agronomy

MDPI

Article

Parameter-Efficient Domain Adaptation and Lightweight Decoding for Agricultural Monocular Depth Estimation

Yanliang Mao [1,2], Wenhao Zhao [1,2] and Liping Chen [1,2,3,*]

[1] College of Information Engineering, Tarim University, Alaer 843300, China; 10757242327@stumail.taru.edu.cn (Y.M.); 10757251216@stumail.taru.edu.cn (W.Z.)
[2] Key Laboratory of Tarim Oasis Agriculture, Tarim University, Ministry of Education, Alaer 843300, China
[3] Key Laboratory of Modern Agricultural Engineering, Tarim University, Alaer 843300, China
* Correspondence: 119970011@taru.edu.cn

Abstract

Reliable monocular depth estimation (MDE) is essential for agricultural robots and unmanned platforms, where low-cost visual perception is required for safe navigation and scene understanding in complex field environments. However, general-purpose depth foundation models remain limited by substantial domain gaps in agriculture, while full fine-tuning of large backbones is computationally expensive and less suitable for deployment on resource-constrained platforms. In this paper, an efficient agricultural MDE framework, termed AgriLoRA-DA, is proposed based on Depth-Anything-V2. Specifically, the pretrained DINOv2 encoder is kept frozen and adapted using LoRA in selected attention projections, while the original Dense Prediction Transformer (DPT) decoder is replaced with a lightweight Lite-FPNHead to reduce decoding overhead and improve deployment efficiency. Experiments conducted on the WE3DS dataset indicate that, although Depth-Anything-V3 provides the strongest zero-shot generalization among the evaluated baselines, target-domain adaptation is still necessary for WE3DS agricultural scenes. After adaptation, AgriLoRA-DA achieves the best overall performance with AbsRel = 0.0133, SqRel = 3.518, RMSE = 132.264, $log10$ = 0.0057, and delta1 = 0.9990, while requiring only 0.19 M (0.87%) trainable parameters. These results suggest that parameter-efficient adaptation and lightweight decoding provide a practical direction for deployable depth estimation in crop-row scenes similar to WE3DS, while broader cross-dataset validation remains an important direction for future work.

Keywords: agricultural monocular depth estimation; domain adaptation; LoRA; lightweight decoder; Depth-Anything-V2

Academic Editors: Jayme Garcia Arnal Barbedo and Mulham Fawakherji

Received: 30 March 2026
Revised: 10 May 2026
Accepted: 12 May 2026
Published: 13 May 2026

1. Introduction

Smart agriculture is rapidly adopting autonomous robots and unmanned platforms for tasks such as navigation, obstacle avoidance, inter-row operations, and field monitoring. Reliable three-dimensional perception is a prerequisite for the safe and efficient operation of these systems in unstructured agricultural environments, where uneven terrain, plant occlusion, and large variations in crop geometry introduce substantial uncertainty. Among available perception solutions, vision-based approaches are particularly attractive for agricultural robotics because of their low cost, ease of integration, and rich semantic information, and they have become a key sensing modality for navigation and scene understanding in agricultural systems [1]. In 3D perception, depth information is typically

acquired using stereo cameras, active RGB-D sensors, or LiDAR. However, these solutions often impose non-negligible hardware and operational burdens in agricultural deployment: stereo methods are sensitive to low-texture regions, repetitive patterns, and calibration drift; active depth sensors may degrade under strong sunlight and specular reflections from foliage; and LiDAR generally increases system cost, power consumption, and mechanical complexity. By contrast, MDE infers scene depth from a single RGB image, enabling low-cost, lightweight, and energy-efficient deployment using standard cameras that are already widely available on agricultural platforms. Recent reviews in agricultural and food applications have likewise identified depth estimation from monocular imagery as an important direction in 3D vision research [2].

Beyond agriculture, MDE has become a fundamental capability in robotics, autonomous driving, augmented and virtual reality, 3D reconstruction, and scene understanding. Recent surveys have shown that, with the growth of model and dataset scales, learning-based depth estimation has achieved substantial gains in robustness and cross-dataset generalization, while its range of applications continues to expand [3]. From a technical perspective, modern MDE benefits from large-scale pretraining and increasingly mature training recipes. For example, mixing heterogeneous datasets and adopting scale-invariant objectives can markedly improve zero-shot cross-dataset transfer, highlighting that data diversity and training strategy are at least as important as network architecture design [4]. Architecturally, Transformer-based dense prediction models, such as those built upon DPT-style decoders [5], exploit global receptive fields to produce more spatially coherent geometry, whereas adaptive discretization methods such as AdaBins [6] model depth distributions through learnable binning and thereby improve metric depth regression.

Despite these advances, agricultural environments remain particularly challenging for general-purpose depth models. Field scenes are characterized by extreme illumination changes, including strong backlighting and canopy shadows, frequent specular reflections from leaf surfaces, complex occlusions caused by stems and overlapping leaves, and pronounced appearance variation across growth stages and cultivars. These factors create a substantial domain gap relative to the distributions represented in mainstream depth benchmarks, which are dominated by urban, indoor, or internet-scale imagery. Consequently, models that generalize well in generic settings may still exhibit unstable depth boundaries, geometrically inconsistent predictions on thin plant structures, and biased depth scales between crop rows when deployed in agricultural environments. Such limitations can directly impair downstream tasks, including robot control, safe-distance maintenance, and physical interaction with plants.

More specifically, existing monocular depth estimation studies for agriculture often require substantial additional effort in data collection, annotation, or multi-module system design tailored to particular crops or operational objectives. Zheng et al. [7] introduced the OrchardDepth dataset for orchard and vineyard environments and retrained a model with consistency regularization between sparse points and dense depth to obtain metric predictions; such a pipeline typically entails dedicated scene-specific data acquisition and repeated retraining. Cui et al. [8] proposed MonoDA, a self-supervised framework that jointly learns depth and pose estimation, thereby reducing the need for explicit depth labels, but still requiring video sequences, geometric consistency constraints, and careful tuning for training stability. For fine-grained agricultural manipulation such as harvesting, Coll-Ribes et al. [9] coupled grape and peduncle detection with depth estimation for robotic operation, a strategy that generally depends on a multi-stage perception pipeline and task-specific structural design. More recently, Zhao et al. [10] incorporated a monocular depth module into an agricultural vision-and-language navigation framework to support decision-making, but the resulting multimodal reasoning and system integration further

increase both training and deployment complexity. Overall, although these methods are effective under their respective settings, they are often characterized by strong scenario dependence and heavy engineering overhead, and may still suffer from scale drift and boundary distortion when transferred across farms, seasons, or sensor configurations.

A notable recent trend is the emergence of foundation depth models. Trained via large-scale data engines and teacher-student distillation, these models aim to provide a unified depth backbone that operates effectively across diverse open-domain imagery. Depth Anything [11] demonstrated that scaling unlabeled data through automatic annotation pipelines can substantially improve the robustness and zero-shot generalization of MDE. Depth Anything V2 [12] further advanced this direction through improved training practices, producing finer-grained and more stable predictions while maintaining high efficiency. Depth Anything 3 [13] extends this paradigm to broader any-view geometric understanding. This trajectory parallels the broader shift in computer vision toward large, promptable foundation models, as exemplified by SAM [14], SAM 2 [15], and SAM 3 [16], which progressively generalize unified pretrained backbones into versatile visual interfaces. