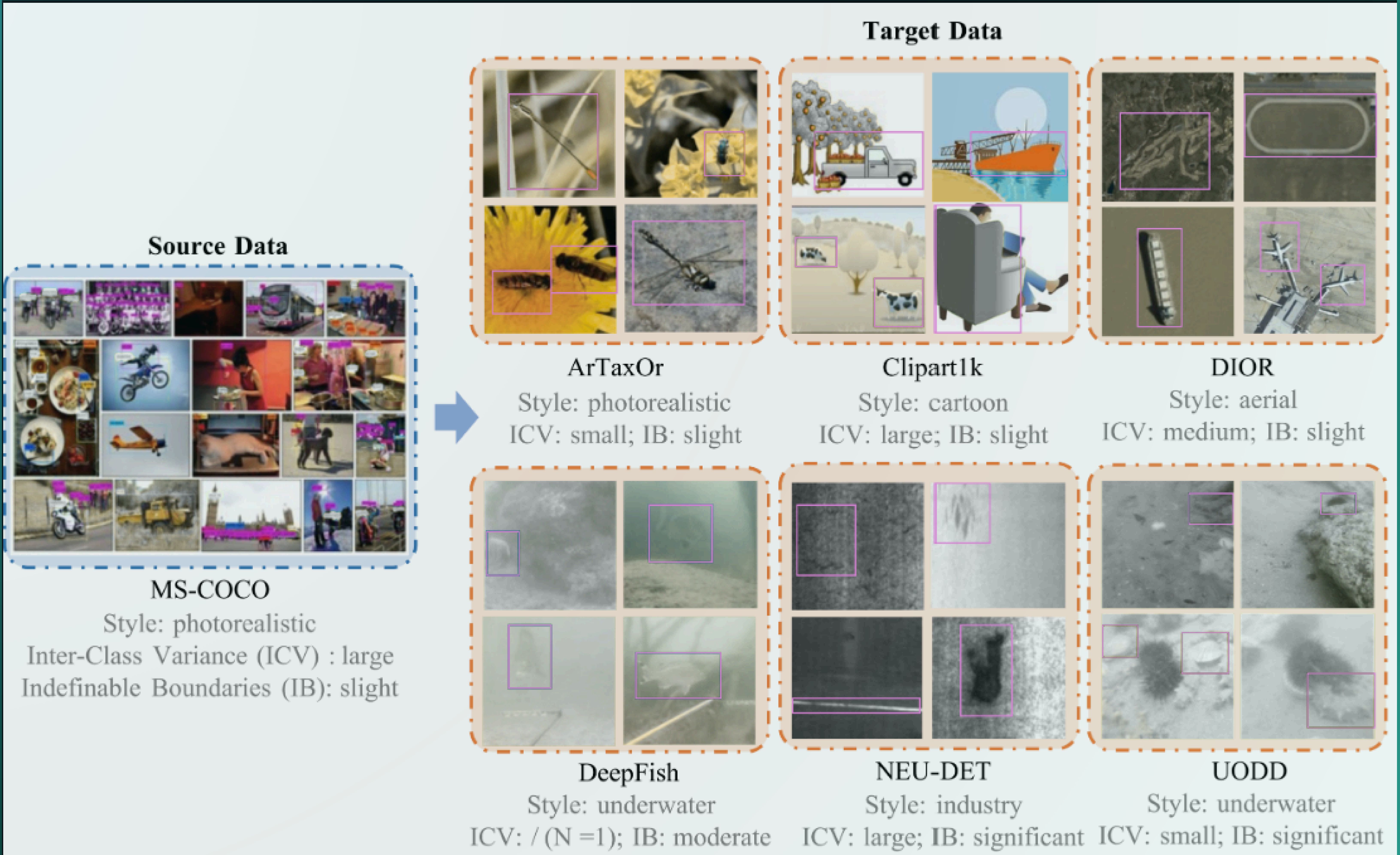
NTIRE2025 Challenge on Cross-Domain Few-Shot Object Detection : Methods & Results

This paper advances research on detecting novel objects in new domains with minimal labeled data, pushing models to generalize effectively across significant domain shifts using diverse datasets and evaluation protocols.



Source Data

MS-COCO

Style: photorealistic

Inter-Class Variance (ICV) : large

Indefinable Boundaries (IB): slight

Target Data

ArTaxOr

Style: photorealistic

ICV: small; IB: slight

Clipart1k

Style: cartoon

ICV: large; IB: slight

DIOR

Style: aerial

ICV: medium; IB: slight

DeepFish

Style: underwater

ICV: / (N =1); IB: moderate

NEU-DET

Style: industry

ICV: large; IB: significant

UODD

Style: underwater

ICV: small; IB: significant

Background

The Cross-Domain Challenge

Most existing FSOD methods assume that the training (source) and testing (target) data come from the same domain (e.g., both are natural images). However, in real-world scenarios, this is rarely the case. For example, a detector trained on natural images (like MS-COCO) may perform poorly when applied to very different domains such as remote sensing or medical imagery.

Why is Cross-Domain FSOD Hard?

- **Domain Shift**: Differences in style, image characteristics, and object boundaries between domains make generalization difficult.
- **Limited Data**: Only a few labeled examples are available for each new class in the target domain.
- **Real-World Relevance:** Addressing domain shift is essential for deploying robust object detectors in varied, unseen environments

Problem Statement & Challenge

Challenge Objective

The NTIRE 2025 CD-FSOD Challenge was launched to push the boundaries of object detection under domain shifts and limited labeled data. The goal is to develop models that can generalize to entirely new domains with only a few labeled examples per class.

Two Tracks: Closed-Source and Open-Source

- **Closed-Source CD-FSOD:**
  - Training is limited to a single source dataset (MS-COCO).
  - The classes in the source and target domains are completely disjoint.
  - Participants must train on the source and adapt to novel target domains with only a few labeled samples per class.
- **Open-Source CD-FSOD:**
  - Participants can use any datasets and large pre-trained models.
  - This track explores the upper bound of model generalization, leveraging foundation models and diverse data.

Challenge Setup & Datasets

The challenge uses an N-way K-shot evaluation protocol, where for each novel class in the target domain, only K labeled examples are provided to assist detection. During development, six diverse target datasets with different styles and object characteristics were used for validation. For the final testing phase, three new unseen datasets-DeepFruits, Carpk, and CarDD-were introduced to evaluate model generalization.

The closed-source track restricts training to the MS-COCO dataset only, while the open-source track allows the use of any source data or pretrained foundation models. The primary evaluation metric is Mean Average Precision (mAP), with a weighted emphasis on 1-shot performance to encourage robustness in extremely low-data scenarios.

