# DETRmining the Cosmos: A Transformer-Based Approach to Galaxy Morphology Detection

*Kumar Chandra*, Max Rodriguez*, Renn Su**

*CS 231N Spring 2025, Department of Computer Science, Stanford University*

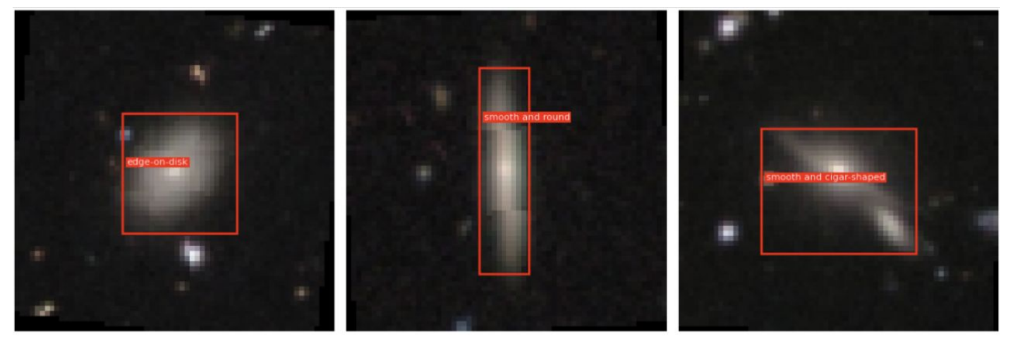
Stanford
Computer Science

## Introduction

We present an end-to-end method for **classifying galaxy morphology** in wide-field astronomical surveys. Our approach uses a multistep detection transformer (DETR) architecture and integrates a convolutional or self-supervised DINO-based backbone. We show that **self-supervised DINO features accelerate convergence and improve classification**, while DETR's end-to-end formulation obviates heuristic region proposals.

### Dataset - Galaxy Zoo 2

- Crowd-sourced labeled datasets of galaxy morphology
- Nearly 300, 000 galaxies from the Sloan Digital Sky Survey
- We use the DemoRings, MNIST, and GZ2 datasets.



*Example images from the GalaxyMNIST dataset with bounding boxes and ground truth labels.*

## Problem Statement

Modern sky surveys produce millions of galaxy images, but classifying and labeling telescopic imagery remains slow and unreliable. Existing methods depend on slow manual labeling or unreliable algorithms that struggle with noise, ambiguity, and scale.

## Experimentation

### Experimentation Overview

Our experiments were designed to rigorously compare CNN and self-supervised DINO features, with the goal of advancing scalable, automated analysis of large-scale astronomical surveys.

### Study Design and Factors

- **Data Preprocessing and Scale:** tested on three datasets with 1k, 10k, 210k images
- **Feature Encoding:** ablation studies between CNN and DINO-based approaches.
- **Transformer effectiveness:** compared transformer-based architectures with CNN baseline
- **Hyperparameter search:** used varying tuning approaches; analyzed use of scheduler.

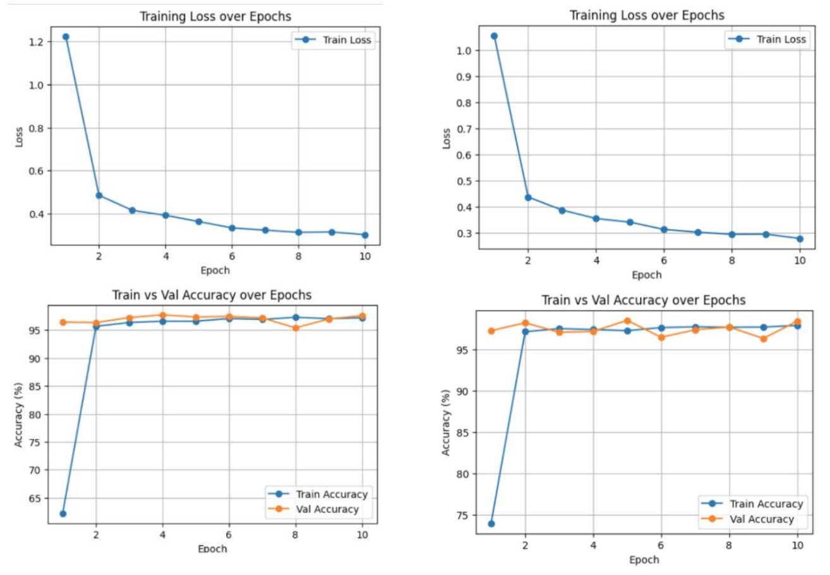


*Figure (Left) Training/validation loss curves without scheduler. (Right) Training/validation loss curves with ReduceLROnPlateau scheduler.*

## Methods

### Architecture Overview

Our architecture takes an **input of a single-frame, RGB-encoded galaxy image.** We then use a DETR-based model (a CNN or self-supervised DINO backbone followed by a transformer decoder) to **output a predicted morphological class and a normalized bounding box** around the primary galaxy.

