CMC UNIVERSITY
Aspire to Inspire the Digital World

ICAI Impacts & Potentials

Sponsors:

TRAINING-FREE MULTI-MODAL ALIGNMENT FOR FINE-GRAINED COUNTERFEIT FRUIT DETECTION

Quynh Nguyen Huu[1], Dat Tran-Anh[2], Ngoc Anh Nguyen Thi[1], Thieu Huy Nguyen[2], Nguyen Huu Gia Bach[3]

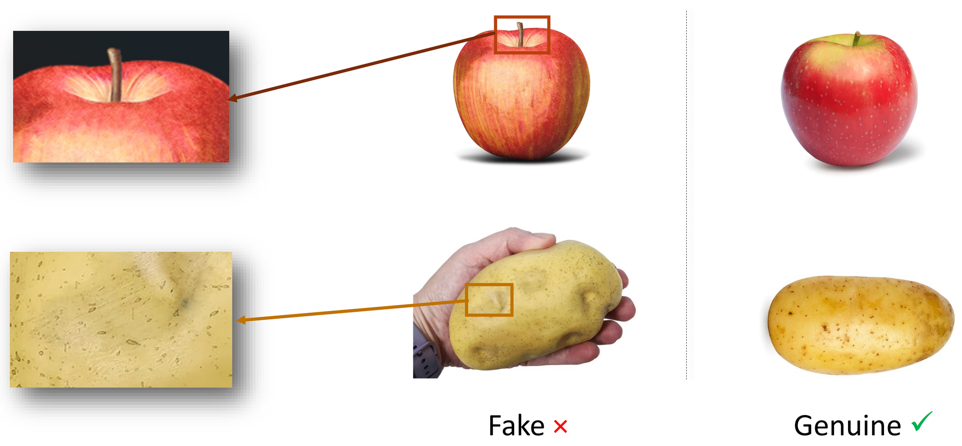
[1] CMC University, Hanoi, Vietnam, [2] Faculty of Information Technology, Thuyloi University, Hanoi, Vietnam Email: dat.trananh@tlu.edu.vn

[3]Yen Hoa High School, Hanoi, Vietnam

PROBLEM DEFINITION AND CONTRIBUTION

Goal:

Detect fine-grained counterfeit fruits in a training-free and incrementally expandable manner.



Fine-grained differences between authentic and fake fruits (texture, gloss, micro-patterns).

Contributions: AMAF-Net (Adaptive Multi-modal Alignment Framework):

- A training-free framework for fine-grained counterfeit fruit detection.
- Enriched semantic representations to capture subtle visual differences.
- Stabilized few-shot visual features using base knowledge.

