Test Time Adaptation

in Remote Sensing

Masters Thesis Phase-1

Ritwik Razneesh Manchi
24m0310

Guide
Prof. Biplab Banerjee

23 Oct 2025

Outline

Problem Statement

- Foundational models like CLIP are transforming computer vision with strong zero-shot capabilities.
- CLIP has enabled powerful zero-shot recognition by aligning images and text in a shared embedding space.
  - Numerous improvements have further advanced prompt learning and model calibration.
- However, direct application of CLIP and its derivatives to Remote Sensing (RS) images faces challenges:
  - Domain gap between RS and natural images.
  - Poor calibration and reduced accuracy on RS benchmarks.
  - Limited labeled data and unique semantic classes in RS.
- Therefore, there is a need for Remote Sensing-specific foundational models and adaptation strategies that can deliver robust, calibrated performance in this domain.

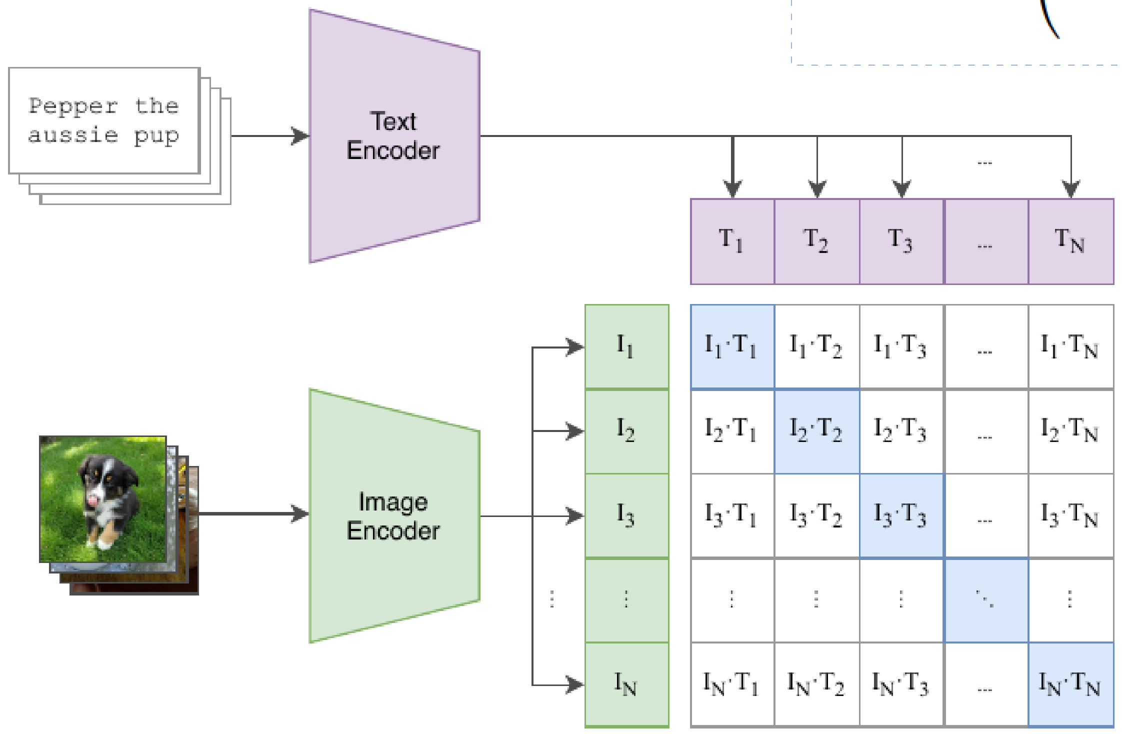
Literature Review

CLIP

Symmetric Cross-Entropy Loss
(a form of InfoNCE loss)

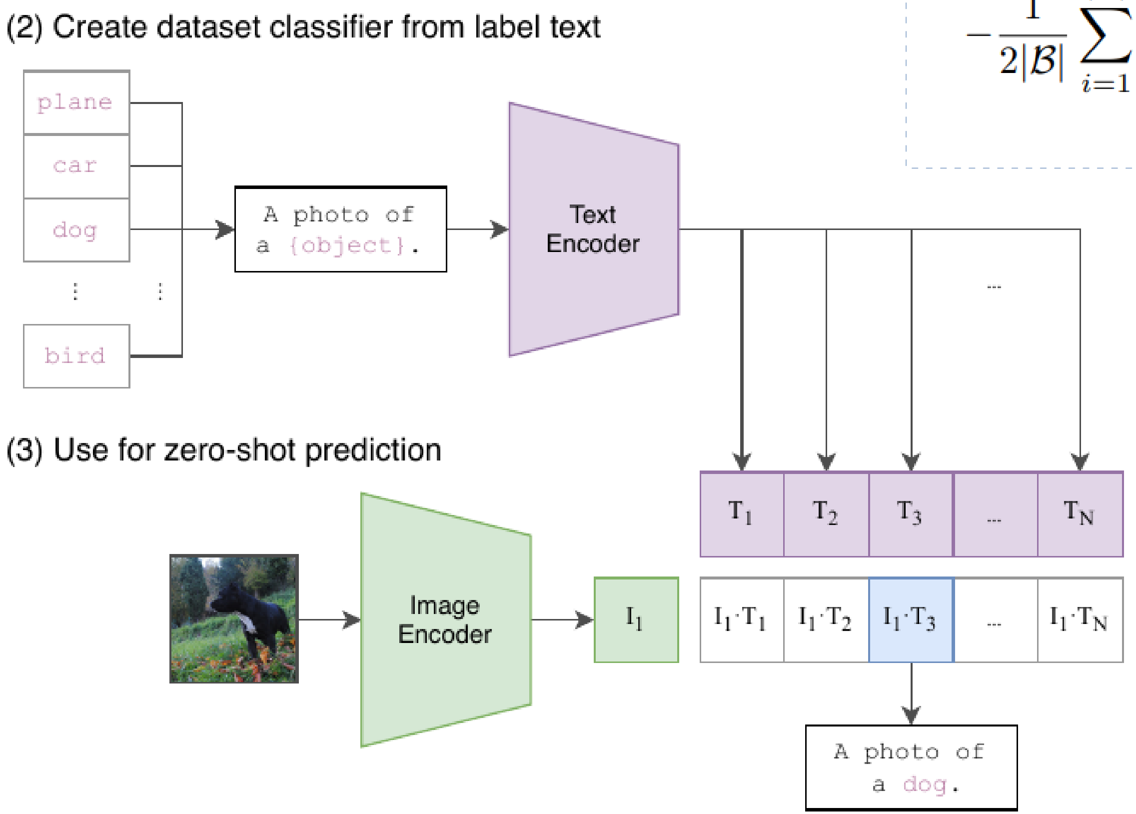
$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\rightarrow\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\rightarrow\text{image softmax}}\right)$$

(1) Contrastive pre-training



Literature Review

CLIP



Symmetric Cross-Entropy Loss
(a form of InfoNCE loss)

$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\to\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\to\text{image softmax}}\right)$$

Limitations :

Manually handwritten, hand-crafted prompts (suboptimal)

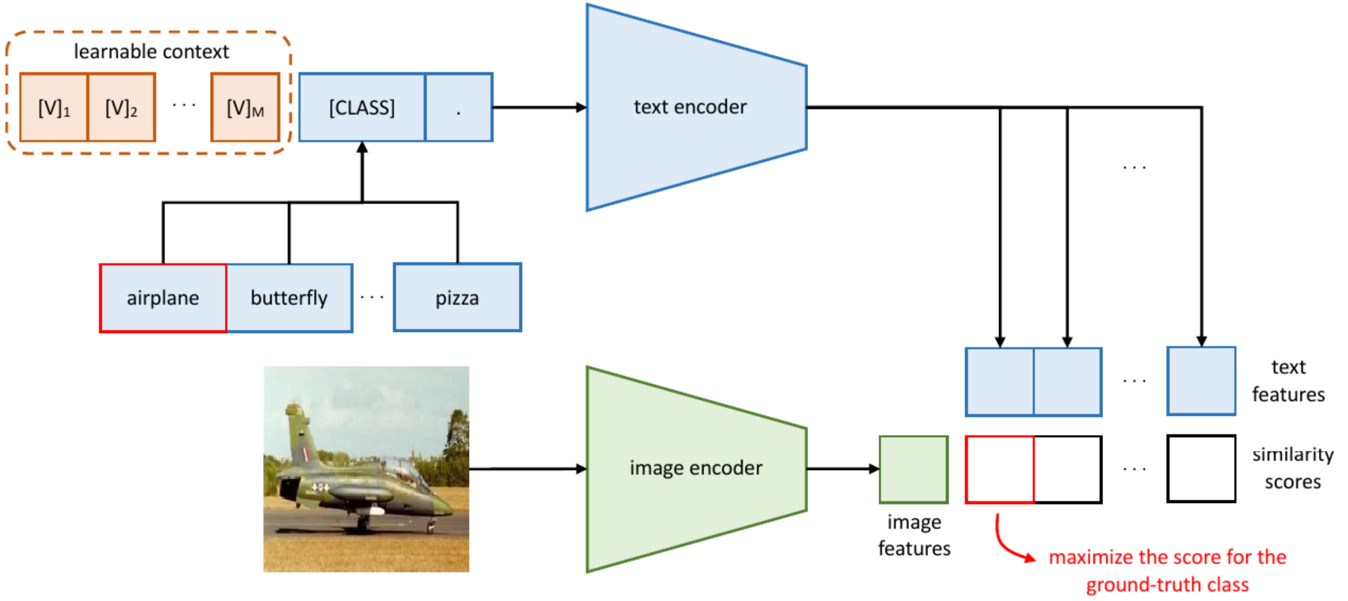
Small change in wording could cause a significant drop in performance

Literature Review
CLIP + CoOp

Cross-Entropy Loss

$$\mathcal{L}_{\text{CoOp}} = -\sum_{i=1}^{M} \mathbf{y}_i \log(p(\mathbf{y}_i|\mathbf{x}_i))$$

$$p(\mathbf{y}_k|\mathbf{x}) = \frac{\exp(\text{sim}(f(\mathbf{x}), g(\mathbf{t}_k))/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(f(\mathbf{x}), g(\mathbf{t}_j))/\tau)}$$



Limitations :

Learns a static set of context vectors shared by all classes

Not generalizable to unseen (new) classes

Overfitting the base classes

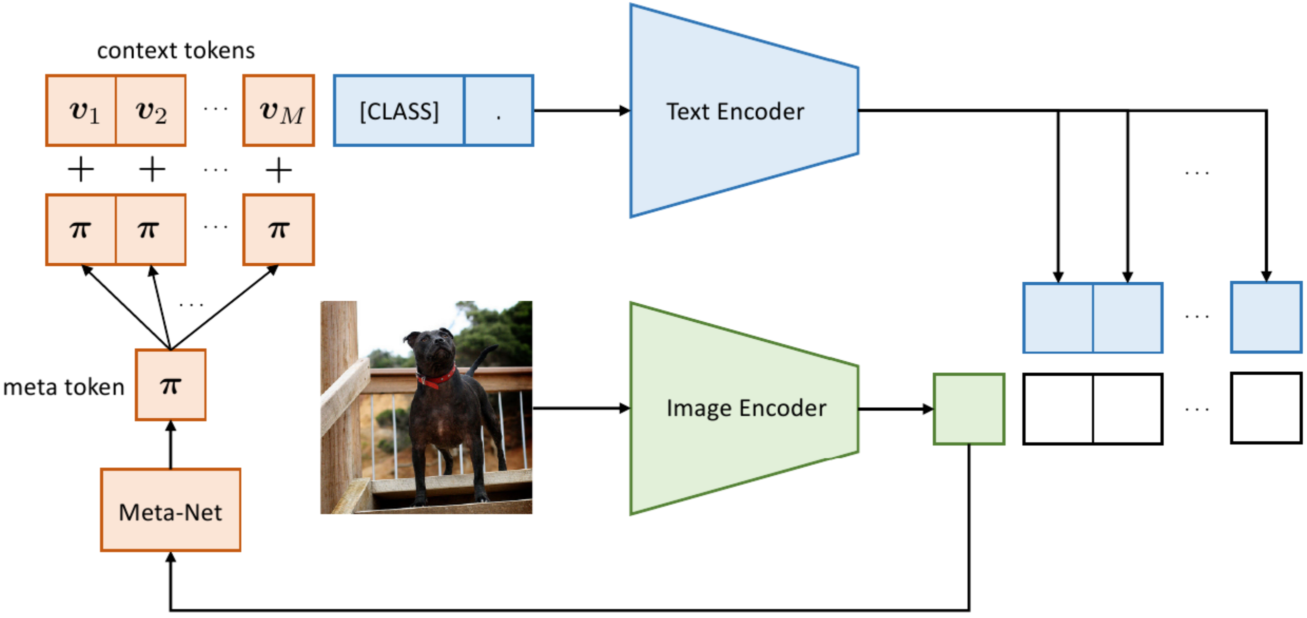
Literature Review

CLIP + CoCoOp

Cross-Entropy Loss

$$\mathcal{L}_{\text{CoOp}} = -\sum_{i=1}^{M} \mathbf{y}_i \log(p(\mathbf{y}_i|\mathbf{x}_i))$$

$$p(\mathbf{y}_k|\mathbf{x}) = \frac{\exp(\text{sim}(f(\mathbf{x}), g(\mathbf{t}_k))/\tau)}{\sum_{j=1}^{K}\exp(\text{sim}(f(\mathbf{x}), g(\mathbf{t}_j))/\tau)}$$



Limitations :

Requires labeled data

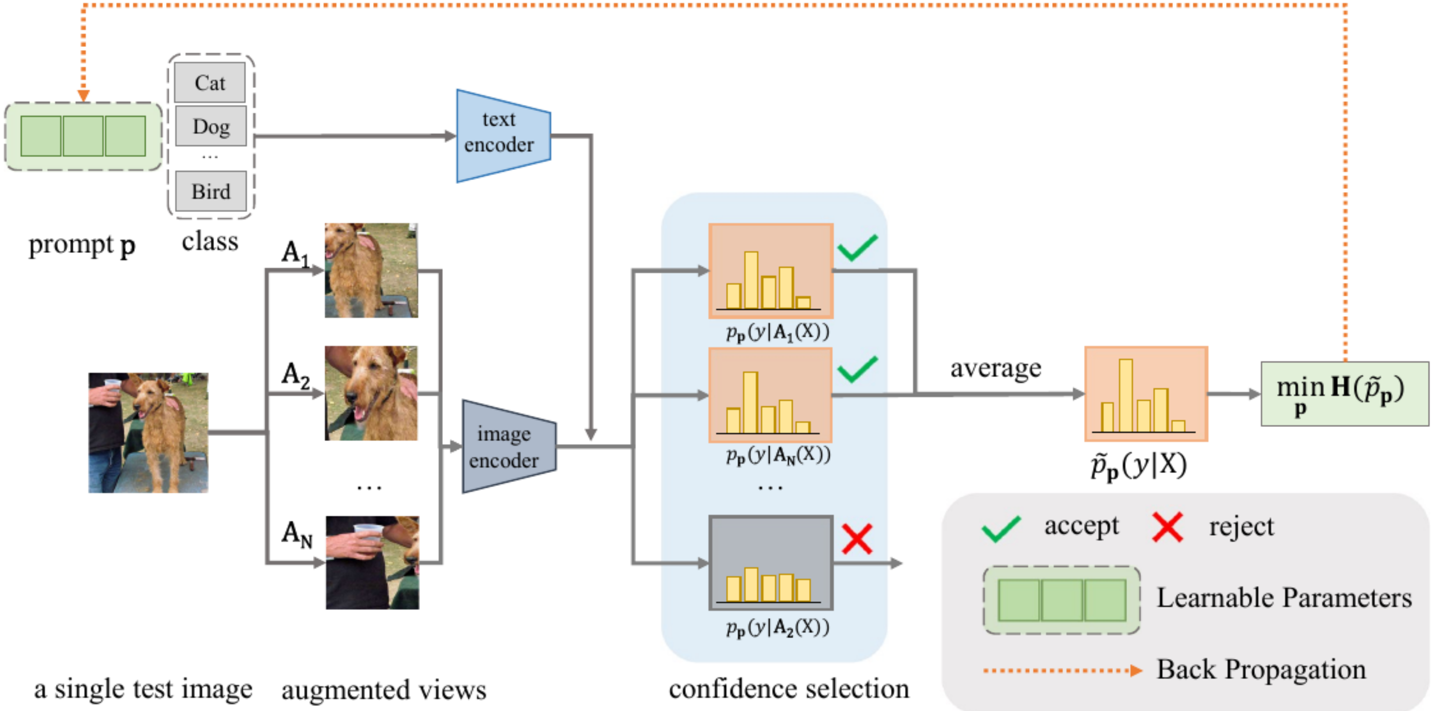
This restriction limits the generality of the underlying foundation models like CLIP

Literature Review
CLIP + TPT

$$p^* = \arg\min_{p} - \sum_{i=1}^{K} \tilde{p}_{p}(y_i|X_{\text{test}}) \log \tilde{p}_{p}(y_i|X_{\text{test}}),$$

Cross-Entropy Loss

$$\tilde{p}_{p(y|X_{\text{test}})} = \frac{1}{\rho N} \sum_{i=1}^{N} \mathbb{1}[\mathbf{H}(p_i) \leq \tau] p_{p}(y|\mathcal{A}_i(X_{\text{test}})),$$



Limitations :

Intrinsically leads to overconfident predictions and consequently increases the Expected Calibration Error (ECE) (i.e., poor calibration)

Compromised the model's reliability

Literature Review

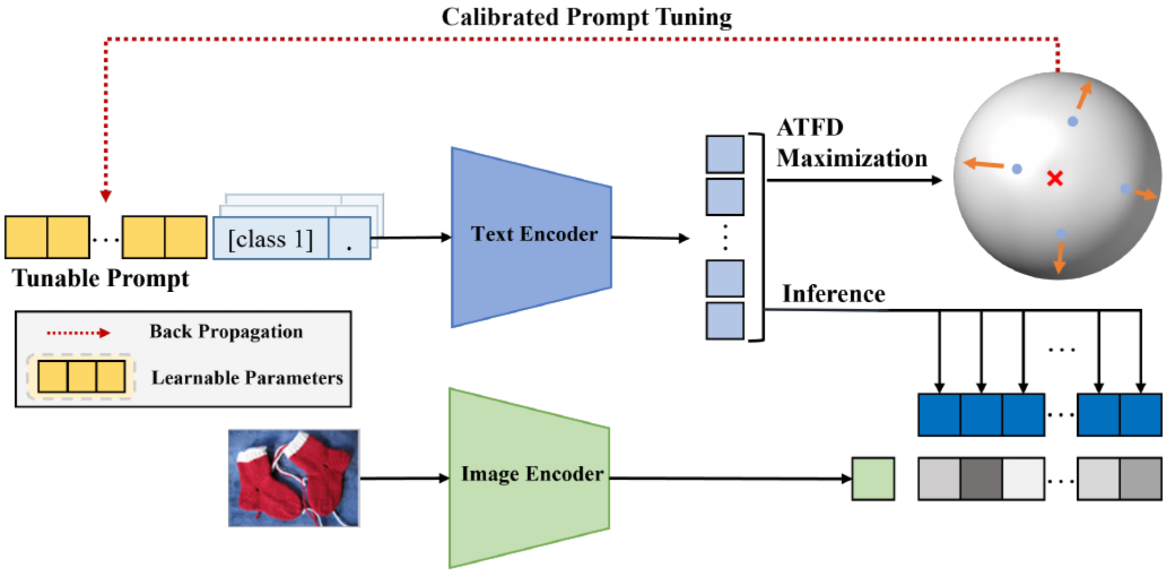
CLIP + C-TPT

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|A_m|}{N} \left| \mathrm{acc}(A_m) - \mathrm{conf}(A_m) \right|,$$

Cross-Entropy Loss + AFTD Maximization

$$\mathbf{p}^* = \arg\min_{\mathbf{p}} [\mathcal{L}_{\text{TPT}} + \lambda \cdot \mathcal{L}_{\text{C-TPT}}(\mathbf{t}_{[\mathbf{p};y_1]}, \mathbf{t}_{[\mathbf{p};y_2]}, \ldots, \mathbf{t}_{[\mathbf{p};y_N]})],$$

$$\mathrm{ATFD}(\mathbf{t}_{[\mathbf{p};y_1]}, \mathbf{t}_{[\mathbf{p};y_2]}, \ldots, \mathbf{t}_{[\mathbf{p};y_N]}) = \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{t}_{\text{centroid}} - \mathbf{t}_{[\mathbf{p};y_i]}||_2.$$



Limitations :

Underutilized the textual feature space, leading to suboptimal calibration

Overlooked the critical correlation between angular separation (cosine similarity) and calibration performance

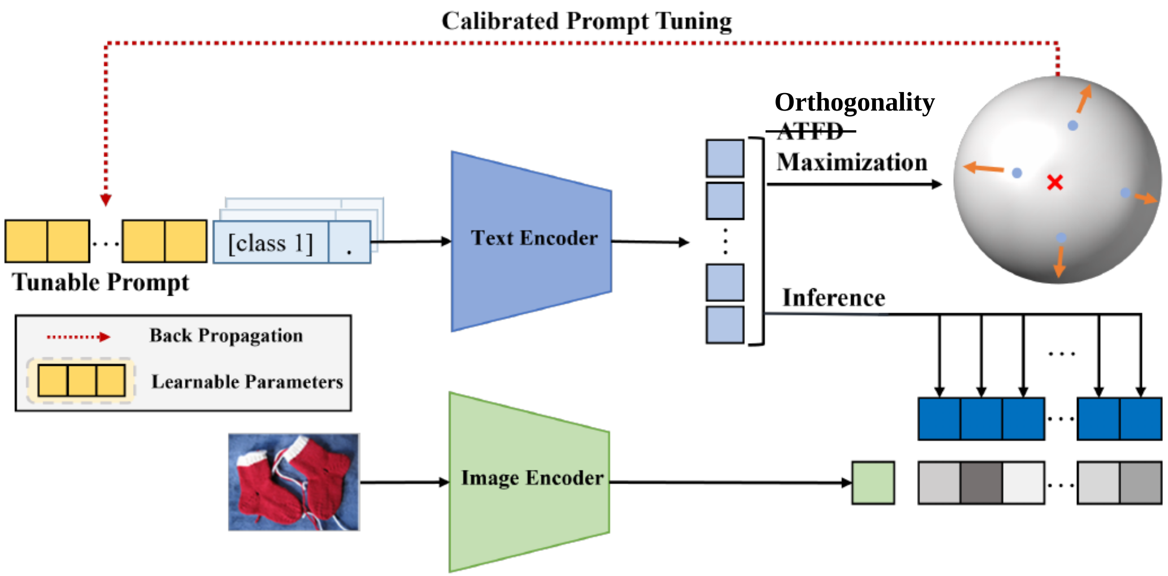
Literature Review

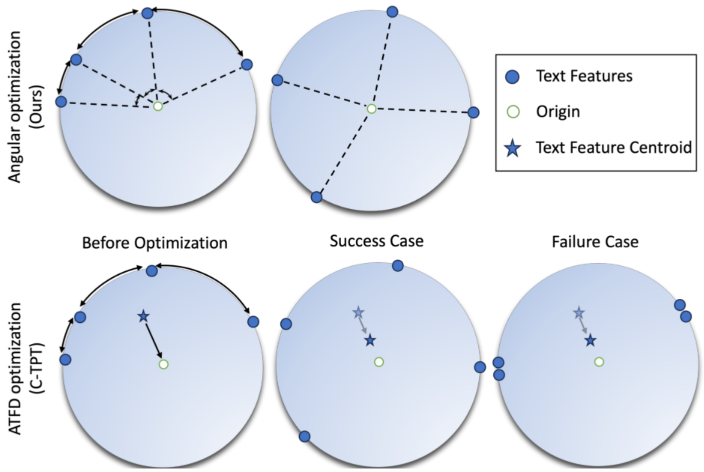
CLIP + O-TPT

Cross-Entropy Loss + Orthogonality Maximization

$$\mathbf{t}^* = \arg\min_{\mathbf{t}} (L_{TPT} + \lambda \|\mathbf{E}\mathbf{E}^T - I_C\|_2^2)$$

$$\text{ECE} = \sum_{m=1}^{M} \frac{|A_m|}{N} \left| \text{acc}(A_m) - \text{conf}(A_m) \right|,$$





Literature Review

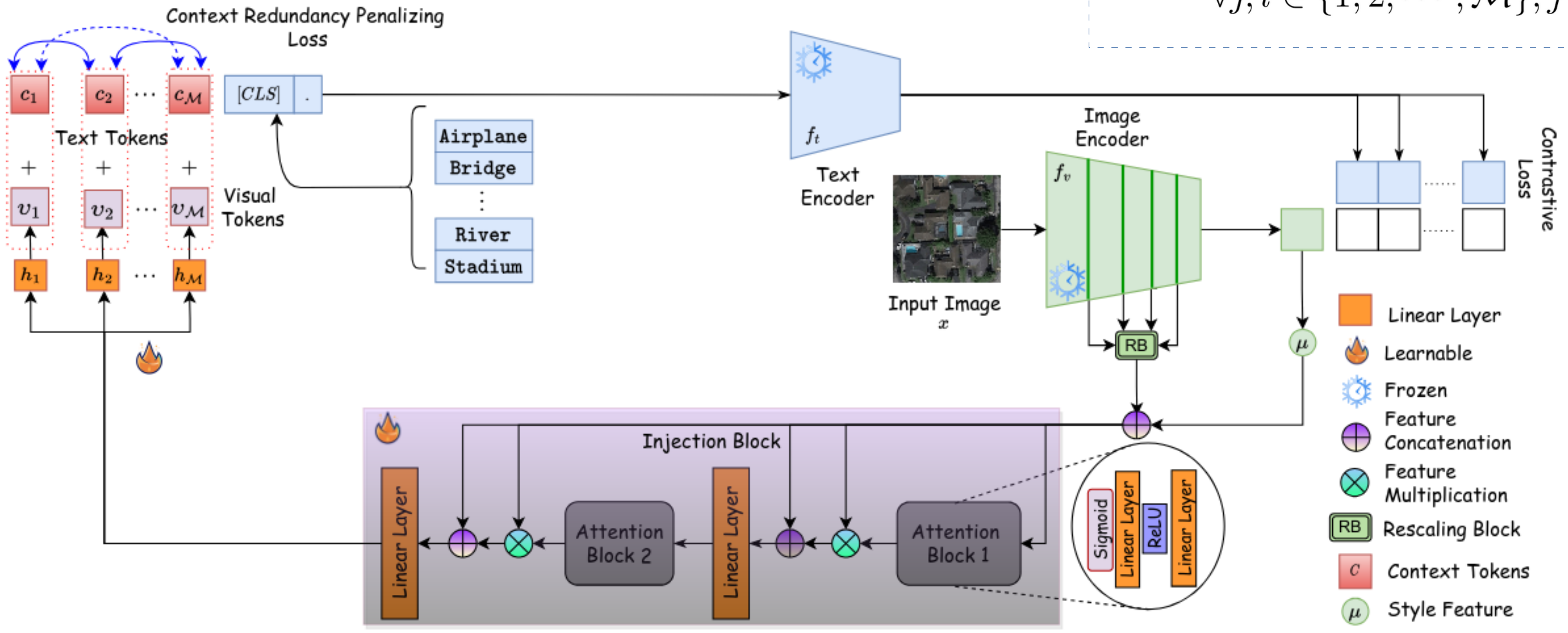
APPLENet

Cross-Entropy Loss + Context Redundancy Penalizing Loss

$$\mathbf{L}_{\mathbf{total}} = \underset{\mathcal{B}_{\phi},\{h_m\}}{\arg\min}[\mathbf{L}_{\mathbf{ce}} + \lambda * \mathbf{L}_{\mathbf{CRP}}]$$

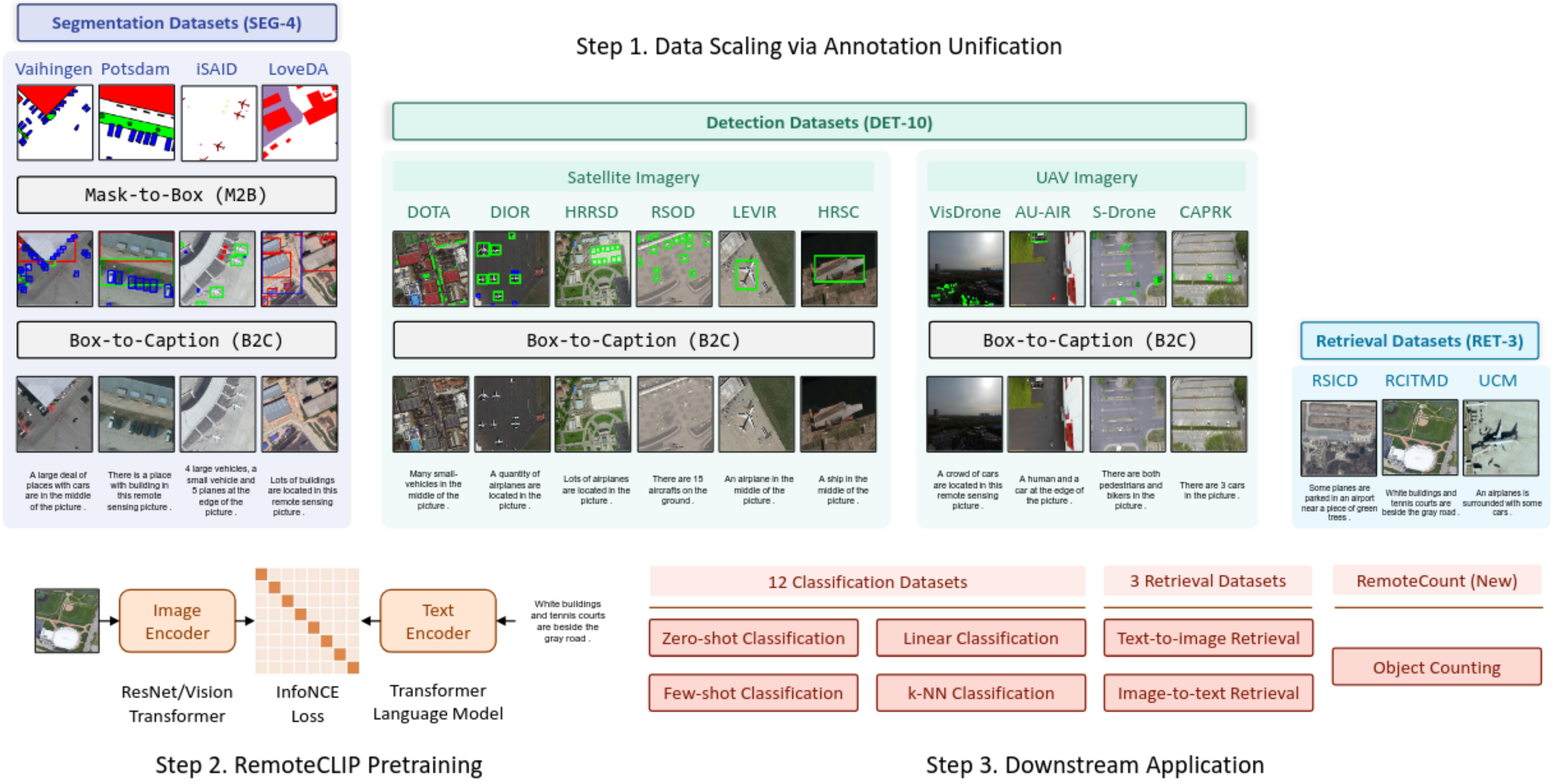
$$\mathbf{L}_{\mathbf{CRP}} = \underset{\mathcal{B}_{\phi},\{h_m\}}{\arg\min} \underset{(x,y)\in\mathcal{P}(\mathcal{D}_s)}{\mathbb{E}} \left|c'_j(x) \cdot c'_l(x) - \mathcal{I}\right|,$$

$$\forall j,l \in \{1,2,\cdots,\mathcal{M}\}, j \neq l, c'_j = c_j + v_j$$



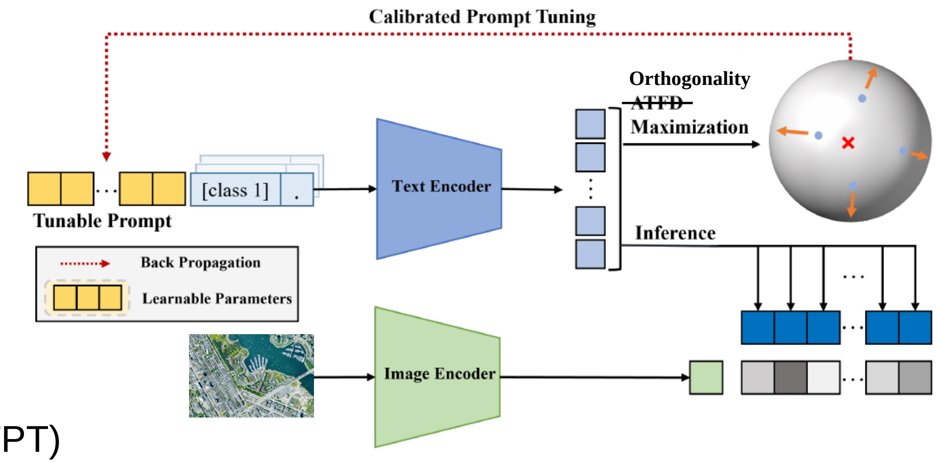
Literature Review

RemoteCLIP



Proposed Design

RemoteCLIP + O-TPT



Backbone:

RemoteCLIP (ViT-B/32) for remote-sensing domain priors

Head:

CLIP-style text prompts with learnable context (O-TPT)

Adaptation:

Test-time prompt tuning (entropy minimization + orthogonality regularization)

Evaluation:

Single-view inference (center-crop). Temperature scaling optional

Datasets:

AID, UCM, WHU-RS19, NWPU-RESISC45, EuroSAT, PatternNet

Metrics :

| Accuracy | ECE (Expected Calibration Error) |

Dataset	Model	Accuracy	ECE
UCM	openCLIP	72.380 %	4.614
	openCLIP + O-TPT	68.761 %	12.163
	RemoteCLIP	90.904 %	3.454
	RemoteCLIP + O-TPT	**91.095 %**	**3.348**
AID	openCLIP	66.600 %	5.541
	openCLIP + O-TPT	63.866 %	9.942
	RemoteCLIP	**87.833 %**	5.206
	RemoteCLIP + O-TPT	87.433 %	**1.998**
WHU-RS19	openCLIP	86.447 %	3.658
	openCLIP + O-TPT	77.631 %	8.525
	RemoteCLIP	**93.552 %**	5.070
	RemoteCLIP + O-TPT	92.763 %	**1.998**
NWPU-RESISC45	openCLIP	66.555 %	7.371
	openCLIP + O-TPT	61.266 %	13.914
	RemoteCLIP	65.822 %	**6.482**
	RemoteCLIP + O-TPT	**66.844 %**	7.865
EuroSAT	openCLIP	42.900 %	21.883
	openCLIP + O-TPT	48.500 %	**16.172**
	RemoteCLIP	34.099 %	18.9075
	RemoteCLIP + O-TPT	**35.499 %**	26.192
PatternNet	openCLIP	**58.95 %**	9.379
	openCLIP + O-TPT	54.447 %	10.476
	RemoteCLIP	51.578 %	**12.724**
	RemoteCLIP + O-TPT	54.736 %	13.578

Limitations

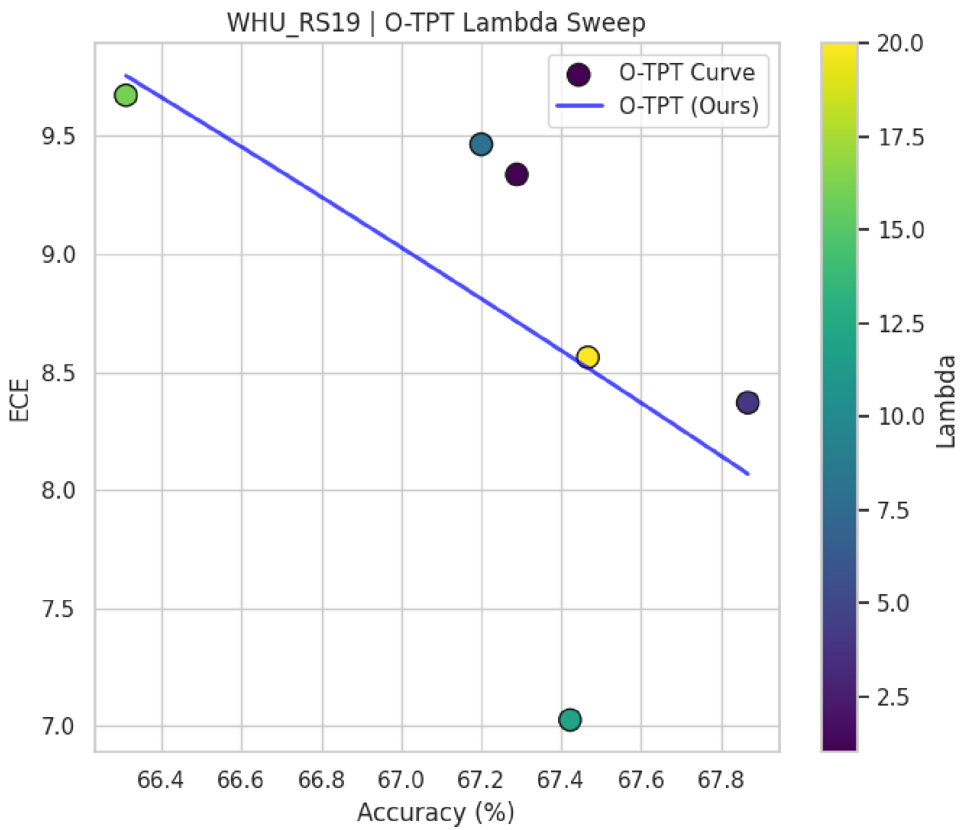
- Scene ambiguity in RS imagery
  - Single-label scenes often contain multiple semantic regions (e.g., residential + agriculture)
- Entropy-only adaptation signal can be brittle
  - Confident views may disagree on class; entropy of the mean remains high, driving noisy updates
- Variance across runs
  - Random crops is causing metric jumps

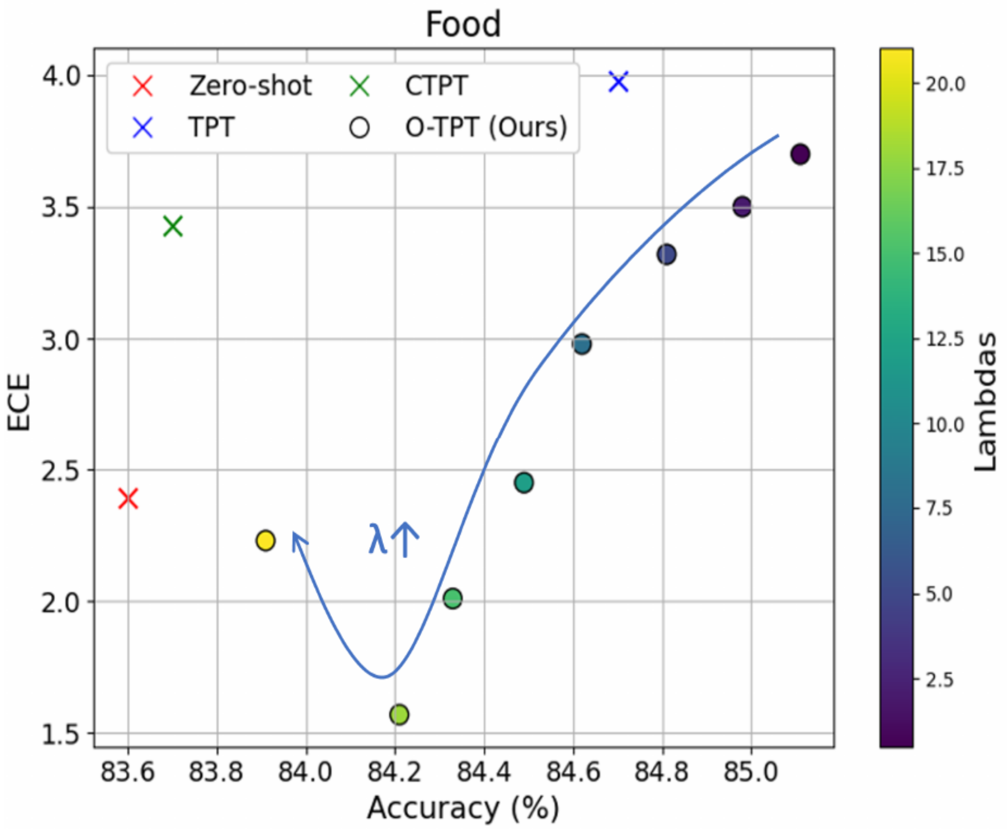
```
[remoteclip][eurosat][otpt] -> {    'top1': 0.2940370440483093,     'ece': 39.39869403839111    }
[remoteclip][eurosat][otpt] -> {    'top1': 0.30000001192092896,    'ece': 18.335366249084473   }
[remoteclip][eurosat][otpt] -> {    'top1': 0.335999995470047,      'ece': 29.383617639541626   }
[remoteclip][eurosat][otpt] -> {    'top1': 0.2849999964237213,     'ece': 33.8981330394749     }
[remoteclip][eurosat][otpt] -> {    'top1': 0.35499998927116394,    'ece': 26.1929452419281     }

[openclip][natural-images][otpt] -> {    'top1': 0.9912499785423279,    'ece': 1.29747549071908     }
[openclip][natural-images][otpt] -> {    'top1': 0.9850000143051147,    'ece': 1.4176726341247559   }
[openclip][natural-images][otpt] -> {    'top1': 0.9925000071525574,    'ece': 0.5291652865707874   }
[openclip][natural-images][otpt] -> {    'top1': 0.987500011920929,     'ece': 0.7389267906546593   }
[openclip][natural-images][otpt] -> {    'top1': 0.9862499833106995,    'ece': 0.9560159407556057   }
```

Limitations (Hypothesis)

- Assumption violation:
  TPT/O-TPT assumes label invariance across augmented views of the same image
- RS scenes break this:
  - Some crops show only agriculture, others residential, others mixed - each “rightfully” yields different class predictions
- Resulting dynamics:
  - Many low-entropy (confident) views but with different classes
  - Averaged prediction becomes multi-modal → high H(mean), inconsistent gradients
  - Leads to noisy λ trends and limited gains

Limitations (Hypothesis)





Future Work

References

I. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever.
Learning Transferable Visual Models From Natural Language Supervision. https://arxiv.org/abs/2103.00020

II. Kaiyang Zhou, Jingkang Yang, Chen Change Loy, Ziwei Liu.
Learning to Prompt for Vision-Language Models. https://arxiv.org/abs/2109.01134

III. Kaiyang Zhou, Jingkang Yang, Chen Change Loy, Ziwei Liu.
Conditional Prompt Learning for Vision-Language Models. https://arxiv.org/abs/2203.05557

IV. Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, Chaowei Xiao.
Test-Time Prompt Tuning for Zero-Shot Generalization in Vision-Language Models. https://arxiv.org/abs/2209.07511

V. Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, Chang D. Yoo.
C-TPT: Calibrated Test-Time Prompt Tuning for Vision-Language Models via Text Feature Dispersion. https://arxiv.org/abs/2403.14119

VI. Ashshak Sharifdeen, Muhammad Akhtar Munir, Sanoojan Baliah, Salman Khan, Muhammad Haris Khan.
O-TPT: Orthogonality Constraints for Calibrating Test-time Prompt Tuning in Vision-Language Models. https://arxiv.org/abs/2503.12096

VII. Mainak Singha, Ankit Jha, Bhupendra Solanki, Shirsha Bose, Biplab Banerjee.
APPLeNet: Visual Attention Parameterized Prompt Learning for Few-Shot Remote Sensing Image Generalization using CLIP. https://arxiv.org/abs/2304.05995

VIII. Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, Jun Zhou.
RemoteCLIP: A Vision Language Foundation Model for Remote Sensing. https://arxiv.org/abs/2306.11029

Thank You!