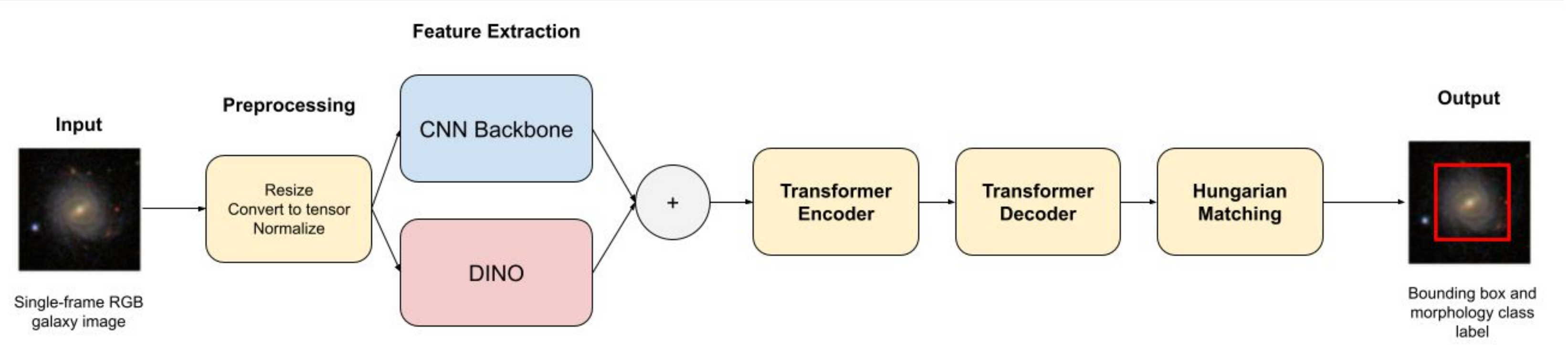


***Figure** (above). Architecture of our end-to-end DETR-based galaxy classification pipeline.*

### Feature Encoding with CNN and DINO

Our experimental architecture uses a backbone to encodes **spatial and semantic patterns** necessary for accurate classification and localization. We investigate two approaches:

1) **CNN Backbone:** A compact, shallow convolutional network that extracts semantically rich features while preserving spatial structure; effective for small, clean datasets.
2) **DINO Backbone:** A self-supervised Vision Transformer pretrained on ImageNet; captures fine-grained textures and global context, improving convergence and robustness on noisy astronomical data.

### Transformers for Galaxy Queries

We flatten the feature map $x \in \mathbb{R}^{B \times c \times H' \times W'}$ produced by either a CNN or DINO backbone. We reshape x into a sequence of tokens: $X \in \mathbb{R}^{B \times s \times D}$, where $S = H' \times W'$, and $D = C$. Our transformer consists of:

- A three-layer encoder with multi-head self-attention and a feedforward network, producing globally contextualized features.
- A three-layer decoder, with self-attention over N object queries and cross-attention to encoder outputs, enabling spatial-semantic reasoning across S positions.

