TORA: Parameter-Efficient Token-Driven Residual Adaptation of Remote Sensing Foundation Models for Multiscale Agricultural Vision

Haocheng Li, Juepeng Zheng, *Member, IEEE*, Shuangxi Miao, Kaiqi Du, Guilong Xiao, Zhiwei Zhang, Lingyuan Zhao, Haohuan Fu, *Fellow, IEEE*, and Jianxi Huang, *Senior Member, IEEE*

***Abstract*—Precision agriculture increasingly relies on robust visual understanding of crops and fields. However, agricultural scenes remain challenging due to large appearance variability, background clutter, and fine-grained multiscale structures. Although remote sensing foundation models (RSFMs) have demonstrated strong generalization in generic Earth observation tasks, adapting them effectively to domain-specific agricultural applications is still nontrivial. Existing parameter-efficient fine-tuning (PEFT) strategies often struggle to balance semantic adaptability across diverse crop phenotypes and domain shifts with spatial fidelity in preserving fine-grained structures and boundaries, which limits their effectiveness on dense prediction tasks. To this end, we propose token-driven residual adaptation (TORA), an efficient adaptation paradigm tailored to multiscale agricultural vision. TORA augments frozen RSFMs with hierarchical learnable semantic prototypes to explicitly model domain-specific semantic patterns and to interact with intermediate representations through tokenwise attention. Meanwhile, we design an instance-aware residual path to inject adaptation signals into backbone features, thereby strengthening domain alignment while effectively mitigating the degradation of spatial details. Furthermore, we establish a comprehensive evaluation benchmark spanning three representative tasks—semantic segmentation, object detection, and image classification—across nine datasets. Extensive experiments demonstrate that TORA trains only approximately 1.5%–3.8% of the backbone parameters, substantially fewer than full fine-tuning (FFT), yet achieves performance that is competitive with or superior to strong PEFT baselines and approaches FFT on multiple datasets.**

***Index Terms*—Agricultural vision, parameter-efficient fine-tuning (PEFT), precision agriculture, remote sensing foundation models (RSFMs).**

I. Introduction

UNDER the dual pressures of global population growth and constrained arable land, precision agriculture increasingly relies on robust visual understanding of crops and fields [1]. Traditional agricultural perception has largely depended on manual inspection or handcrafted feature pipelines, which often struggle in real-world field conditions with substantial appearance variability, cluttered backgrounds, severe occlusions, and illumination variations [2]. Consequently, these approaches have difficulty achieving both high accuracy and strong generalization for demanding applications such as crop phenotyping and fine-grained disease recognition [3], [4]. The challenge is further amplified in multiplatform agricultural sensing spanning proximal, UAV, and satellite observations.

Deep learning has been widely adopted in remote sensing and has improved task-specific applications such as land-cover recognition and parcel extraction [5]. However, many models remain task-specific and require customized architectures across categories, regions, and monitoring objectives, limiting cross-task transfer [6]. In addition, large-scale annotations are expensive for remote sensing imagery, which hinders sustained scaling of fully supervised training [7]. These challenges have motivated remote sensing foundation models (RSFMs) [8], which are pretrained on massive corpora in a self-supervised manner (e.g., RingMo [9], SatMAE [10], and SkySense [11]) and show strong transferability across downstream tasks. Nevertheless, existing RSFMs are largely developed for generic Earth observation scenarios, and their potential in agricultural applications remains underexplored. Efficiently adapting

Received 18 March 2026; revised 26 May 2026; accepted 28 May 2026. Date of publication 2 June 2026; date of current version 18 June 2026. This work was supported in part by the National Natural Science Foundation of China under Grant T2125006, Grant 42530109, and Grant 42401415; in part by Shenzhen Science and Technology Program under Grant KCXFZ20240903093759004 and Grant KJZD20230923115106012; and in part by Guangdong Science and Technology Program under Grant 2025B0101080001. *(Corresponding authors: Juepeng Zheng; Jianxi Huang.)*

Haocheng Li, Shuangxi Miao, Kaiqi Du, and Guilong Xiao are with the College of Land Science and Technology, China Agricultural University, Beijing 100083, China, and also with the Key Laboratory of Remote Sensing for Agri-Hazards, Ministry of Agriculture and Rural Affairs, Beijing 100083, China (e-mail: haoc_li@cau.edu.cn).

Juepeng Zheng is with the School of Artificial Intelligence, Sun Yat-sen University, Zhuhai 519082, China, and also with the National Supercomputing Center in Shenzhen, Shenzhen 518055, China (e-mail: zhengjp8@mail.sysu.edu.cn).

Zhiwei Zhang is with the School of Artificial Intelligence, Sun Yat-sen University, Zhuhai 519082, China (e-mail: zhangzhw65@mail2.sysu.edu.cn).

Lingyuan Zhao is with Huantian Wisdom Technology Company Ltd., Meishan 620564, China.

Haohuan Fu is with Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China, also with the National Supercomputing Center in Shenzhen, Shenzhen 518055, China, and also with the Ministry of Education Key Laboratory for Earth System Modeling and the Department of Earth System Science, Tsinghua University, Beijing 100084, China (e-mail: haohuan@tsinghua.edu.cn).

Jianxi Huang is with the College of Land Science and Technology, China Agricultural University, Beijing 100083, China, also with the Key Laboratory of Remote Sensing for Agri-Hazards, Ministry of Agriculture and Rural Affairs, Beijing 100083, China, also with the Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu 611756, China, and also with the School of Electronic Engineering, Guangxi University of Science and Technology, Liuzhou 545006, China (e-mail: jxhuang@cau.edu.cn).

Digital Object Identifier 10.1109/TGRS.2026.3699401

RSFMs to heterogeneous agricultural scenes under real-world production settings, thus, remains nontrivial.

Agricultural vision applications are highly fragmented, spanning crop classification [12], disease recognition under laboratory-to-field domain shifts [13], [14], field parcel and boundary delineation [15], [16], and emerging agricultural scene understanding benchmarks [17], [18]. Although RSFMs exhibit strong transferability, fully fine-tuning them for every task, sensor modality, and region is often impractical in real-world deployments. Parameter-efficient fine-tuning (PEFT), therefore, provides a more scalable solution for task-specific adaptation. However, agricultural remote sensing imposes more stringent requirements on PEFT than generic vision scenarios. On the one hand, the same agricultural target may exhibit drastically different visual scales and appearances across proximal, UAV, and satellite platforms, requiring strong scale adaptivity [19]. On the other hand, many agricultural applications rely on dense prediction, where success depends not only on semantic discrimination but also on preserving fine-grained spatial structures, such as crop boundaries, narrow parcel edges, leaf textures, and densely distributed small instances [20], [21]. In practice, insufficient spatial fidelity often manifests as broken parcel boundaries, blurred object contours, and missed small targets, which directly limits reliable agricultural interpretation. Existing PEFT paradigms often struggle to satisfy these requirements simultaneously: prompt-based methods mainly inject task-specific cues in token space but usually lack explicit constraints on intermediate spatial representations, which may weaken high-frequency details critical for delineating agricultural boundaries and textures [22]; in contrast, structure-oriented adapters typically rely on globally shared transformations, which are often insufficiently instance-adaptive for agricultural scenes with highly entangled foreground and background, such as crop–weed mixtures, orchard canopies under shadows, or farmland parcels embedded in heterogeneous surroundings [23]. This tension between semantic adaptability and spatial sensitivity limits the effectiveness of existing PEFT methods in agricultural remote sensing. To address this issue, we propose token-driven residual adaptation (TORA), a PEFT framework tailored to multiscale agricultural vision, aiming to better balance semantic adaptability and spatial sensitivity.

TORA is built upon RSFMs and is compatible with Transformer-based backbones. At each layer, it introduces learnable hierarchical tokens that interact with intermediate representations to generate instance-adaptive residual signals, enabling fine-grained adaptation while keeping most pretrained parameters frozen. With interlayer parameter sharing, TORA updates only approximately 1.5%–3.8% of the total parameters. As illustrated in Fig. 1, TORA achieves a favorable balance between parameter efficiency and performance compared with representative PEFT baselines. Therefore, we summarize our main contributions as follows.

1) We propose TORA, a parameter-efficient TORA framework for agricultural vision tasks. With RSFMs largely frozen, TORA enhances multiscale spatial awareness and instance-adaptive modeling using only a small set of trainable parameters.

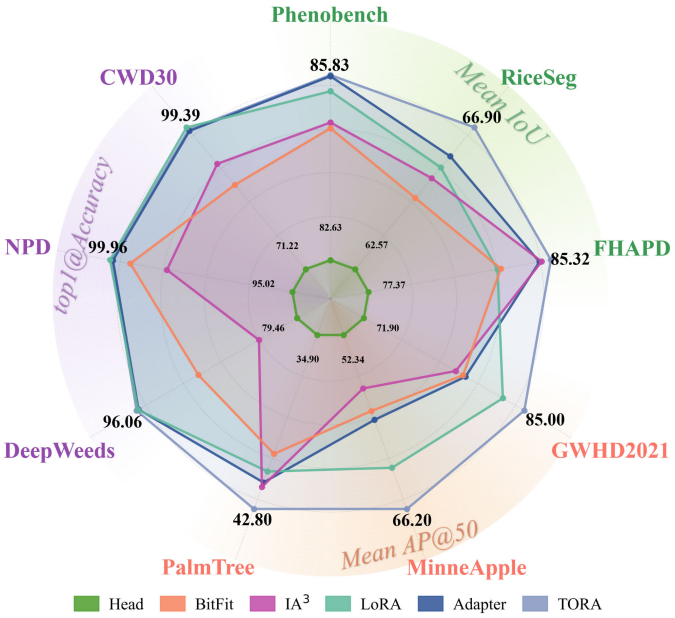

Fig. 1. TORA-integrated SatMAE++ achieves superior performance across nine datasets and three task categories, consistently outperforming five PEFT baselines.

2) We bridge semantic prompting and structure-oriented adaptation within a unified PEFT framework. TORA leverages learnable tokens and residual feature modulation to integrate semantic priors with spatial fidelity, enabling more balanced adaptation in complex agricultural remote sensing scenarios.
3) We establish a unified multitask agricultural benchmark covering image classification, semantic segmentation, and object detection. Results show that with approximately 1.5%–3.8% trainable backbone-adaptation parameters relative to full-backbone fine-tuning, TORA substantially narrows the gap to full fine-tuning (FFT) and achieves competitive or superior performance compared with representative PEFT baselines in most evaluated settings.

The remainder of this article is organized as follows. Section II reviews related work. Section III presents TORA in detail. Section IV reports the experimental results and analysis. Finally, Section VI concludes this article.

II. Related Work

A. Remote Sensing Foundation Models

Foundation models (FMs) have substantially improved the transferability of visual representations through self-supervised learning (SSL) on large-scale data [24], leading to the rapid development of RSFMs in Earth observation [25], where masked image modeling (MIM) [26], [27], [28] plays a central role. Driven by the multispectral, multitemporal, and multiscale nature of remote sensing data, a growing body of work has explored RSFMs from different perspectives. RingMo extends the MAE paradigm to optical remote sensing pretraining and demonstrates strong cross-task transferability [9]; SatMAE enhances remote sensing

scene understanding by explicitly modeling multispectral and temporal characteristics [10], and SatMAE++ further expands pretraining to multiscale settings with a reconstruction module to improve multiscale representation quality [29]. Focusing on scale robustness, ScaleMAE and Cross-Scale MAE introduce explicit scale-aware mechanisms and frequency-domain reconstruction to learn more stable multiscale representations [30], [31]. From the perspective of data construction and pretraining strategies, recent studies have further improved RSFMs through large-scale datasets, dynamic masking, and multitask pretraining [32], [33], [34]. Moreover, SkySense targets realistic multimodal spatiotemporal observation by pretraining on multitemporal optical and SAR data, enabling unified modeling across modalities and scales through spatiotemporal encoding and contrastive learning, and achieving competitive performance on a broad range of downstream tasks [11]. More recently, SkySense++ further advances this line by introducing a semantic-enhanced multimodal RSFM with progressive pretraining, improving generalization and few-shot transfer across diverse Earth observation tasks and domains [35]. Meanwhile, RSFM research for agricultural remote sensing is also emerging, where specialized models such as AgriFM are designed for crop recognition, farmland segmentation, and mapping [36].

However, in agricultural remote sensing, specialized foundation models typically require curating new large-scale pretraining datasets and repeating resource-intensive training [25], which makes it difficult to comprehensively meet the complex demands of multisource, multitemporal, and multitask applications. By contrast, leveraging existing general-purpose RSFMs and achieving cost-effective, reusable adaptation to agricultural scenarios is more practical. Therefore, we focus on PEFT and its potential for efficient adaptation in remote sensing and agricultural settings.

B. Parameter-Efficient Fine-Tuning

PEFT aims to adapt pretrained models to downstream tasks by freezing the backbone parameters and introducing only a small set of learnable parameters for efficient task adaptation [37]. Existing PEFT approaches can be broadly categorized into four groups: selective fine-tuning, structure-augmentation (module insertion), prompt learning, and reparameterization [38]. Specifically, selective fine-tuning updates only a limited subset of existing parameters, such as biases or normalization layers, as in BitFit [39] and LayerNorm tuning [40], which is simple to implement and incurs minimal parameter overhead. Structure-augmentation methods insert lightweight modules into the network and optimize only the added parameters to modulate the feature flow in a residual manner, such as Adapter [41] and AdaptFormer [42]. Prompt-learning methods introduce a small set of learnable prompt tokens into the input token sequence to steer task adaptation without modifying the backbone architecture; a representative example is visual prompt tuning (VPT) [22]. Building on this idea, recent work further explores multigranularity token learning to improve cross-domain robustness in semantic segmentation [43]. Reparameterization methods learn structured weight updates on top of frozen weights to inject task-specific information, including low-rank adaptations such as LoRA and DoRA [44], [45], while IA^3 modulates intermediate activations via channelwise multiplicative scaling factors within attention and feed-forward blocks [46]. These methods have been extensively validated on natural-image and general vision benchmarks, demonstrating favorable tradeoffs among parameter budget, training efficiency, and adaptation performance.

In contrast, systematic studies of PEFT in remote sensing remain relatively limited [47], and explorations targeting agricultural remote sensing are even scarcer. Agricultural data exhibit substantial cross-platform and cross-region variability [48] and frequently involve dense prediction tasks such as semantic segmentation and object detection, where lightweight adaptation must not only transfer semantic knowledge but also preserve spatial structures and fine-grained details. Under strong scale variation and complex background interference, generic PEFT designs often yield unstable gains, highlighting the need for PEFT mechanisms that are better aligned with the characteristics of agricultural remote sensing to enable cost-effective and generalizable task adaptation.

C. Adaptation Strategies for Dense Prediction

Building upon the aforementioned RSFMs, PEFT provides a cost-effective pathway to adapt foundation backbones to downstream tasks. In agricultural remote sensing, dense prediction tasks such as semantic segmentation and object detection demand not only high-level semantic understanding but also fine-grained modeling of pixel-level spatial structures and multiscale instances [49], [50], [51] thereby imposing additional constraints on PEFT [52], [53], [54], [55].

Head-only tuning can substantially reduce training costs, yet it is often insufficient to bridge the gap between generic backbone representations and the multiscale requirements of dense prediction [26]. Prompt-based methods introduce semantic priors by prepending learnable tokens; however, this guidance is injected implicitly in the token space, and high-frequency spatial details may be diluted as features propagate through deeper layers, limiting gains for small-object recognition and fine boundary delineation [52]. By contrast, structure-augmentation methods modulate feature flows in a residual manner (e.g., via bottlenecks or auxiliary branches) and have, therefore, become a more common choice for dense prediction [23]. In the remote sensing domain, recent approaches such as AiRs [56] and TEA [57] further combine adapter-style designs with RSFMs, achieving a better tradeoff between performance and efficiency. Complementary efforts explore unified multimodal fine-tuning frameworks that jointly adapt foundation backbones across heterogeneous remote sensing modalities for semantic segmentation [58]. Recent studies have also explored adapting foundation models such as SAM for remote sensing change detection [59], [60] and for semantic segmentation with explicit object and boundary constraints [61], further highlighting the importance of task-aware adaptation mechanisms for dense remote sensing tasks. Nonetheless, these methods still lack sufficient instance-aware and spatially selective adaptation, which limits their effectiveness in agricultural scenes with highly entangled foreground and background.

From the perspective of information injection and feature interaction, existing dense prediction approaches can be broadly grouped into two categories. One line emphasizes prompt injection but lacks explicit intermediate feature modulation, while the other relies on residual modulation yet remains limited in instance-aware, spatially selective adaptation; few methods unify both types of information within a single framework. Motivated by this observation, we propose TORA, which combines hierarchical learnable tokens with instance-aware residual modulation under an extremely low parameter budget, enabling fine-grained adaptation to the multiscale spatial structures required by dense prediction. Consequently, TORA better supports multiscale agricultural vision tasks.

III. Method

In this section, we present the proposed methodology. We first describe the RSFM backbone and its pretraining objective and then detail how TORA is integrated in a parameter-efficient manner, including its core residual generation mechanism.

A. Preliminaries

1) Transformer-Based RSFM Backbone: We adopt two representative pretrained RSFMs as backbones: SatMAE++ (ViT-based) and SkySense (SwinV2-based). Given an input multispectral image $X \in \mathbb{R}^{H\times W\times C}$, where $H \times W$ denotes the spatial resolution and C is the number of spectral channels, the backbone first maps X to an initial token sequence Z_0 via a patch-embedding stem. Although these two architectures differ in their downsampling schemes and attention scopes, their encoders can be uniformly formulated as a stack of Transformer blocks, enabling a unified integration of TORA across backbones. For the lth Transformer block, the input feature sequence is denoted as $Z_{l-1} \in \mathbb{R}^{N_l\times D_l}$, where N_l is the number of tokens (e.g., patch tokens in ViT or window-partitioned tokens in SwinV2) and D_l is the feature dimension at block l. For ViT-style backbones, the channel width is constant, i.e., $D_l \equiv D$ for all l. For hierarchical backbones such as SwinV2, blocks are grouped into S stages; we denote by $s(l) \in \{1,\ldots,S\}$ the stage index of block l, and the feature dimension is stagewise constant, i.e., $D_l = D^{(s(l))}$. If a class token is used, it can be treated as a special token in Z_{l-1} and exploited for classification. We adopt the standard prenorm Transformer formulation

$$\begin{aligned} \bar{Z}_l^{\mathrm{b}} &= Z_{l-1} + \mathrm{MSA}\left(\mathrm{LN}\left(Z_{l-1}\right)\right) \\ Z_l^{\mathrm{b}} &= \bar{Z}_l^{\mathrm{b}} + \mathrm{FFN}\left(\mathrm{LN}\left(\bar{Z}_l^{\mathrm{b}}\right)\right) \end{aligned} \tag{1}$$

where MSA corresponds to global self-attention in ViT and window-based self-attention in SwinV2. This generic block structure provides a unified interface for TORA, which injects residual adaptation signals at each layer conditioned on the intermediate features.

2) Problem Formulation: In FFT, given pretrained parameters Θ_{pre}, the model minimizes the task loss by updating all backbone and head parameters, which incurs substantial computation/storage costs and may increase overfitting risk. In contrast, we study parameter-efficient adaptation for agricultural vision tasks. For each dataset $t \in \{1,\ldots,T\}$, corresponding to a task type $\tau(t) \in \{\mathrm{cls}, \mathrm{seg}, \mathrm{det}\}$, we denote the training set as $\mathcal{D}_t = \{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^{n_t}$.

We freeze the backbone parameters Θ_{pre} and optimize the TORA parameters Θ_{TORA} together with a dataset-specific head ϕ_t for each dataset independently

$$\begin{aligned} &\Theta_{\mathrm{TORA}}^{*(t)},\ \phi_t^* \\ &\quad = \underset{\Theta_{\mathrm{TORA}},\phi_t}{\arg\min} \sum_{i=1}^{n_t} \ell_{\tau(t)}\left(g_{\phi_t}\left(f\left(x_i^{(t)};\Theta_{\mathrm{pre}},\Theta_{\mathrm{TORA}}\right)\right), y_i^{(t)}\right) \end{aligned} \tag{2}$$

where $f(\cdot)$ denotes the feature extractor comprising the frozen RSFM backbone augmented with TORA and $g_{\phi_t}(\cdot)$ is the decoding head for dataset t. We enforce parameter efficiency by constraining the trainable budget relative to the backbone size

$$\frac{|\Theta_{\mathrm{TORA}}|}{|\Theta_{\mathrm{pre}}|} \le \rho, \quad \rho \ll 1. \tag{3}$$

Under this constraint, our goal is to maximize generalization performance in multisource and multiscale agricultural scenarios.

B. Overall Architecture of TORA

As illustrated in Fig. 2, TORA is a lightweight, plug-and-play residual adaptation module inserted in parallel with every Transformer block of the frozen RSFM backbone in our default configuration, which is used in all main experiments. By operating on the intermediate token sequence Z_{l-1}, TORA can be integrated with different backbone architectures in a unified manner: whether the backbone uses global attention in ViT (e.g., SatMAE++) or window-based attention in SwinV2 (e.g., SkySense), TORA treats intermediate features as a generic token sequence and processes them consistently. This design allows TORA to recalibrate pretrained spatial representations without interfering with Swin's window partitioning and cyclic shifting.

1) Layerwise Semantic Prototypes: To accommodate diverse visual patterns in agricultural scenes, we maintain a set of learnable, layer-specific tokens in TORA, denoted as $\mathcal{T} = \{\mathcal{T}_1,\ldots,\mathcal{T}_L\}$. Specifically, the prototypes at layer l are represented as $\mathcal{T}_l \in \mathbb{R}^{K\times D_l}$, where K is the number of tokens per layer and D_l is the feature dimension. These tokens are initialized independently of the input and shared across images, serving as learnable semantic prototypes (an external memory) that capture domain-specific, fine-grained cues (e.g., crop textures, lesion patterns, and parcel boundaries). As fine-tuning proceeds, they are optimized in a task-driven manner to form hierarchical domain knowledge, thereby compensating for the limited agricultural semantics of the frozen backbone.