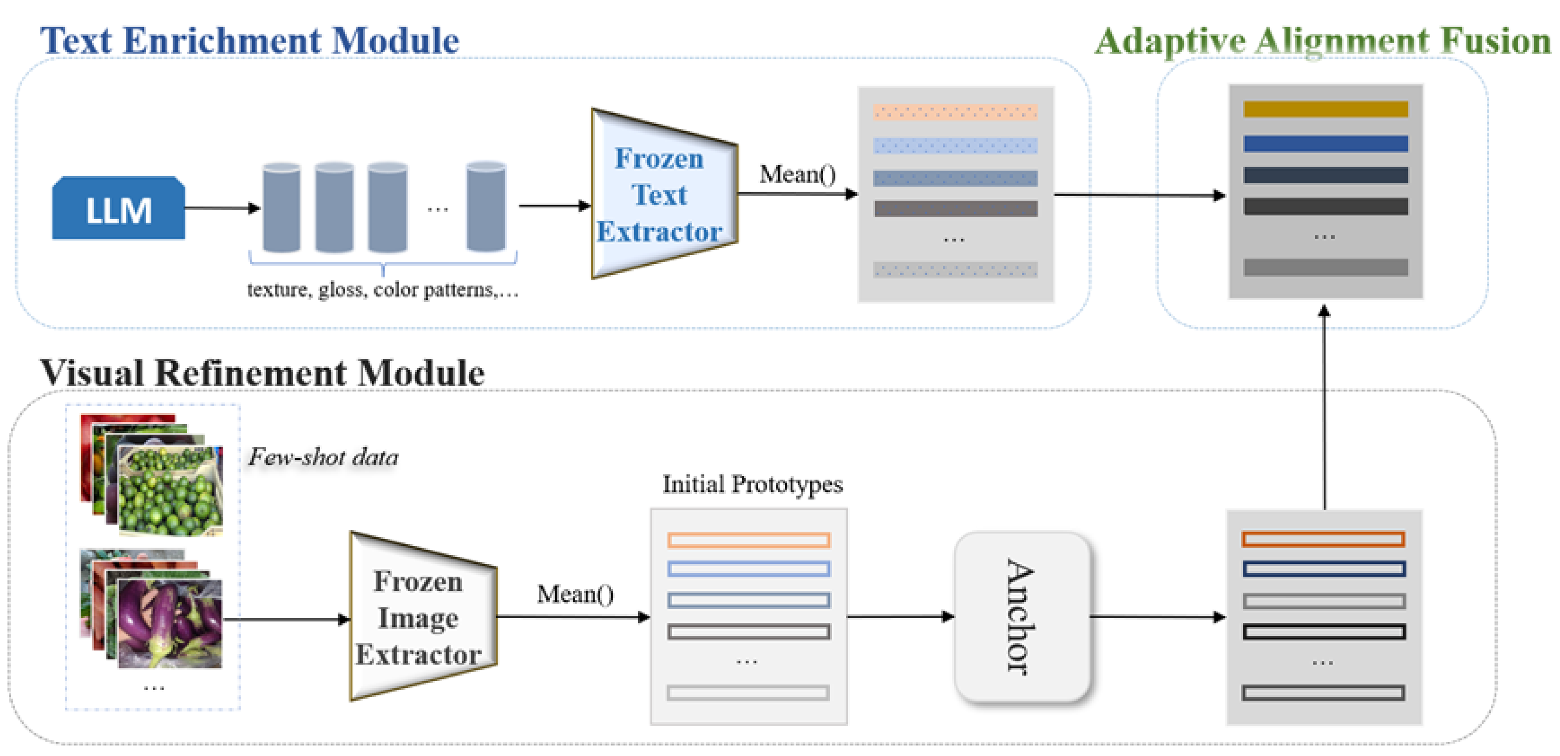
METHODOLOGY

Assumption: Access to a frozen Vision–Language Models and a small number of support images per new class. No fine-tuning or retraining is allowed in any session.

Process:

Encode → Refine → Fuse → Predict

Model Architecture: Images and enriched texts are encoded with frozen backbones, refined using base knowledge, and fused adaptively to produce stable, training-free fine-grained predictions.



EXPERIMENTATION

Dataset: Evaluate on four FSCIL benchmarks: MiniImageNet, CUB-200 (fine-grained bird species), EuroSAT (remote sensing), and FG-Fruit (authentic vs. counterfeit fruits). This setup covers both generic and highly fine-grained scenarios.

MiniImageNet. 100 classes — *CUB-200. 200 classes* — *FG-Fruit. 50 classes*

Visual Backbones:

- ResNet-50/101.
- ViT-B/16, ViT-B/32, ViT-L/14.

LLMs for Semantic Prompts:

- Attribute-based descriptions.
- Capture subtle cues (texture, gloss, micro-patterns).

QUANTITATIVE RESULTS

Method	MiniImageNet		CUB-200		EuroSAT		FG-Fruit	
	Acc_{avg} (%)	Drop (%)	Acc_{avg} (%)	Drop (%)	Acc_{avg} (%)	Drop (%)	Acc_{avg} (%)	Drop (%)
CLIP-Zero	84.4	**-0.24**	87.0	-0.30	86.7	**-0.41**	87.0	**-0.29**
CoOp	86.8	-0.11	89.3	-0.13	88.6	0.14	89.3	0.22
CoCoOP	87.4	0.021	89.5	**-0.41**	88.8	0.21	88.9	-0.16
FSIL-VL	86.1	-0.15	89.0	0.12	89.1	-0.18	88.4	0.26
Ours	**87.9**	-0.11	**90.6**	0.20	**89.6**	-0.12	**90.2**	0.13

Performance comparison across four FSCIL datasets in the 5-shot setting

Overall Performance:

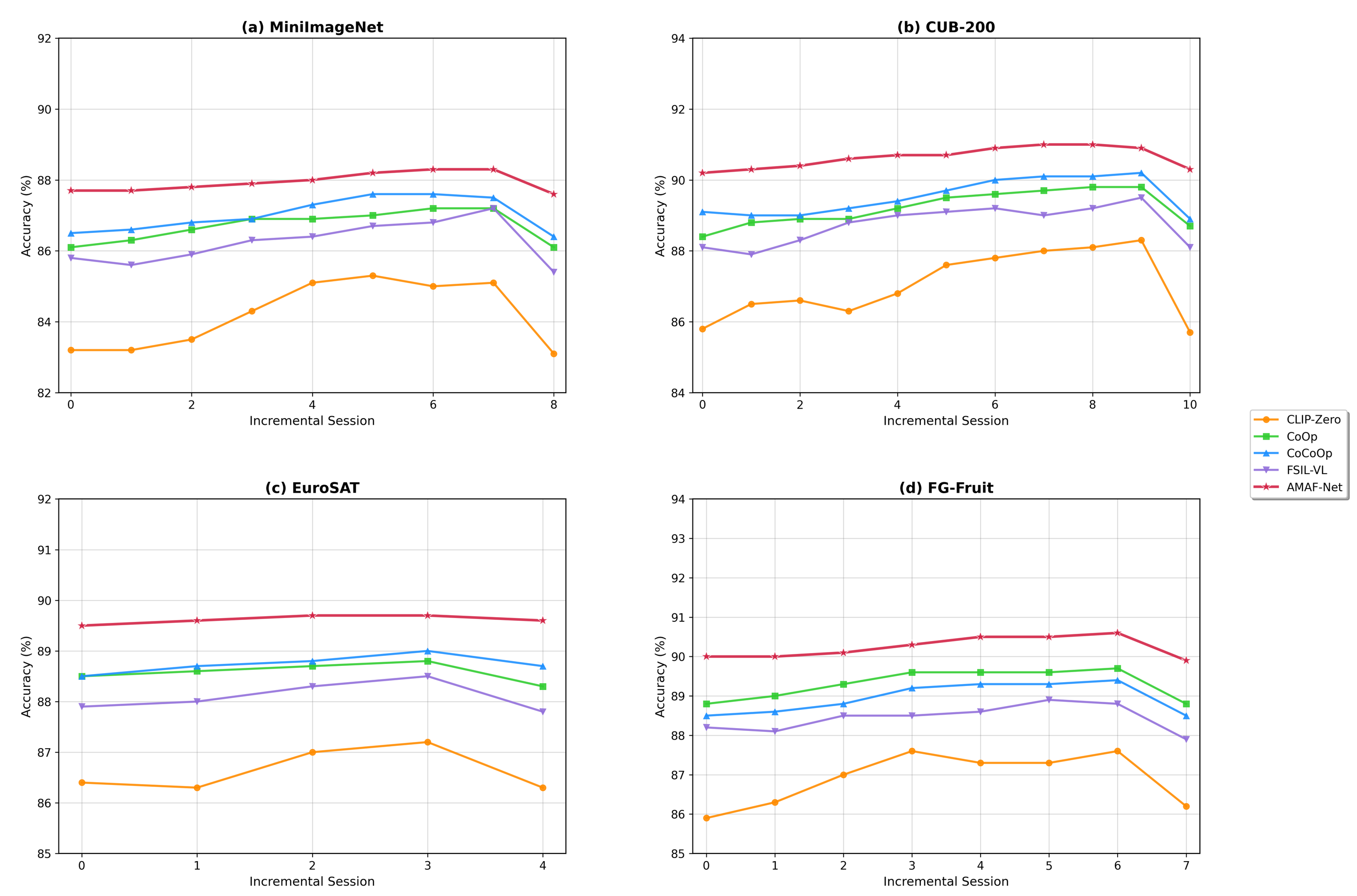
- Highest accuracy across all four FSCIL benchmarks.
- Stable performance under the training-free condition.

Fine-Grained Ability:

- 90.2% accuracy on FG-Fruit.
- 90.6% accuracy on CUB-200.
- Strong at capturing subtle visual differences.

ANALYSIS & DISCUSSION

Analysis of quantitative results: Achieves the highest average accuracy across all four benchmarks, outperforming all training-free and VLM-based baselines.



Consistent Across All Sessions:

- AMAF-Net consistently outperforms baselines.
- Prototype refinement slows accuracy degradation.
- Stable across all incremental sessions.

BIBLIOGRAPHY

References

Qian, L., Yu, T. and Yang, J., 2023. Multi-Scale Feature Fusion of Covariance Pooling Networks for Fine-Grained Visual Recognition. Sensors, 23(8). https://doi.org/10.3390/s23083970.

Tian, S., Li, L., Li, W., Ran, H., Ning, X. and Tiwari, P., 2024. A survey on few-shot class-incremental learning. Neural Networks, https://doi.org/10.1016/j.neunet.2023.10.039.

Trongtirakul, T., Agaian, S., Oulefki, A. and Panetta, K., 2023. Method for Remote Sensing Oil Spill Applications Over Thermal and Polarimetric Imagery. IEEE Journal of Oceanic Engineering, 48(3). https://doi.org/10.1109/JOE.2023.3245759.

Wang, H., Huang, R. and Zhang, J., 2022. Research Progress on Vision–Language Multimodal Pretraining Model Technology. Electronics (Switzerland), https://doi.org/10.3390/electronics11213556.

Scan to access the full paper

Presented at the first International Conference on Artificial Intelligence: Impacts and Potentials (ICAI-IP) on 4 December 2025 in Hanoi, Vietnam.