### Hungarian Matching Algorithm

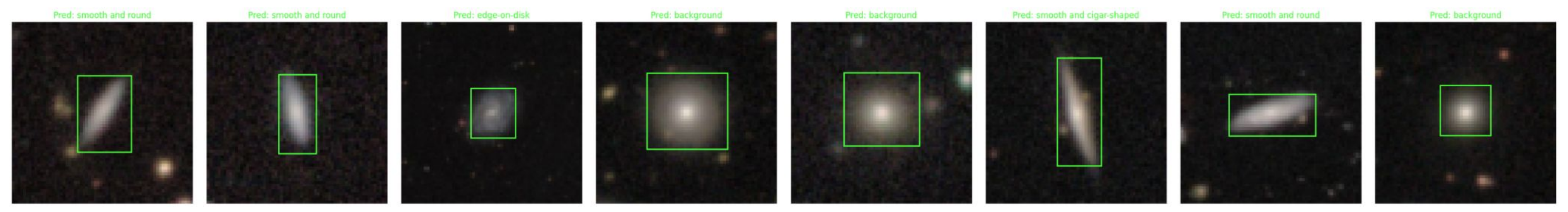


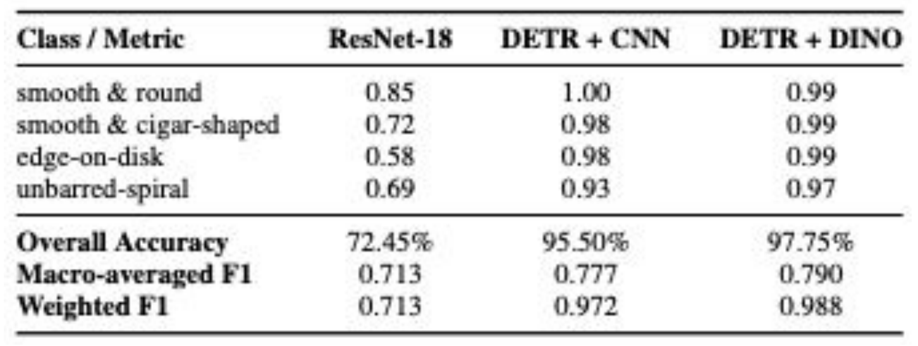
***Figure** (above). Bounding boxes learned for gz2 holdout data.*

- Assigns predicted queries to ground truth objects by minimizing a cost combining classification error, L1 box distance, and GIoU
- Enables end-to-end supervision for both localization and classification in DETR.

## Evaluation

### Quantitative Analysis

**DETR + DINO** outperforms baselines, achieving up to **97.75%** accuracy on MNIST and **95.36%** on Galaxy Zoo 2 holdout data.

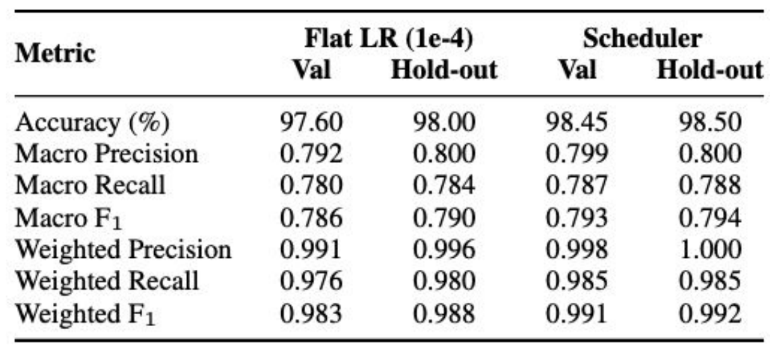


| Class / Metric | ResNet-18 | DETR + CNN | DETR + DINO |
|---|---|---|---|
| smooth & round | 0.85 | 1.00 | 0.99 |
| smooth & cigar-shaped | 0.72 | 0.98 | 0.99 |
| edge-on-disk | 0.58 | 0.98 | 0.99 |
| unbarred-spiral | 0.69 | 0.93 | 0.97 |
| **Overall Accuracy** | 72.45% | 95.50% | 97.75% |
| **Macro-averaged F1** | 0.713 | 0.777 | 0.790 |
| **Weighted F1** | 0.713 | 0.972 | 0.988 |

***Figure** ( Above). Comparison of per -class F1 -scores and overall metrics across three models on the MNIST dataset.*

ResNet-18 **baseline** achieves only **72.45%,** on MNIST, highlighting the strength of transformer-based detection.