2) Residual Integration Mechanism: During the forward pass, TORA takes the token sequence entering the lth Transformer block, denoted as Z_{l-1}, and produces an additive residual that is injected into the frozen backbone output Z_l^{b}. Specifically,

$$Z_l = \begin{cases} Z_l^{\mathrm{b}} + \gamma_l\left[\mathbf{0}_{1\times D_l}; R_l\right], & \text{(with class token)} \\ Z_l^{\mathrm{b}} + \gamma_l R_l, & \text{(without class token)} \end{cases} \tag{4}$$

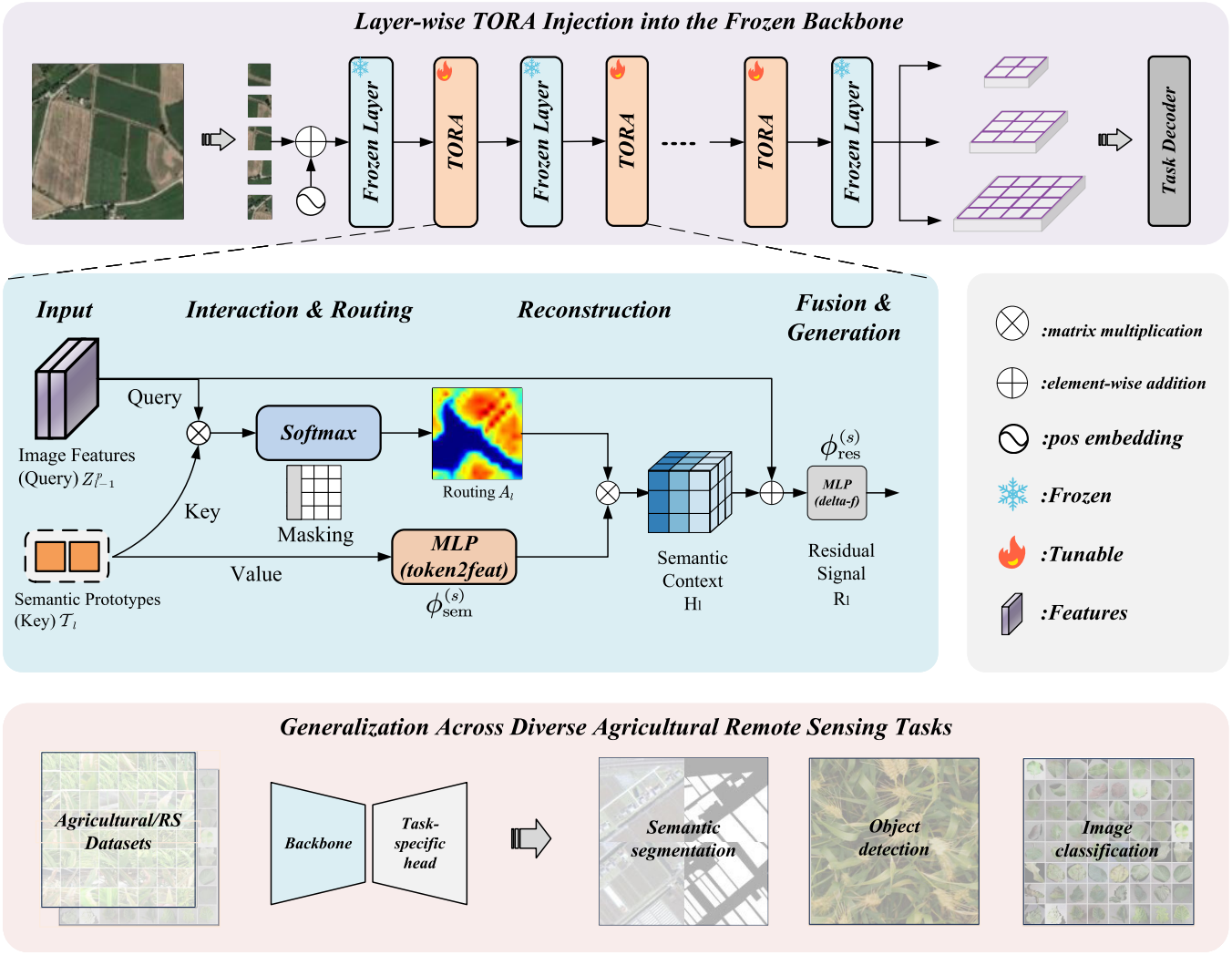

Fig. 2. Overall architecture of the proposed TORA. The pretrained backbone is kept frozen, and a lightweight TORA module is inserted in parallel with every transformer block. At each block, the normalized spatial tokens $\tilde{Z}_{l-1}^p$ interact with learnable semantic prototypes $\mathcal{T}_l$ to produce the routing matrix $\mathbf{A}_l$ and semantic context $\mathbf{H}_l$. The residual projection $\phi_{\text{res}}^{(s)}$ maps $\mathbf{H}_l + \tilde{Z}_{l-1}^p$ to a residual signal R_l, which is added back to the backbone stream. The adapted features are then forwarded to task-specific heads for semantic segmentation, object detection, and image classification.

where $R_l \in \mathbb{R}^{N_l^p \times D_l}$ is the patch-token residual produced by $F_{\text{TORA}}(\cdot)$ (detailed in Section III-C). For backbones with a class token, we denote $Z_{l-1} = [z_{\text{cls}}; Z_{l-1}^p]$ and inject the residual only into patch tokens by prepending a zero row $[\mathbf{0}_{1\times D_l}; R_l]$ so that the class token is unaffected by the *TORA residual*. For architectures without a class token (e.g., Swin), we set $Z_{l-1}^p \equiv Z_{l-1}$ and directly inject R_l. The scalar γ_l is a learnable layerwise scaling factor initialized to a small value (e.g., 10^{-3}) to preserve the stability of pretrained representations in early training.

We denote by $s(l) \in \{1, \ldots, S\}$ the stage index of block l, which determines the stage-shared projection parameters $\Theta_{\text{proj}}^{(s(l))}$; the concrete sharing scheme for hierarchical and ViT-style backbones is detailed in the Parameter Efficiency Strategy paragraph in the following. To ensure spatial fidelity and architectural compatibility, TORA adopts an architecture-aware processing strategy.

1) For ViT-style backbones, the token sequence includes a class token z_{cls}. We apply TORA only to patch tokens and keep the class token unaffected by prepending a zero row as in (4), thereby preserving the original classification pathway and semantic pooling behavior.
2) For Swin-style backbones, TORA operates directly on window-partitioned tokens and injects residual adaptation at each layer. Since this process is decoupled from the backbone's window-based attention computation, it requires no modification to window partitioning or the shift/reverse-shift operations.

With this architecture-aware design, TORA enables noninvasive feature adaptation across different spatial organizations without altering the topology of the pretrained backbone.

3) Parameter Efficiency Strategy: To minimize the trainable footprint, TORA shares the projection modules in F_{TORA} in a dimension-consistent manner. Specifically, the semantic prototypes $\{\mathcal{T}_l\}_{l=1}^L$ are kept layer-specific to capture hierarchical characteristics, whereas the projection parameters are shared within each stage. For hierarchical backbones (e.g., SwinV2), we instantiate $\{\Theta_{\text{proj}}^{(s)}\}_{s=1}^S$ and reuse $\Theta_{\text{proj}}^{(s)}$ for all blocks in

stage s to ensure channel-dimension consistency. For ViT-style backbones with constant feature dimension, this reduces to a single shared Θ_{proj} across all layers. In both cases, the number of prototypes per layer is held constant at $K = 100$ in all main experiments; what differs across stages of hierarchical backbones is the prototype feature dimension $D^{(s)}$, which inherits the per-stage channel width of the backbone. This design preserves parameter and storage efficiency while enabling consistent adaptation from low-level textures to high-level parcel structures.

C. Token-Driven Residual Generation

After describing the overall architecture of TORA and its stagewise parameter sharing strategy, we now elaborate on the internal mechanism of the core generation function F_{TORA}. This module transforms static semantic prototypes into dynamic residual signals via two steps: explicit semantic interaction and instance-aware feature refinement.

1) Explicit Semantic Interaction: Let $Z_{l-1}^p \in \mathbb{R}^{N_l^p \times D_l}$ denote the *spatial* token sequence entering the lth block. For ViT-style backbones, we decompose $Z_{l-1} = [z_{\text{cls}}; Z_{l-1}^p]$; for backbones without a class token (e.g., SwinV2), we simply have $Z_{l-1}^p \equiv Z_{l-1}$. Given layerwise semantic prototypes $\mathcal{T}_l \in \mathbb{R}^{K \times D_l}$, we establish dense associations between spatial tokens and prototypes within deep layers. For stable routing, we use prenormalized spatial tokens as queries

$$\tilde{Z}_{l-1}^p = \text{LN}\left(Z_{l-1}^p\right) \in \mathbb{R}^{N_l^p \times D_l}. \tag{5}$$

Specifically, we formulate $\tilde{Z}_{l-1}^p$ as queries and $\mathcal{T}_l$ as keys to compute the semantic affinity matrix

$$\mathbf{A}_l = \text{Softmax}\left(\frac{\tilde{Z}_{l-1}^p \mathcal{T}_l^\top}{\sqrt{D_l}} + \mathbf{B}_l\right) \tag{6}$$

where Softmax$(\cdot)$ is applied rowwise over the K prototypes and $\mathbf{B}_l \in \mathbb{R}^{N_l^p \times K}$ denotes a masking bias matrix in which invalid tokens (e.g., padded positions) are assigned $-\infty$ before Softmax. For ViT without padding, we set $\mathbf{B}_l = \mathbf{0}$.

To prevent attention collapse and foster interpretable routing distributions, we adopt a *null-token* strategy. Concretely, we designate the first prototype $T_{l,0}$ as a null prototype and apply stop-gradient only to its corresponding pre-Softmax affinity logits, i.e., the first column of $\tilde{Z}_{l-1}^p T_l^\top / \sqrt{D_l} + B_l$. This operation blocks the gradient path from the routing logits to $T_{l,0}$, preventing the null prototype from being directly optimized through competitive matching scores. Importantly, this design does not freeze the null-prototype embedding itself: $T_{l,0}$ remains a trainable parameter and is updated through the semantic reconstruction branch $H_l = A_l \phi_{\text{sem}}^{(s(l))}(T_l)$. Therefore, the null prototype can still serve as a learnable representational sink for unmatched or noisy tokens, while the remaining $K-1$ prototypes are encouraged to specialize in discriminative agricultural semantics, such as crop textures, lesion patterns, and parcel boundaries.

2) Instance-Aware Feature Refinement: We instantiate two lightweight stagewise shared projection modules, $\phi_{\text{sem}}^{(s)}$ and $\phi_{\text{res}}^{(s)}$, to map prototypes and fuse residuals. We denote their parameters collectively as $\Theta_{\text{proj}}^{(s)}$. Based on the semantic routing matrix $\mathbf{A}_l \in \mathbb{R}^{N_l^p \times K}$, we first aggregate the projected prototypes $\hat{\mathcal{T}}_l = \phi_{\text{sem}}^{(s(l))}(\mathcal{T}_l) \in \mathbb{R}^{K \times D_l}$, yielding a semantic-context representation

$$\mathbf{H}_l = \mathbf{A}_l \hat{\mathcal{T}}_l \in \mathbb{R}^{N_l^p \times D_l}. \tag{7}$$

While $\mathbf{H}_l$ injects strong semantic priors from the prototype space, relying solely on it may smooth out fine-grained spatial details (e.g., edges and textures).

To balance semantic guidance and spatial fidelity, we introduce an internal residual fusion path. Specifically, we fuse the semantic context $\mathbf{H}_l$ with the *spatial* token features via elementwise addition and generate a stagewise shared adaptation residual using $\phi_{\text{res}}^{(s(l))}$

$$R_l = \phi_{\text{res}}^{(s(l))}\left(\mathbf{H}_l + \tilde{Z}_{l-1}^p\right) \in \mathbb{R}^{N_l^p \times D_l} \tag{8}$$

where $\tilde{Z}_{l-1}^p$ denotes the prenormalized spatial tokens (i.e., patch tokens) and N_l^p is the number of spatial tokens at layer l.

Notably, R_l is content-adaptive due to the dynamic routing matrix $\mathbf{A}_l$, providing guided corrections to spatial representations rather than a fixed linear mixture.

Finally, the residual is scaled by a learnable layerwise scalar γ_l and injected into the backbone output Z_l^{b} to obtain the adapted features

$$Z_l = Z_l^{\text{b}} + \gamma_l \cdot \mathcal{P}(R_l) \tag{9}$$

where γ_l is initialized to a small value (e.g., 10^{-3}) to preserve the stability of pretrained representations during early training. Here, $\mathcal{P}(\cdot)$ is a padding operator that keeps the class token *unaffected by the TORA residual* for ViT-style backbones: $\mathcal{P}(R_l) = [\mathbf{0}_{1\times D_l}; R_l]$ if $Z_{l-1} = [z_{\text{cls}}; Z_{l-1}^p]$ and $\mathcal{P}(R_l) = R_l$ otherwise (e.g., Swin-style backbones without a class token).

In Section IV, we conduct extensive experiments on image classification, semantic segmentation, and object detection to demonstrate the effectiveness of the proposed method.

