employs a dual-branch Transformer-inspired architecture that detects individual equipment (e.g., tanks, separators) and outlines facility boundaries.

Using high-resolution Maxar satellite images, I curated a dataset and trained this lightweight model, which achieved 93% accuracy on both facility and equipment classification. The model also integrates a second-stage architecture to handle large-scale spatial contexts, using sliding windows and segmentation fusion.

This project pushes on the question: How can we achieve localized, pixel-accurate recognition in visually cluttered, low-signal environments?



Figure 1: Facility Segmentation Results in DJ Basin Colorado

Why This Matters?

These efforts are driven by the broader need for **trustworthy AI vision systems** that operate reliably under constraints—whether it's limited data, ambiguous visual input, or high operational stakes. Building models that can generalize with less supervision, interpret complex scenes, and operate efficiently is key to unlocking real-world computer vision across industries.

By addressing challenges in motion understanding, ambiguous textures, small object segmentation, and spatial context integration, my research contributes towards building more resilient, flexible, and accurate computer vision systems.

I am still actively working toward this goal—refining architectures, expanding datasets, and developing training strategies that can bring us closer to truly fine-grained, pixel-level visual understanding.

Publications

Sonu Dileep, *Automated Recognition of Oil and Gas Production Infrastructure using Satellite Imagery*, International Journal of Applied Earth Observation and Geoinformation (2025) (Under Review)