Main Open-Source Track Methods

1. MoveFree
2. AI4EarthLab
3. IDCFS

Team 1: MoveFree

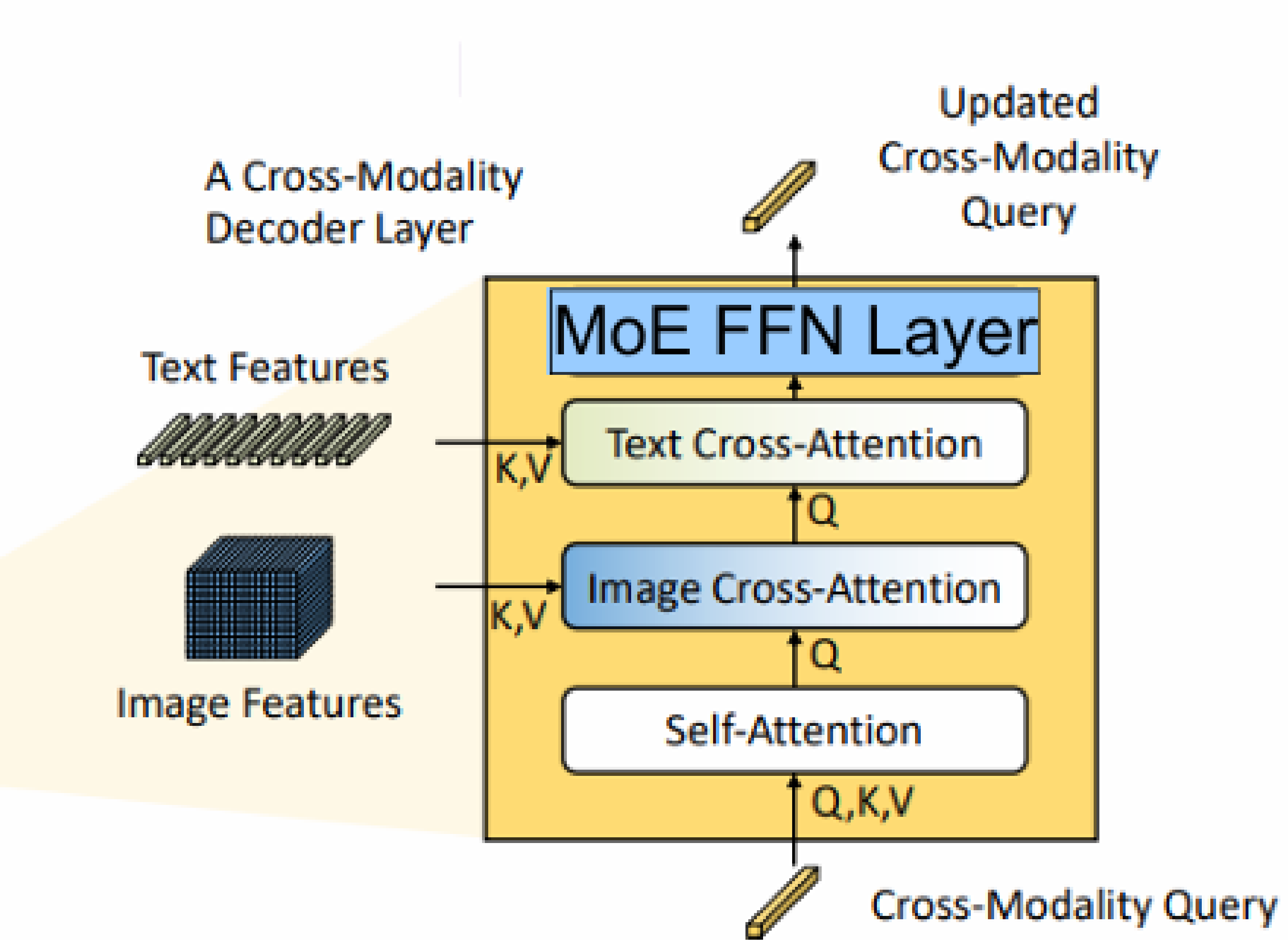
- Open set object detector used (Grounding DINO)
- Zero Shot detection possible due to open set object detectors
- “Shot” detection (zero-shot, one-shot, few-shot, many-shot)
- zero-shot, one-shot, few-shot powerful
- Problem 1 : Missing labels can degrade model performance
- Solution 1 : Self-training (with each cycle, model gets better at making accurate predictions; more reliable and complete labels)
- Problem 2 : Very little data to learn from, and learning and testing environments are different (cross-domain)
- Solution 2 : MoE (Mixture of Experts) (System picks the best expert for each task, depending on what the input looks like)

Training Details

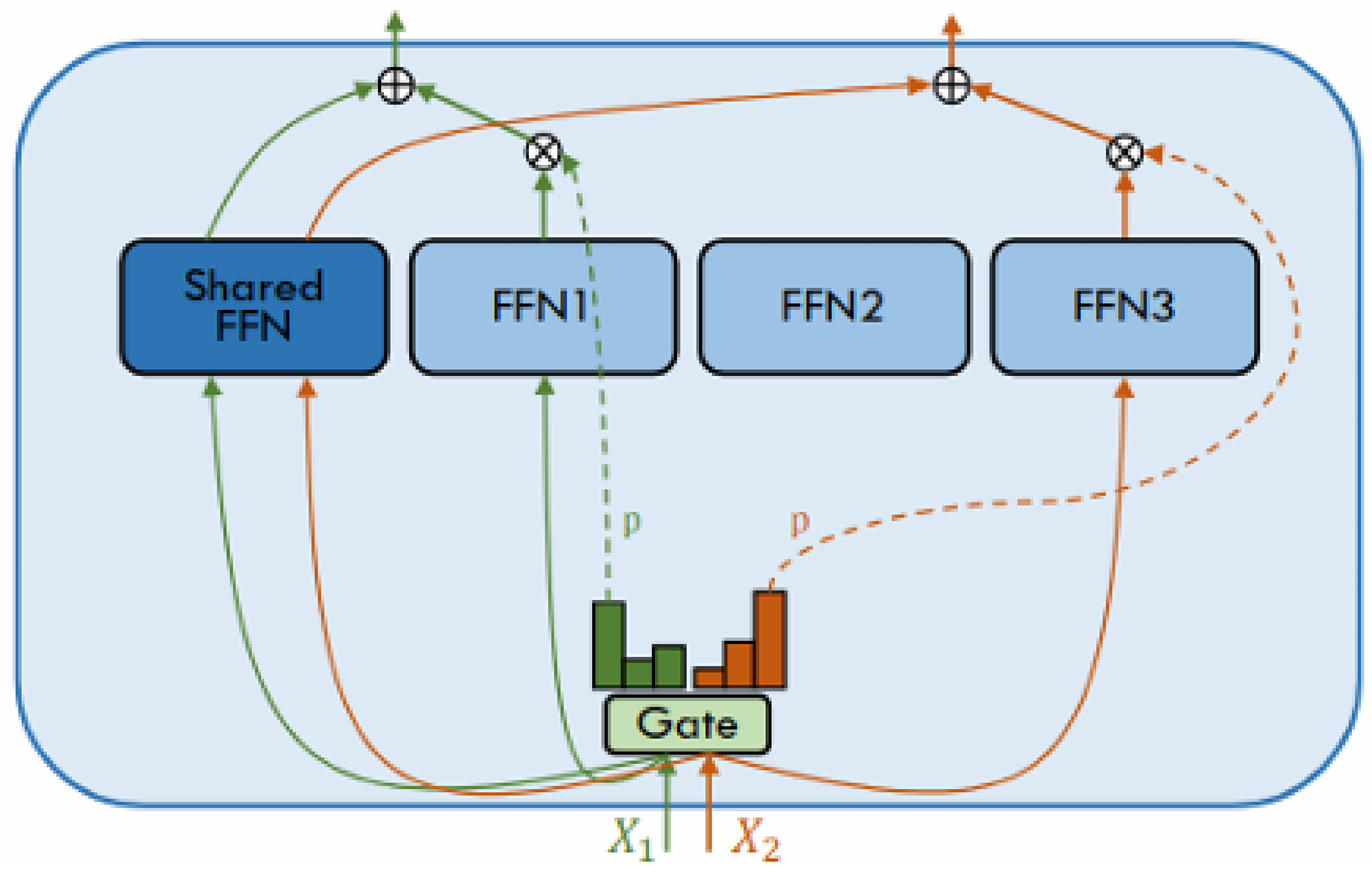
Two-Stage Fine-Tuning

- **Stage 1**: Fine-tune the pre-trained model with all parameters trainable, except for the language encoder (kept frozen because it was already working well so there was no need to retrain it).
- **Stage 2**: Introduce MoE architecture in decoder layer, fine-tuning only the MoE components and detection head. All other parts of the model that were fine tuned in stage 1 are now frozen.

1. MoE Cross-Modality Decoder Layer



2. MoE FFN Layer



Team 2: AI4EarthLab

- Leveraged foundation model Grounding DINO (Swin-B) pretrained on diverse large-scale datasets.
- Built a composite augmentation pipeline:
    1. CachedMosaic
    2. YOLOXHSVRAug
    3. RandomFlip
    4. CachedMixUp
    5. RandomSize
    6. CachedCrop
- Used coarse grain validation sets for hyperparameter tuning.
- Achieved improved robustness and adaptability in CD-FSOD.

Training Details

- Conducted experiments on NVIDIA A100 GPUs, with 8 × 50 experiment groups per round.
- Selected optimal step sizes based on historical performance.
- Applied milestone-based learning rate scheduling (1, 5, 9 epochs).
- Used 900 detection queries and a maximum text token length of 256.
- Employed a BERT-based text encoder with BPE tokenization.
- Network composed of 6-layer feature enhancer and 6-layer cross-modality decoder.
- Loss function combined classification loss. L1 box, and GIoU losses.
- Used Hungarian matching weights: 2.0 (classification), 5.0 (L1), 2.0 (GIoU); final loss weights: 1.0, 5.0, 2.0.

Team 3: IDCFS

- Pseudo-Label Driven Vision-Language Grounding
- Combines GLIP (Grounded Language-Image Pretraining) and Grounding DINO for Cross-Domain Few-Shot Object Detection (CD-FSOD).
- Iterative Pseudo-Labeling:
    - Generates high-confidence pseudo-labels during fine-tuning.
    - Iteratively refines model understanding with these pseudo-labeled samples.
- Model Ensemble with Confidence-Reweighted NMS:
    - Combines GLIP and Grounding DINO outputs.
    - Confidence scores are reweighted, and Non-Maximum Suppression (NMS) is applied to filter overlaps.

Training Details

1. GLIP Fine-Tuning:
   - Full model fine-tuned with a learning rate of 2e-5 for better cross-domain detection.
2. Iterative Pseudo-Labeling Cycle:
   - Predict → Generate Pseudo-Labels → Retrain → Repeat.
   - Enhances model understanding with minimal labeled data.
3. LoRA-based DINO Fine-Tuning:
   - Efficiently adapted for new domains using low-memory updates.
4. Model Ensemble:
   - Combines GLIP + DINO predictions with confidence-weighted NMS for higher accuracy.

Specialized Close-Source Track Method

1. X-Few

Team 1: X-Few

- Introduced Instance Feature Caching (IFC) to store support features.
- Used feature matching for better classification and localization.
- Applied contrastive alignment to reduce domain gap.
- Built on CD-ViTO with lightweight, adaptive modules.
- Enhanced generalization in few-shot, cross-domain settings.

Training Details

- Trained on MS-COCO (closed-source), using 1, 5, 10-shot settings.
- Used RTX A800 GPU, batch size 16, lr: 1e-3 / 1e-4.
- Fine-tuned for 40–200 epochs based on dataset.
- Best mAP scores: 50.98 (D1), 28.00 (D2), 33.00 (D3).
- Final Score: 125.90 – Ranked 1st in closed-source track.

Our Implementation Approach

Proposed Method

Dataset Preparation

- Cloned ETS GitHub repository
- Downloaded COCO 2017 annotations & validation images
- Extracted files and organized dataset directories

Few-Shot Subset Creation

- Created a 5-shot dataset using pycocotools
- Selected 5 annotated images per category
- Ensured unique images with corresponding annotations

Data Visualization

- Loaded few-shot annotations using COCO API
- Displayed random sample images with bounding boxes
- Verified annotations with category names

Model Setup

- Installed transformers & timm
- Loaded pre-trained DETR (facebook/detr-resnet-50)
- Used DetrImageProcessor for pre/post-processing

Inference & Detection

- Selected an image from val2017
- Preprocessed & passed it through DETR model
- Extracted predictions: labels, scores, bounding boxes

Result Visualization

- Drew detection boxes on the image using matplotlib
- Displayed class names and confidence scores
- Confirmed object detection accuracy visually

Training Details

- Model Used: facebook/detr-resnet-50
- Pre-training Dataset: COCO
- The model was used in evaluation/inference mode without any additional training or fine-tuning.
- DetrImageProcessor from HuggingFace was used to handle image pre-processing and post-processing.
- No fine-tuning or training of the model was done in this project. The DETR model used was pre-trained.

Group Members

1. Nigarish Rehman Sarmad 22K-8723
2. Unaiza Ahmed Khan 22K-4121
3. Khoula Adil 22K-8733
4. Farheen Fatima 22K-4045
5. Uroosha Zehra Abidi 22K-4048

THANK YOU