IV. Experiments

To comprehensively evaluate TORA in multisource agricultural remote sensing, we adopt a unified evaluation protocol spanning three task types, evaluating dense prediction first (segmentation and detection) and using classification as a complementary test of task generality. Our evaluation has two primary objectives: 1) to examine TORA's benefits for dense prediction, particularly in modeling fine-grained spatial structures and multiscale instances in agricultural scenes and 2) to assess whether TORA can retain the robustness of pretrained representations for global discriminative tasks under a strict parameter budget. We compare against representative PEFT baselines and use FFT as a strong high-capacity reference baseline, reporting results in terms of both predictive accuracy and parameter efficiency. The detailed hyperparameter configurations of the baselines and TORA are provided in Table I.

A. Datasets and Task Description

In this study, we establish a comprehensive evaluation benchmark comprising three task categories and nine datasets, spanning three sensing platforms—satellite, UAV, and proximal cameras. The imagery resolution spans orders

TABLE I

DETAILED HYPERPARAMETER CONFIGURATIONS. WE REPORT THE SPECIFIC SETTINGS FOR BOTH BASELINES AND TORA FOR REPRODUCIBILITY AND FAIR COMPARISON. ALL METHODS, INCLUDING TORA, ARE APPLIED IN PARALLEL WITH EVERY TRANSFORMER BLOCK OF THE FROZEN BACKBONE IN ALL MAIN EXPERIMENTS; NO LAYER SUBSET SELECTION IS USED

Method	Key hyper-parameters	Adapted modules / position
BitFit	–	Bias terms in all linear layers
LN-Tuning	–	Affine parameters of all LayerNorms
IA^3	–	Channel-wise scaling on K, V and FFN_{act}
Adapter	bottleneck ratio $= 1/16$, Act=GELU	Sequential after FFN (Pfeiffer config)
LoRA	r=16, α=32, Dropout $= 0.0$	Parallel to $\{W_q, W_v\}$ in Self-Attention
TORA (Ours)	K=100 per layer (both backbones, all stages); shared projection=True; dynamic scaling=True	Residual injected to block features ViT: global tokens; Swin: window-local tokens

TABLE II

STATISTICAL INFORMATION OF TASK DATASETS

	Dataset	Platform	Train Set	Test Set	Size	Category
Seg.	PhenoBench	UAV	1407	772	1024×1024	3
	RiceSeg	Proximal	2462	616	256×256	6
	FHAPD	Satellite	55184	13798	256×256	2
Det.	GWHD2021	UAV/Proximal	5198	1317	1024×1024	1
	MinneApple	Proximal	670	332	720×1280	1
	PalmTree	Satellite	22129	2408	512×512	1
Cls.	DeepWeeds	Proximal	14007	3502	256×256	9
	NPD	Proximal	70295	17572	256×256	38
	CWD30	Proximal	175816	43954	512×512	30

of magnitude, from centimeter-level observations capturing leaf textures to meter-level imagery covering parcel-scale structures, enabling a systematic assessment of model generalization under highly heterogeneous agricultural scenarios. The detailed dataset configurations are summarized in Table II.

1) Semantic Segmentation:

1) *PhenoBench [62]:* A high-resolution UAV dataset of sugar beet fields with image size 1024 × 1024 pixels. Key challenges include heavy leaf occlusion and weed interference, demanding robust separation of overlapping foliage and precise boundary delineation.
2) *RiceSeg [63]:* A cross-region rice semantic segmentation dataset consisting of proximal images collected from multiple countries and varieties. It features highly unstructured conditions, such as complex canopies at reproductive stages, tiny panicles, and water-surface reflections, necessitating robustness to background clutter and specular artifacts.
3) *FHAPD [15]:* A farmland-parcel extraction dataset derived from Gaofen (GF-1/2) satellite imagery. The nadir-view perspective and coarse-to-medium resolution require the model to delineate parcel boundaries from macroscopic texture patterns while distinguishing subtle scale-related appearance variations.

2) Object Detection:

1) *GWHD2021 [64]:* A global wheat head detection benchmark containing 270k annotated instances. It poses a stringent test for dense small-object detection due to extremely high object density and substantial appearance variation across genotypes and acquisition conditions.
2) *MinneApple [65]:* An apple detection dataset collected in complex orchard environments. Challenges such as severe occlusions from branches/leaves, frequent fruit overlap, and highly variable canopy shadows require robust feature extraction under heavy occlusion and illumination changes.
3) *PalmTree [66]:* A cross-region oil palm detection dataset annotated from high-resolution satellite imagery. It evaluates the ability to localize sparse tree crowns over large-scale remote sensing scenes characterized by cluttered backgrounds and substantial geographic variability.

3) Image Classification:

1) *DeepWeeds [67]:* A large-scale weed classification dataset collected from Australian rangelands. Highly similar textures and color tones between weeds and native pasture grasses, together with strong illumination variation, make it a challenging benchmark for fine-grained discriminability.
2) *New Plant Diseases (NPDs) [68]:* An extended version of PlantVillage [69] with 87k leaf images covering 38 healthy and diseased categories. It is widely used to evaluate disease-specific feature learning under controlled leaf-centric imagery.
3) *CWD30 [70]:* A large-scale crop–weed classification dataset for open-field environments with 210k images across 30 categories. Its multiview and multigrowth-stage characteristics require robust cross-view and cross-stage generalization.

B. Experiment Settings

1) Common Implementation Details: To assess the architectural generality of TORA across different RSFMs, we adopt SatMAE++ and SkySense as backbones and compare TORA with representative fine-tuning strategies under a unified evaluation protocol. For fair comparison, we employ the same task heads across all methods: UperNet for segmentation, RetinaNet for detection, and a linear classifier for classification. All experiments are implemented in PyTorch using

TABLE III

SEMANTIC SEGMENTATION RESULTS ON PHENOBENCH, RICESEG, AND FHAPD WITH SATMAE++ AND SKYSENSE BACKBONES. Δ DENOTES THE MIOU DIFFERENCE RELATIVE TO FFT UNDER THE SAME BACKBONE. "# PARAM" REPORTS TRAINABLE BACKBONE-SIDE PARAMETERS ONLY AND EXCLUDES THE TASK-SPECIFIC HEAD, WHICH IS IDENTICAL ACROSS ALL COMPARED METHODS

Method	# Param	SatMAE++					# Param	SkySense				
		PhenoB.	RiceSeg	FHAPD	Avg	Δ		PhenoB.	RiceSeg	FHAPD	Avg	Δ
Traditional Finetuning												
FULL	316M	86.09	67.12	86.82	80.01	0.00	654M	87.30	70.56	87.18	81.68	0.00
HEAD	0M	82.63	62.57	77.37	74.19	−5.82	0M	85.98	68.22	81.66	78.62	−3.06
PEFT Methods												
BITFIT	0.271M	84.91	64.75	83.14	77.60	−2.41	0.292M	87.41	70.08	83.98	80.49	−1.19
LN	0.100M	84.94	65.11	84.06	78.04	−1.97	0.162M	87.23	70.30	83.78	80.43	−1.24
IA^3	0.147M	85.01	65.35	84.92	78.43	−1.58	0.199M	87.29	69.90	82.81	80.00	−1.68
LORA	6.291M	85.55	65.68	82.99	78.07	−1.94	8.471M	86.76	70.18	84.20	80.38	−1.30
ADAPTER	6.442M	85.81	66.02	84.82	78.88	−1.13	13.56M	87.53	**71.20**	84.20	80.98	−0.70
TORA(OURS)	4.655M	**85.83**	**66.90**	**85.32**	**79.35**	**−0.66**	24.58M	**87.71**	70.47	**86.08**	**81.42**	**−0.26**

TABLE IV

OBJECT DETECTION RESULTS ON GWHD2021, MINNEAPPLE, AND PALM TREE WITH SATMAE++ AND SKYSENSE BACKBONES (REPORTED IN AP_{50}). "# PARAM" REPORTS TRAINABLE BACKBONE-SIDE PARAMETERS ONLY AND EXCLUDES THE TASK-SPECIFIC HEAD, WHICH IS IDENTICAL ACROSS ALL COMPARED METHODS

Method	# Param	SatMAE++					# Param	SkySense				
		GWHD	MApple	PTree	Avg	Δ		GWHD	MApple	PTree	Avg	Δ
Traditional Finetuning												
FULL	316M	88.30	68.90	44.00	67.07	0.00	654M	88.90	67.70	43.70	66.77	0.00
HEAD	0M	71.90	52.34	34.90	53.05	−14.02	0M	83.40	55.27	40.40	59.69	−7.08
PEFT Methods												
BITFIT	0.271M	80.00	58.40	40.30	59.57	−7.50	0.292M	86.50	58.70	42.00	62.40	−4.37
LN	0.100M	79.90	60.40	40.10	60.13	−6.93	0.162M	84.30	58.40	41.80	61.50	−5.27
IA^3	0.147M	79.40	56.60	41.80	59.27	−7.80	0.199M	84.50	57.70	41.70	61.30	−5.47
LORA	6.291M	83.25	62.92	41.10	62.42	−4.64	8.471M	**88.50**	66.20	42.70	65.87	−0.90
ADAPTER	6.442M	80.20	59.10	41.60	60.30	−6.77	13.56M	88.20	66.60	41.90	65.57	−1.20
TORA(OURS)	4.655M	**85.00**	**66.20**	**42.80**	**64.67**	**−2.40**	24.58M	88.14	**69.11**	**42.72**	**66.66**	**−0.11**

the OpenMMLab toolchain (MMSegmentation, MMDetection, and MMPretrain). We enable automatic mixed precision (AMP) and synchronized batch normalization (SyncBN) to improve training efficiency and stability. All models are trained with distributed data parallel (DDP) on a workstation equipped with four NVIDIA GeForce RTX 4090 GPUs. Unless otherwise specified, the backbone weights are kept frozen, and we use a consistent hyperparameter search budget across all compared methods. For consistency, the "# Param" column in Tables III–V reports trainable backbone-side parameters only and excludes the task-specific head; the same task-head architecture is used for all compared methods within each task category, but each method trains its own head weights independently. For Full fine-tuning, this column, therefore, corresponds to all backbone parameters, while, for PEFT methods, it corresponds to the parameters introduced or selected by each adaptation strategy on top of the frozen backbone.

2) Task-Specific Settings: We follow a unified experimental pipeline across the three downstream task categories. Backbone structural hyperparameters are kept identical to the pretrained configurations, while training-related hyperparameters are tuned in a fine-grained manner to account for the different convergence characteristics of the backbones.

a) Semantic segmentation: We set the input resolution of PhenoBench to 1024 × 1024 while utilizing a unified resolution of 256 × 256 for RiceSeg and FHAPD. Standard data augmentations are applied during training, including multiscale random resizing, random cropping, horizontal flipping, and photometric distortion. We adopt the AdamW optimizer (weight decay = 0.05) for all experiments. The learning rate schedules are tailored to the backbone characteristics: for

TABLE V

IMAGE CLASSIFICATION RESULTS ON DEEPWEEDS, NPD, AND CWD30 WITH SATMAE++ AND SKYSENSE BACKBONES (REPORTED IN ACC@1). "# PARAM" REPORTS TRAINABLE BACKBONE-SIDE PARAMETERS ONLY AND EXCLUDES THE TASK-SPECIFIC HEAD, WHICH IS IDENTICAL ACROSS ALL COMPARED METHODS

Method	# Param	SatMAE++					# Param	SkySense				
		DeepW.	NPD	CWD30	Avg	Δ		DeepW.	NPD	CWD30	Avg	Δ
Traditional Finetuning												
FULL	316M	96.46	99.96	99.41	98.61	0.00	654M	96.20	99.82	99.63	98.55	0.00
HEAD	0M	79.46	95.02	71.22	81.90	−16.71	0M	93.52	98.21	93.71	95.15	−3.40
PEFT Methods												
BITFIT	0.271M	89.67	99.41	87.98	92.35	−6.26	0.292M	96.26	99.69	94.13	96.69	−1.86
LN	0.100M	87.19	99.01	84.89	90.36	−8.25	0.162M	96.17	99.74	94.34	96.75	−1.80
IA^3	0.147M	83.40	98.42	92.15	91.32	−7.29	0.199M	94.55	99.44	94.20	96.06	−2.49
LORA	6.291M	95.92	**99.96**	**99.39**	**98.42**	**−0.19**	8.471M	96.86	99.93	98.66	98.48	−0.07
ADAPTER	6.442M	95.86	99.88	98.70	98.15	−0.46	13.56M	97.11	99.77	97.39	98.09	−0.46
TORA(OURS)	4.655M	**96.06**	99.91	98.91	98.29	−0.32	24.58M	**97.66**	**99.93**	**99.08**	**98.89**	**0.34**

ViT, we use cosine annealing with a base learning rate of 1.25×10^{-4}; for Swin, we employ a Poly schedule (power = 1.0) with a base learning rate of 5×10^{-4}. A linear warmup of 2k iterations is utilized for both backbones, with a total training duration of 20k iterations. Due to GPU memory constraints, per-GPU batch sizes are adjusted according to resolution: for the high-resolution PhenoBench (1024×1024), and the batch sizes are set to four and two for ViT and Swin, respectively; for the other datasets (256×256), they are increased to 8 and 4. We report mIoU as the primary evaluation metric.

b) Object detection: We resize MinneApple images to 1280×1280, maintain GWHD2021 at 1024×1024 and set Palm Tree to 512×512 to balance object scale and computational cost. Standard augmentations, including random horizontal flipping and photometric distortion, are applied during training. We use the AdamW optimizer (weight decay = 0.05) and train for 48 epochs with a cosine annealing schedule following a 5-epoch linear warm-up. Tailored to backbone-specific convergence behaviors, the base learning rate is set to 5×10^{-5} for ViT with a layerwise learning rate decay of 0.9 and 3×10^{-5} for Swin. Due to GPU memory constraints, the per-GPU batch size is 4 for MinneApple/GWHD2021 and 8 for Palm Tree with ViT, and 2 and 4 with Swin, respectively. For Swin, we further employ four-step gradient accumulation (effective batch size $\times 4$) to stabilize optimization. We report AP_{50} (average precision at IoU = 0.5) as the evaluation metric.

c) Image classification: All input images are processed to 224×224 using RandomResizedCrop during training. We train for 100 epochs, utilizing a 5-epoch linear warm-up followed by cosine annealing. The AdamW optimizer is adopted for all models, with backbone-specific hyperparameters tailored to the distinct scales and convergence behaviors of SatMAE++ and SkySense. Specifically, for ViT, we set the learning rate to 2×10^{-4}, the weight decay to 0.01, and the gradient clipping threshold to 1.0; for Swin, we use a learning rate of 5×10^{-6}, a weight decay of 5×10^{-4}, and a gradient clipping threshold of 5.0. In addition, random horizontal flipping is applied for augmentation. Due to GPU memory constraints, the per-GPU batch size is set to 64 for ViT and 32 for Swin. We report Acc@1 as the evaluation metric.

C. Compared Methods

To benchmark TORA, we compare it against seven fine-tuning strategies that span conventional fine-tuning and representative PEFT paradigms.

1) *General Baselines:*
   a) *Full Fine-tuning (Full)*: Updates all parameters of the pretrained backbone and heads. This setting serves as a strong performance upper bound but incurs the highest computational and storage costs.
   b) *Head-Tuning (Head)* [26]: Freezes the pretrained backbone and trains only the task-specific heads (e.g., classifiers or decoders). It serves as a minimal adaptation baseline and a reference for parameter efficiency.
2) *Representative PEFT Methods:*
   a) *Adapter [41]:* An additive method that sequentially inserts bottleneck modules into Transformer blocks to modulate features via residual connections.
   b) *LoRA [44]:* A reparameterization method that models weight updates as a low-rank decomposition $\Delta W = BA$, applied to specific linear layers (e.g., W_q, W_v) in attention or FFN blocks.
   c) *BitFit [39]:* A lightweight strategy that fine-tunes only the bias terms of the network, introducing one of the smallest parameter footprints among PEFT methods.

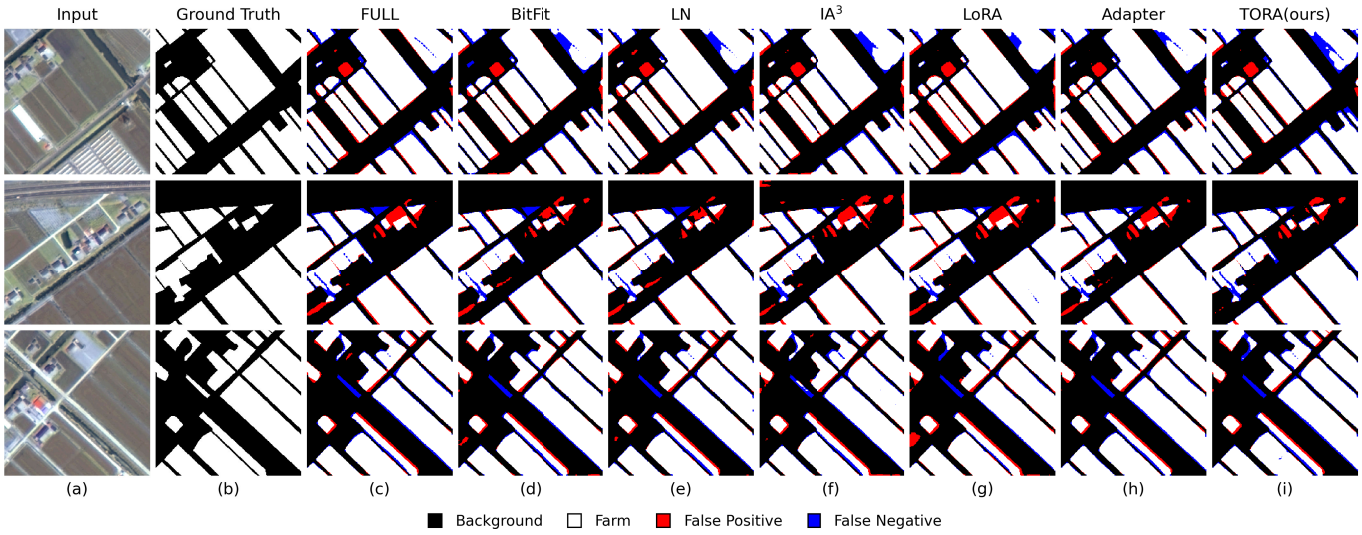

Fig. 3. Visualized comparisons on FHAPD (SkySense): (a) input image, (b) ground truth, (c) FULL, (d) BitFit, (e) LN-tuning, (f) IA^3, (g) LoRA, (h) Adapter, and (i) TORA (ours). Red and blue denote false positives (predicted farm but GT background) and false negatives (predicted background but GT farm), respectively (black: background; white: farm).

d) *LN-Tuning [40]:* Updates only the affine parameters (γ, β) of LayerNorm layers while keeping all other weights frozen.

e) *IA^3 [46]:* Introduces learnable vectors to elementwise rescale the activations in attention (keys and values) and feed-forward networks, achieving adaptation with extremely low parameter overhead.

To ensure fairness and reproducibility, we summarize the key hyperparameters of TORA and all baselines in Table I. The baselines strictly follow their original papers and public implementations, with only minimal, explicitly reported adjustments to ensure compatibility with ViT and Swin backbones. Specifically, we did not perform a per-dataset hyperparameter sweep: structural hyperparameters are fixed to widely adopted defaults (adapter bottleneck ratio of 1/16 and LoRA $r = 16$ with $\alpha = 32$), while BitFit, LN-Tuning, and IA^3 have no structural hyperparameters and only inherit the shared training settings. For TORA, $K = 100$ is chosen based on the sensitivity analysis on PhenoBench as discussed in Section V-B, and reused across all datasets without per-dataset retuning. The optimizer family, learning-rate schedule, total iterations, and augmentation pipeline are kept identical across all compared methods within each task category (see Section IV-B) so that no method is given a tuning advantage over the others.

D. Results

1) Semantic Segmentation: Table III provides a detailed comparison between TORA and representative PEFT baselines under two backbones, SatMAE++ (ViT) and SkySense (SwinV2). Overall, TORA effectively narrows the mIoU gap to FFT in most evaluated settings while retaining a small trainable parameter budget.

Specifically, with the SatMAE++ backbone, TORA achieves an average mIoU of 79.35%, outperforming other PEFT methods while introducing only 4.66M additional parameters—approximately 30% fewer than Adapter and LoRA. Notably, on RiceSeg, where domain shifts are more pronounced, TORA reaches 66.90% mIoU, surpassing Adapter (66.02%). This indicates that our residual-based explicit feature refinement helps preserve spatial structural consistency, thereby substantially improving cross-domain adaptation.

With the SkySense backbone, TORA further demonstrates strong compatibility with hierarchical window-based attention architectures. It achieves an average mIoU of 81.42%, outperforming the second-best Adapter (80.98%) and closely approaching (FFT, 81.68%) with a gap of only 0.26 percentage points. Moreover, on the small-scale PhenoBench dataset, TORA attains 87.71%, slightly exceeding FFT (87.30%), suggesting that the parameter-constrained residual adaptation may provide an implicit regularization effect and, thus, improve robustness in challenging agricultural scenes. Although Adapter is marginally better on RiceSeg, TORA delivers more pronounced gains on datasets such as FHAPD (+1.88% mIoU) and yields the highest average mIoU on this backbone, suggesting a generally favorable tradeoff across the multisource agricultural benchmarks rather than dataset-specific dominance.

As shown in Fig. 3, TORA yields cleaner parcel boundaries with fewer false positives (red) and false negatives (blue) than representative PEFT baselines, approaching the visual quality of FFT. BitFit/LN/IA^3 often miss thin strips and break elongated parcels, resulting in false negatives along long boundaries, whereas LoRA/Adapter is more prone to boundary bleeding and background confusion, introducing scattered false alarms near high-contrast edges. In contrast, TORA better preserves boundary continuity and sharp corners while suppressing fragmented errors, which aligns with its mIoU gains on FHAPD and indicates improved spatial fidelity under dense prediction.

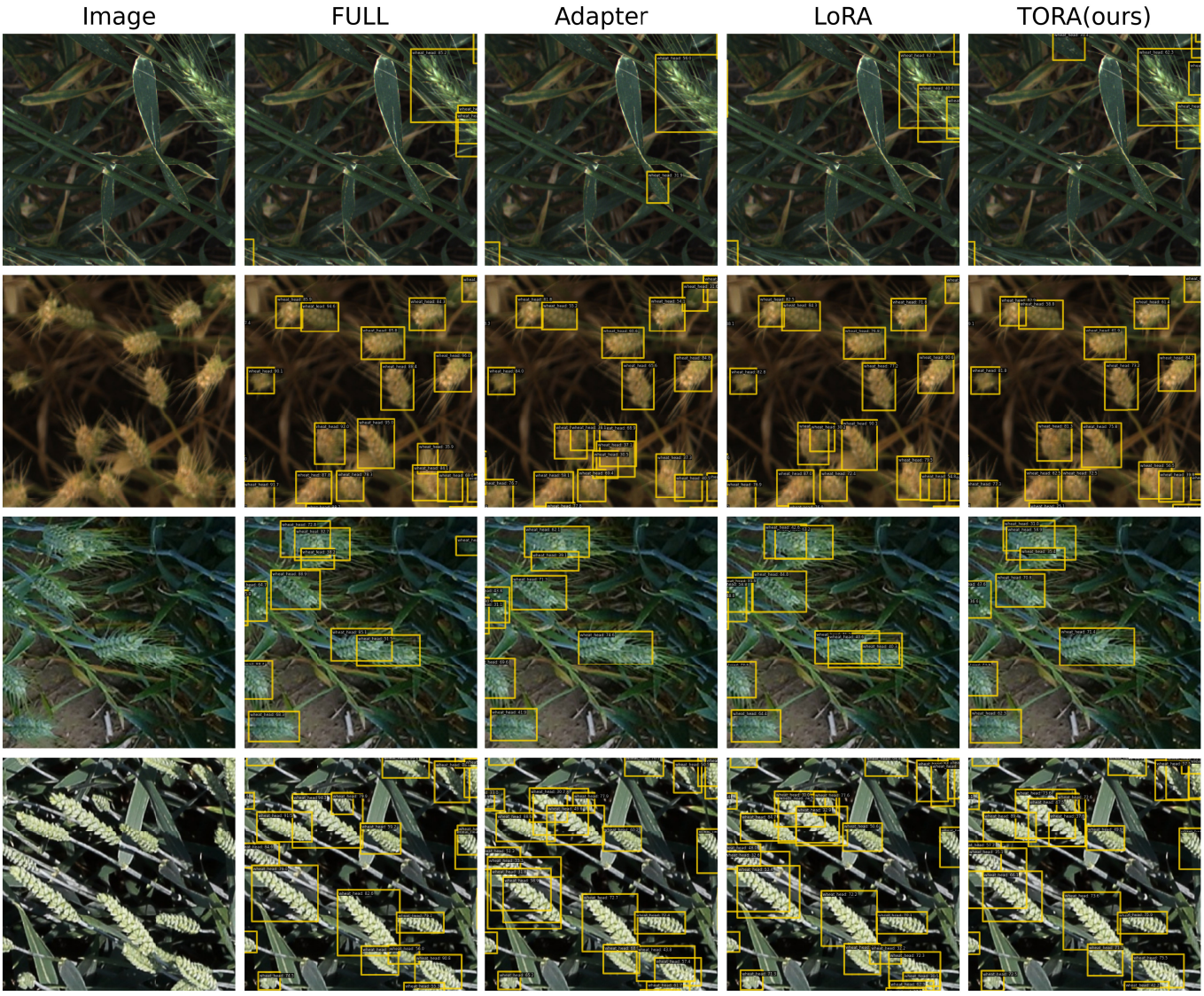

Fig. 4. Qualitative comparisons on GWHD2021 (wheat-head detection) with a SatMAE++ backbone. From left to right: image, full, Adapter, LoRA, and TORA (ours). Yellow boxes denote predicted wheat-head instances with confidence scores.

2) Object Detection: To assess the effectiveness of TORA for instance-level localization in dense prediction, we evaluate it on three object detection datasets, with results summarized in Table IV. Overall, TORA improves performance across datasets under both backbones with a small trainable budget and reduces the gap to FFT in most settings, indicating strong cross-dataset and cross-backbone robustness.

With the SatMAE++ backbone, head tuning yields an average AP_{50} of 53.05%, lagging significantly behind FFT (67.07%). This performance gap indicates that updating the detection head alone is insufficient to adapt pretrained representations to localization-intensive scenarios characterized by small objects and heavy occlusions. In contrast, TORA introduces only 4.66M trainable parameters yet improves the average AP_{50} to 64.67% (+11.62% over HEAD), achieving the best overall performance among PEFT baselines. Specifically, it obtains 85.00%, 66.20%, and 42.80% AP_{50} on GWHD2021, MinneApple, and Palm Tree, respectively. Moreover, despite using $\approx$ 30% fewer trainable parameters than LoRA and Adapter, TORA attains a higher mean AP_{50}. This consistency aligns with our hierarchical prototype-driven residual refinement, which effectively enhances localization-relevant features under challenging small-object and complex-background conditions.

Fig. 4 shows qualitative comparisons on GWHD2021 with the SatMAE++ backbone. Overall, TORA produces detection results that are visually closest to FFT, maintaining high coverage of wheat heads in cluttered and heavily occluded regions while avoiding excessive redundant predictions. In dense spike clusters, Adapter and LoRA tend to introduce more scattered low-confidence boxes and occasional duplicate/overlapping detections, and they may also yield less consistent box localization around partially occluded heads. By contrast, TORA yields a more balanced prediction set with cleaner localization in crowded areas and fewer spurious detections on background textures (e.g., leaves and stems). These

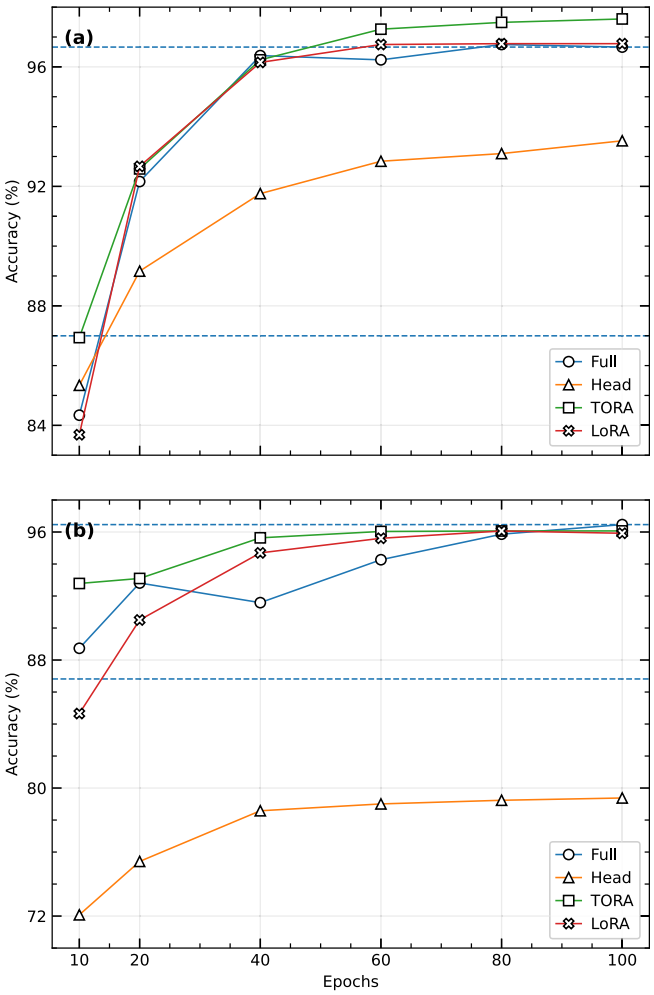

Fig. 5. Accuracy curves on DeepWeeds across training epochs using two pretrained RSFM backbones: (a) SkySense and (b) SatMAE++. We compare FFT, head-only training (Head), and PEFT methods with frozen backbones.

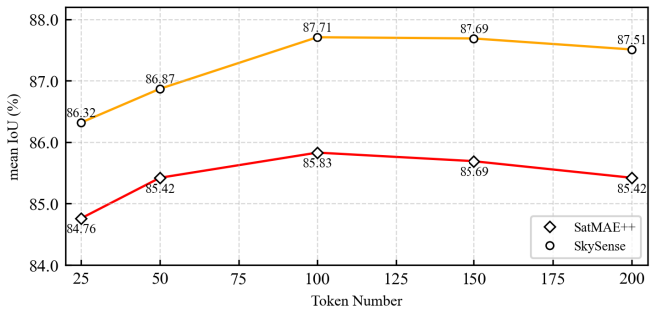

Fig. 6. Sensitivity to the number of semantic prototypes K on PhenoBench.

visual observations are consistent with the AP_{50} improvements reported in Table IV, suggesting that TORA's prototype-driven residual refinement enhances localization-relevant features for small and densely packed instances under complex farmland backgrounds.

With the SkySense (SwinV2) backbone, TORA achieves detection performance nearly on par with FFT. Specifically, while FFT obtains an average AP_{50} of 66.77%, TORA reaches 66.66%, narrowing the gap to a negligible 0.11%. At the dataset level, TORA attains 69.11% on MinneApple and 42.72% on Palm Tree, ranking first among all PEFT methods. On GWHD2021, it remains highly competitive (88.14%), trailing slightly behind LoRA (88.50%) and FFT (88.90%). Although Adapter also performs well with SkySense, TORA achieves the highest average AP_{50} among the compared PEFT methods on this backbone while remaining within 0.11% of FFT.

3) Image Classification: Finally, we evaluate TORA on image classification, with results reported in Table V. Since FFT already achieves near-saturated accuracy under both backbones in this setting, we primarily focus on the parameter efficiency and training dynamics of different methods in this performance-saturated regime.

With the SatMAE++ backbone, image classification primarily depends on global discriminative cues; consequently, the benefit of TORA in preserving fine-grained spatial cues is less pronounced than in dense prediction tasks. Nevertheless, TORA achieves a stable accuracy–efficiency tradeoff. It attains an average accuracy of 98.29% and delivers the best performance among PEFT baselines on the more challenging DeepWeeds dataset (96.06%), substantially outperforming lightweight selective fine-tuning methods. Although its average accuracy is slightly lower than LoRA (98.42%) under this backbone, TORA achieves comparable final performance with a smaller trainable budget. Taking DeepWeeds as an example, the convergence curves in Fig. 5(b) further reveal distinct training dynamics across methods. HEAD-only consistently remains the weakest and improves slowly, indicating that optimizing only the classifier is insufficient to bridge the gap between pretrained representations and downstream data. In contrast, both TORA and LoRA rapidly enter the high-accuracy regime within 10–40 epochs. Notably, TORA approaches FFT earlier and exhibits smoother late-stage convergence, whereas FFT shows larger mid-training fluctuations before gradually catching up. These observations suggest that, in ViT-based classification, TORA achieves FFT performance with a more robust training trajectory.

With the SkySense backbone, TORA exhibits a more pronounced advantage, achieving an average accuracy of 98.89% across the three classification datasets. It ranks first among all PEFT methods and even surpasses (FFT, 98.55%) by 0.34 percentage points. In comparison, LoRA and Adapter attain average accuracies of 98.48% and 98.09%, respectively. The DeepWeeds convergence curves in Fig. 5(a) further corroborate these results: HEAD-only is constrained by its limited trainable capacity, converging slowly with a noticeably lower performance ceiling, whereas TORA adapts rapidly within the first 20 epochs and maintains performance on par with, or slightly higher than, FFT throughout the mid-to-late training stage. Overall, even under the larger-capacity SwinV2 backbone and in this near-ceiling classification regime, TORA delivers equal or slightly better accuracy than FFT while updating only a small fraction of the backbone parameters and

TABLE VI

ABLATION STUDY OF TORA COMPONENTS ON FHAPD AND PHENOBENCH WITH SATMAE++ AS BACKBONE. Δ DENOTES THE MIOU CHANGE RELATIVE TO THE FULL MODEL ON EACH DATASET

Variant	FHAPD		PhenoBench	
	mIoU	Δ	mIoU	Δ
TORA (full)	85.32	0.00	85.83	0.00
w/o semantic routing	84.57	-0.75	84.79	-1.04
w/o spatial shortcut	84.12	-1.20	85.28	-0.55

exhibits faster and smoother convergence than the compared methods.

V. DISCUSSION

A. Ablation Experiments

To assess the contribution of each component in TORA, we conduct ablation studies on semantic segmentation. As a highly demanding dense prediction task, semantic segmentation imposes strict requirements on both spatial detail preservation and semantic parsing, thereby providing a direct testbed for TORA's key designs on multiscale spatial adaptation and instance-aware residual modulation. Unless otherwise specified, all ablations are performed with a *frozen SatMAE++ backbone* under an identical training protocol, and results are reported on FHAPD and PhenoBench (see Table VI).

1) Impact of Semantic Prototypes: We ablate the prototype-based semantic routing branch, including the semantic prototypes $\mathcal{T}_l$ and the affinity matrix $\mathbf{A}_l$, by replacing the prototype–feature interaction with a simple linear residual modulation branch. As shown in Table VI, removing $\mathcal{T}_l$ consistently degrades performance on both datasets, with mIoU dropping by 0.75 on FHAPD (85.32 $\rightarrow$ 84.57) and 1.04 on PhenoBench (85.83 $\rightarrow$ 84.79). These results indicate that linear modulation alone provides limited task-specific semantic calibration under a frozen backbone. In contrast, the learnable prototypes act as an external memory of explicit semantic anchors, enabling discriminative tokenwise routing and improving robustness across agricultural benchmarks.

2) Necessity of the Inner Residual Path: To assess the role of preserving spatial details during feature reconstruction, we ablate the inner residual connection in (8) by replacing $R_l = \phi_{\text{res}}^{(s(l))}(\mathbf{H}_l + \tilde{Z}_{l-1}^p)$ with $R_l = \phi_{\text{res}}^{(s(l))}(\mathbf{H}_l)$, where $\mathbf{H}_l$ denotes the prototype-aggregated semantic context and $\tilde{Z}_{l-1}^p$ provides the local spatial shortcut. This variant leads to consistent performance drops on both FHAPD and PhenoBench (see Table VI), decreasing mIoU by 1.20 (85.32 $\rightarrow$ 84.12) and 0.55 (85.83 $\rightarrow$ 85.28), respectively. The results suggest that updates generated solely from $\mathbf{H}_l$ may be insufficient to retain fine-grained spatial cues, potentially inducing overly smooth reconstructions. Reintroducing $\tilde{Z}_{l-1}^p$ provides a local structural reference to the residual generator, allowing context-guided refinement that remains grounded in the original spatial details. This design is conducive to more accurate boundary delineation and better preservation of small-scale structures across datasets.

TABLE VII

CONVERGENCE EFFICIENCY ANALYSIS ON SEMANTIC SEGMENTATION. WE COMPARE MIOU UNDER THE FULL SCHEDULE (20K ITERATIONS) AND A SHORTENED SCHEDULE (8K ITERATIONS, ~40% BUDGET)

Method	Training Schedule		Δ (Drop)
	20k iters	8k iters	
Traditional Tuning			
Full	86.82	83.36	−3.46
Head	77.37	75.79	−1.58
PEFT Methods			
BitFit	83.14	80.47	−2.67
LN	84.06	79.87	−4.19
IA^3	84.92	81.79	−3.13
LoRA	82.99	78.52	−4.47
Adapter	84.82	81.70	−3.12
TORA (Ours)	**85.32**	**82.63**	**−2.69**

B. Parameterization and Efficiency

Beyond the componentwise ablations, we discuss the performance–efficiency tradeoff of TORA and highlight practical design choices for compute-constrained adaptation.

1) Sensitivity to the Number of Prototypes K: Using PhenoBench as a representative benchmark, we investigate the effect of the prototype count K on both SatMAE++ and SkySense backbones by varying $K \in 25, 50, 100, 150, 200$ while keeping all other settings fixed. As shown in Fig. 6, increasing K consistently improves performance at first, suggesting that a richer prototype dictionary provides more expressive semantic anchors for tokenwise routing. The gain peaks around $K = 100$ for both backbones (SatMAE++: 85.83 mIoU; SkySense: 87.71 mIoU), after which performance saturates and slightly declines when using larger K (e.g., $K \geq 150$). This behavior indicates that $K = 100$ offers sufficient granularity to capture dominant semantic patterns in agricultural scenes, whereas an overly large dictionary may introduce redundant prototypes and increase optimization difficulty with limited additional discriminative benefit. Considering both accuracy and efficiency, we set $K = 100$ as the default in all main experiments.

2) Sensitivity to Training Budget: We further evaluate convergence efficiency on FHAPD by comparing the standard schedule (20k iterations) with a reduced schedule (8k iterations, ≈40% budget) in Table VII. Under the 8k setting, TORA achieves 82.63% mIoU, competitive with BitFit trained for the full 20k iterations (83.14%) and significantly outperforming IA^3 (81.79%) and Adapter (81.70%) under the same reduced budget.

In terms of robustness to budget reduction, FFT drops from 86.82% to 83.36% (−3.46%), whereas TORA decreases by only 2.69 points, retaining 96.85% of its 20k-iteration performance (versus 96.01% for FFT). These results indicate faster convergence of TORA, which aligns with our design: explicit semantic interaction and residual refinement likely provide informative priors early in training, making TORA

well-suited for compute-constrained or rapid-iteration agricultural applications.

3) Architecture-Dependent Parameter Overhead: The trainable footprint of TORA depends on the architecture of the underlying backbone. In TORA, the layerwise semantic prototypes contribute a term related to $\sum_l KD_l$, while the projection modules contribute architecture-dependent parameters determined by the feature dimensions and the projection-sharing scheme. For constant-width ViT backbones such as SatMAE++, all Transformer blocks share the same feature dimension, so the projection modules can be shared globally across layers, leading to a compact trainable footprint of 4.655M parameters. In contrast, hierarchical backbones such as the SwinV2-based SkySense contain multiple stages with different channel widths. To maintain channel-dimension compatibility, TORA uses stagewise projection modules, and the projection cost increases in high-dimensional stages. Consequently, TORA reaches 24.58M trainable parameters on SkySense, which is larger than LoRA (8.471M) and Adapter (13.56M), although it remains substantially smaller than FFT (654M). This behavior reflects the architecture-dependent nature of TORA's parameterization. Accordingly, TORA should be regarded as parameter-efficient primarily relative to FFT, rather than necessarily being the lightest PEFT method on every backbone.

C. Visualization

We visualize representative activation maps in Fig. 7 to further interpret the learned representations. As shown, TORA exhibits high-response regions that align more closely with semantically meaningful target areas, and it presents a clearer and more continuous high-to-low response transition along object boundaries. This indicates improved structural fidelity and reduced background interference. In contrast, FFT and LN-tuning show more pronounced background activation spillover in multiple examples, where strong responses extend beyond the target regions to surrounding textures and irrelevant areas, leading to a more diffuse response distribution. These observations suggest that TORA's semantics-guided residual modulation can enhance task-relevant semantic responses while suppressing spurious background cues, thereby encouraging more object-centric and discriminative representations. Overall, the qualitative evidence supports that TORA remains capable of reconstructing high-level semantic features and preserving spatial structural consistency under a frozen-backbone setting, providing a more reliable feature basis for downstream vision tasks.

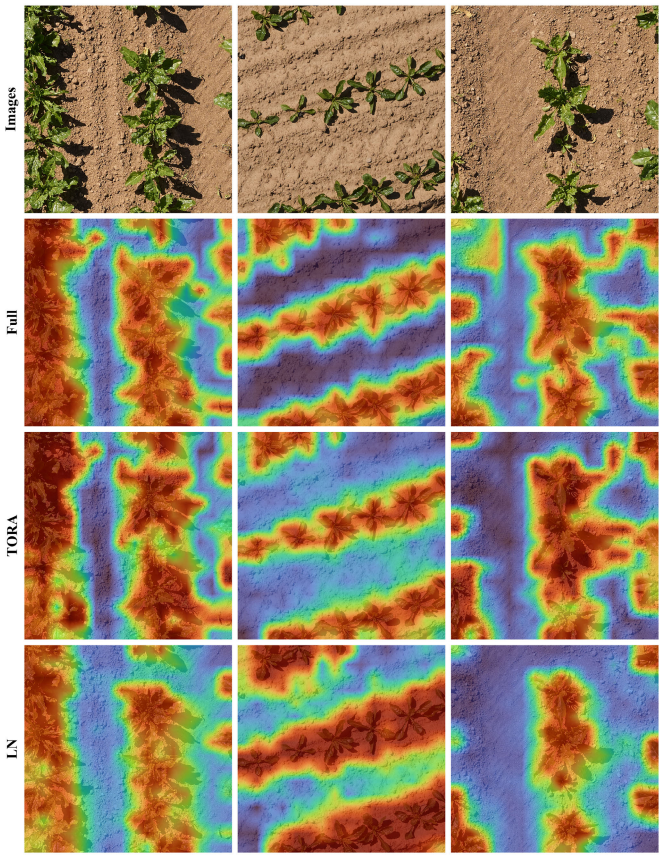

Fig. 7. Visualization of last-layer feature responses from SkySense fine-tuned with different PEFT methods. The first row shows the input images, and the subsequent rows present the corresponding feature maps produced by full fine-tuning, TORA, and LN-tuning, respectively (warmer colors indicate stronger responses).

D. Limitations and Future Work

While TORA achieves strong accuracy–efficiency tradeoffs across tasks and scales, our study is still limited to dataset-specific adaptation on mainly single-date optical imagery and two representative RSFM backbones. As analyzed in Section V-B, the parameter overhead of TORA is architecture-dependent. It is relatively compact on constant-width ViT backbones, but stagewise projections on hierarchical wide backbones such as the SwinV2-based SkySense can make it less parameter-economical than low-rank methods such as LoRA although it remains substantially smaller than full fine-tuning. Future work will investigate more lightweight designs for hierarchical architectures, extend TORA to multimodal and multitemporal inputs (e.g., multispectral time series and SAR), broaden evaluation to more diverse geographies, crop types, and acquisition conditions, and explore more realistic annotation settings for operational agricultural mapping.

VI. Conclusion

In this article, we presented TORA, a parameter-efficient adaptation framework for multisource and multiscale agricultural vision with frozen remote-sensing foundation models. TORA couples hierarchical semantic prototypes with a token-driven residual modulation pathway, aiming to enhance task-specific semantic calibration while retaining the spatial fidelity required by dense prediction. We conducted extensive evaluations across image classification, semantic segmentation, and object detection on diverse agricultural benchmarks to assess the generality and practicality of the proposed design. With trainable backbone-adaptation parameters amounting to approximately 1.5%–3.8% of the full-backbone parameter budget in our evaluated settings, TORA provides a compact and extensible adaptation mechanism relative to full

fine-tuning although its relative parameter economy among PEFT methods remains architecture-dependent. The empirical evidence supports the feasibility of prototype-driven residual adaptation as a general paradigm for transferring pretrained RSFM representations to heterogeneous agricultural environments under constrained training budgets. Despite these findings, this study primarily focuses on single-modality settings and the considered benchmark tasks and backbones. Future work will extend TORA to multimodal and multitemporal inputs (e.g., multispectral and time-series observations) and investigate lightweight deployment on resource-constrained platforms to better support practical precision agriculture.

REFERENCES

[1] *In Brief to the State of Food and Agriculture 2022: Leveraging Automation in Agriculture for Transforming Agrifood Systems*, Food and Agriculture Organization of the United Nations, Rome, Italy, 2022.

[2] S. Ghazal, A. Munir, and W. S. Qureshi, "Computer vision in smart agriculture and precision farming: Techniques and applications," *Artif. Intell. Agricult.*, vol. 13, pp. 64–83, Sep. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2589721724000266

[3] K. M. Murphy, E. Ludwig, J. Gutierrez, and M. A. Gehan, "Deep learning in image-based plant phenotyping," *Annu. Rev. Plant Biol.*, vol. 75, no. 1, pp. 771–795, Jul. 2024. [Online]. Available: https://www.annualreviews.org/content/journals/10.1146/annurev-arplant-070523-042828

[4] A. Upadhyay et al., "Deep learning and computer vision in plant disease detection: A comprehensive review of techniques, models, and trends in precision agriculture," *Artif. Intell. Rev.*, vol. 58, no. 3, Jan. 2025, Art. no. 92.

[5] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag. (replaces Newsletter)*, vol. 5, no. 4, pp. 8–36, Dec. 2017. [Online]. Available: https://ieeexplore.ieee.org/document/8113128/

[6] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271619301108

[7] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag., Replaces Newslett.*, vol. 4, no. 2, pp. 22–40, Jun. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7486259/

[8] S. Lu et al., "Vision foundation models in remote sensing: A survey," *IEEE Geosci. Remote Sens. Mag., Replaces Newslett.*, vol. 13, no. 3, pp. 190–215, Sep. 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10916803/

[9] X. Sun et al., "RingMo: A remote sensing foundation model with masked image modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5612822. [Online]. Available: https://ieeexplore.ieee.org/document/9844015/

[10] Y. Cong et al., "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," 2022, *arXiv:2207.08051*.