Furthermore, **DINO backbone improves generalization,** with a consistent **2–3% accuracy boost over CNN backbones** on all datasets. We also observed performance improvement through using a scheduler (see below).



| Metric | Flat LR (1e-4) | | Scheduler | |
|---|---|---|---|---|
| | Val | Hold-out | Val | Hold-out |
| Accuracy (%) | 97.60 | 98.00 | 98.45 | 98.50 |
| Macro Precision | 0.792 | 0.800 | 0.799 | 0.800 |
| Macro Recall | 0.780 | 0.784 | 0.787 | 0.788 |
| Macro $F_1$ | 0.786 | 0.790 | 0.793 | 0.794 |
| Weighted Precision | 0.991 | 0.996 | 0.998 | 1.000 |
| Weighted Recall | 0.976 | 0.980 | 0.985 | 0.985 |
| Weighted $F_1$ | 0.983 | 0.988 | 0.991 | 0.992 |

***Figure** (Above). DETR with DINO performance comparison on MNIST with scheduler*

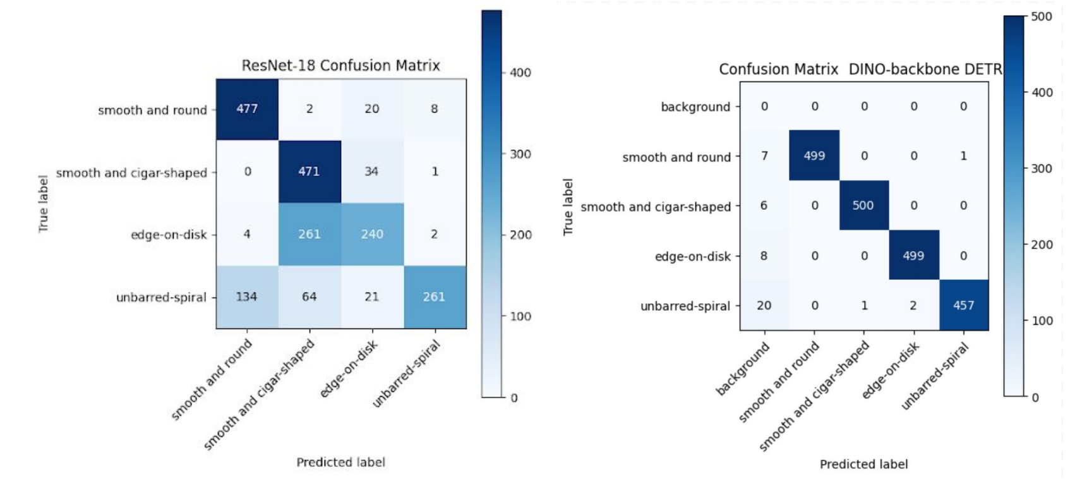
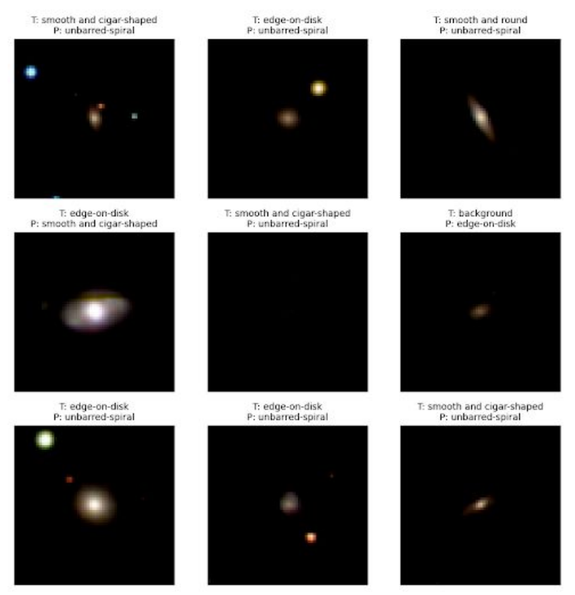


***Figure** ( Left) ResNet-18 validation confusion matrix. (Right) DETR + CNN validation confusion matrix.*

### Qualitative Analysis



We analyzed qualitative trends in misclassifications and found:

- Failure cases often align with class boundary ambiguity. (e.g. distinguishing edge-on disks from faint spirals)
- Human labeling inconsistencies or lack of consensus in Galaxy Zoo votes.
- Low signal-to-noise images or occlusions that experienced human annotators would find ambiguous.

***Figure**. Misclassified MNIST images from the DINO pipeline with predicted class and true class.*

## Discussion & Future Research

### Discussion

- Transformer-based models demonstrate strong potential for galaxy morphology classification, particularly when leveraging self-supervised features to capture fine-grained structural patterns.

### Future Research

- Investigate the applicability of DETR galaxy detection methods to other forms of imagery such as spectroscopic surveys, x-ray surveys, etc.
- Develop a pre-filtering pipeline to automatically discard empty or corrupted images pre-training.