[11] X. Guo et al., "SkySense: A multi-modal remote sensing foundation model towards universal interpretation for Earth observation imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 27662–27673. [Online]. Available: https://ieeexplore.ieee.org/document/10655854/

[12] A. Joshi, B. Pradhan, S. Gite, and S. Chakraborty, "Remote-sensing data and deep-learning techniques in crop mapping and yield prediction: A systematic review," *Remote Sens.*, vol. 15, no. 8, p. 2014, Apr. 2023. [Online]. Available: https://www.mdpi.com/2072-4292/15/8/2014

[13] M. Shoaib, A. Sadeghi-Niaraki, F. Ali, I. Hussain, and S. Khalid, "Leveraging deep learning for plant disease and pest detection: A comprehensive review and future directions," *Frontiers Plant Sci.*, vol. 16, Feb. 2025, Art. no. 1538163. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpls.2025.1538163/full

[14] K. Zhan, Y. Peng, M. Liao, and Y. Wang, "Domain generalization plant leaf disease recognition: Toward from laboratory to field," *Eng. Appl. Artif. Intell.*, vol. 156, Sep. 2025, Art. no. 111168.

[15] H. Zhao et al., "A large-scale VHR parcel dataset and a novel hierarchical semantic boundary-guided network for agricultural parcel delineation," *ISPRS J. Photogramm. Remote Sens.*, vol. 221, pp. 1–19, Mar. 2025. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271625000395

[16] J. Zheng et al., "A comprehensive review of agricultural parcel and boundary delineation from remote sensing images: Recent progress and future perspectives," *IEEE Geosci. Remote Sens. Mag. (replaces Newsletter)*, early access, Feb. 13, 2026, doi: 10.1109/MGRS.2026.3658493.

[17] Q. Li et al., "Can large multimodal models understand agricultural scenes? benchmarking with AgroMind," in *Proc. NeurIPS Datasets Benchmarks Track*, 2025. [Online]. Available: https://openreview.net/forum?id=zZn2Kk5CLt

[18] H. Li et al., "AgroVG: A large-scale multi-source benchmark for agricultural visual grounding," 2026, *arXiv:2605.22034*.

[19] A. Alexopoulos et al., "Complementary use of ground-based proximal sensing and airborne/spaceborne remote sensing techniques in precision agriculture: A systematic review," *Agronomy*, vol. 13, no. 7, p. 1942, Jul. 2023. [Online]. Available: https://www.mdpi.com/2073-4395/13/7/1942

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2014, *arXiv:1411.4038*.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing Computer-Assisted Intervention—MICCAI*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241. [Online]. Available: http://link.springer.com/10.1007/978-3-319-24574-4_28

[22] M. Jia, "Visual prompt tuning," in *Computer Vision—ECCV*, vol. 13693. Cham, Switzerland: Springer, 2022, pp. 709–727. [Online]. Available: https://link.springer.com/10.1007/978-3-031-19827-4_41

[23] Z. Chen et al., "Vision transformer adapter for dense predictions," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Kigali, Rwanda, May 2023.

[24] M. Awais et al., "Foundation models defining a new era in vision: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 4, pp. 2245–2264, Apr. 2025, doi: 10.1109/TPAMI.2024.3506283.

[25] A. Xiao et al., "Foundation models for remote sensing and Earth observation: A survey," *IEEE Geosci. Remote Sens. Mag., Replaces Newslett.*, vol. 13, no. 4, pp. 297–324, Dec. 2025. [Online]. Available: https://ieeexplore.ieee.org/document/11097335/

[26] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 15979–15988. [Online]. Available: https://ieeexplore.ieee.org/document/9879206/

[27] Z. Xie et al., "Simmim: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9653–9663.

[28] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in *Proc. Int. Conf. Learn. Represent.*, 2022. [Online]. Available: https://openreview.net/forum?id=p-BhZSz59o4

[29] M. Noman et al., "Rethinking transformers pre-training for multi-spectral satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 27811–27819. [Online]. Available: https://ieeexplore.ieee.org/document/10657832/

[30] C. J. Reed et al., "Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4065–4076. [Online]. Available: https://ieeexplore.ieee.org/document/10377166/

[31] M. Tang, A. Cozma, K. Georgiou, and H. Qi, "Cross-scale MAE: A tale of multi-scale exploitation in remote sensing," 2024, *arXiv:2401.15855*.

[32] F. Wang et al., "Harnessing massive satellite imagery with efficient masked image modeling," 2024, *arXiv:2406.11933*.

[33] D. Wang et al., "SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model," in *Proc. 37th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2023. [Online]. Available: https://openreview.net/forum?id=jHrgq55ftl

[34] D. Wang et al., "MTP: Advancing remote sensing foundation model via multitask pretraining," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 11632–11654, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10547536/

[35] K. Wu et al., "A semantic-enhanced multi-modal remote sensing foundation model for Earth observation," *Nature Mach. Intell.*, vol. 7, no. 8, pp. 1235–1249, Aug. 2025. [Online]. Available: https://www.nature.com/articles/s42256-025-01078-8

[36] W. Li et al., "AgriFM: A multi-source temporal remote sensing foundation model for agriculture mapping," 2025, *arXiv:2505.21357*.

[37] N. Ding et al., "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Mach. Intell.*, vol. 5, no. 3, pp. 220–235, Mar. 2023. [Online]. Available: https://www.nature.com/articles/s42256-023-00626-4

[38] L. Wang et al., "Parameter-efficient fine-tuning in large language models: A survey of methodologies," *Artif. Intell. Rev.*, vol. 58, no. 8, p. 227, May 2025. [Online]. Available: https://link.springer.com/10.1007/s10462-025-11236-4

[39] E. Ben Zaken, Y. Goldberg, and S. Ravfogel, "BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 1–9. [Online]. Available: https://aclanthology.org/2022.acl-short.1

[40] S. Basu, S. Hu, D. Massiceti, and S. Feizi, "Strong baselines for parameter-efficient few-shot fine-tuning," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 10, pp. 11024–11031.

[41] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, pp. 2790–2799. [Online]. Available: https://proceedings.mlr.press/v97/houlsby19a.html

[42] S. Chen et al., "AdaptFormer: Adapting vision transformers for scalable visual recognition," 2022, *arXiv:2205.13535*.

[43] M. Liao et al., "Multi-granularity token learning for domain generalized semantic segmentation," *IEEE Trans. Multimedia*, no. 99, pp. 1–14, 2026, doi: 10.1109/TMM.2026.3678041.

[44] E. Hu et al., "LoRA: Low-rank adaptation of large language models," Tech. Rep., 2022.

[45] S.-Y. Liu et al., "DoRA: Weight-decomposed low-rank adaptation," in *Proc. 41st Int. Conf. Mach. Learn.*, vol. 235, pp. 32100–32121. [Online]. Available: https://proceedings.mlr.press/v235/liu24bn.html

[46] H. Liu et al., "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 1950–1965. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/0cde695b83bd186c1fd456302888454c-Paper-Conference.pdf

[47] D. Yin et al., "Remote sensing tuning: A survey," *Comput. Vis. Media*, vol. 11, no. 5, pp. 897–937, Oct. 2025, doi: 10.26599/CVM.2025.9450490.

[48] J. Zheng et al., "Open-set domain adaptation for scene classification using multi-adversarial learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 208, pp. 245–260, Feb. 2024, doi: 10.1016/j.isprsjprs.2024.01.015.

[49] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271616300144

[50] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114417. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420310836

[51] J. Zheng, S. Yuan, W. Li, H. Fu, L. Yu, and J. Huang, "A review of individual tree crown detection and delineation from optical remote sensing images: Current progress and future," *IEEE Geosci. Remote Sens. Mag. (replaces Newsletter)*, vol. 13, no. 1, pp. 209–236, Mar. 2025, doi: 10.1109/MGRS.2024.3479871.

[52] W. Liu, X. Shen, C.-M. Pun, and X. Cun, "Explicit visual prompting for low-level structure segmentations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19434–19445. [Online]. Available: https://ieeexplore.ieee.org/document/10203162/

[53] X. Ma, X. Zhang, X. Ding, M.-O. Pun, and S. Ma, "Decomposition-based unsupervised domain adaptation for remote sensing image semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5645118.

[54] Anonymous. (2024). *CVPT: Cross-Attention Help Visual Prompt Tuning Adapt Visual Task*. [Online]. Available: https://openreview.net/forum?id=pNQB78UDp4

[55] L. Ren, C. Chen, L. Wang, and K. Hua, "DA-VPT: Semantic-guided visual prompt tuning for vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 4353–4363.

[56] L. Hu, H. Yu, W. Lu, D. Yin, X. Sun, and K. Fu, "AiRs: Adapter in remote sensing for parameter-efficient transfer learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5605218. [Online]. Available: https://ieeexplore.ieee.org/document/10385180/

[57] L. Hu, W. Lu, H. Yu, D. Yin, X. Sun, and K. Fu, "TEA: A training-efficient adapting framework for tuning foundation models in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5648118. [Online]. Available: https://ieeexplore.ieee.org/document/10740309/

[58] X. Ma, X. Zhang, M.-O. Pun, and B. Huang, "A unified framework with multimodal fine-tuning for remote sensing semantic segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5405015.

[59] D. Zhang, F. Wang, L. Ning, Z. Zhao, J. Gao, and X. Li, "Integrating SAM with feature interaction for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4513011. [Online]. Available: https://ieeexplore.ieee.org/document/10737290/

[60] J. Gao, D. Zhang, F. Wang, L. Ning, Z. Zhao, and X. Li, "Combining SAM with limited data for change detection in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, 2025, Art. no. 5614311. [Online]. Available: https://ieeexplore.ieee.org/document/10902491/

[61] X. Ma, Q. Wu, X. Zhao, X. Zhang, M.-O. Pun, and B. Huang, "SAM-assisted remote sensing imagery semantic segmentation with object and boundary constraints," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5636916.

[62] J. Weyler et al., "PhenoBench - a large dataset and benchmarks for semantic image interpretation in the agricultural domain," 2023, *arXiv:2306.04557*.

[63] J. Zhou et al., "Global Rice multiclass segmentation dataset (RiceSEG): Comprehensive and diverse high-resolution RGB-annotated images for the development and benchmarking of Rice segmentation algorithms," *Plant Phenomics*, vol. 7, no. 3, Sep. 2025, Art. no. 100099. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2643651525001050

[64] E. David et al., "Global wheat head detection 2021: An improved dataset for benchmarking wheat head detection methods," *Plant Phenomics*, vol. 2021, Sep. 2021, Art. no. 9846158.

[65] N. Häni, P. Roy, and V. Isler, "MinneApple: A benchmark dataset for apple detection and segmentation," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 852–858, Apr. 2020.

[66] J. Zheng et al., "Cross-regional oil palm tree counting and detection via a multi-level attention domain adaptation network," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 154–177, Sep. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0924271620301830

[67] A. Olsen et al., "DeepWeeds: A multiclass weed species image dataset for deep learning," *Sci. Rep.*, vol. 9, no. 1, p. 2058, Feb. 2019. [Online]. Available: https://www.nature.com/articles/s41598-018-38343-3

[68] *New Plant Diseases Dataset*. Accessed: Dec. 22, 2025. [Online]. Available: https://www.kaggle.com/datasets/vipoooool/new-plant-diseases-dataset

[69] D. P. Hughes and M. Salathè, "An open access repository of images on plant health to enable the development of mobile disease diagnostics," 2015, *arXiv:1511.08060*.

[70] T. Ilyas et al., "CWD30: A new benchmark dataset for crop weed recognition in precision agriculture," *Comput. Electron. Agricult.*, vol. 229, Feb. 2025, Art. no. 109737. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0168169924011281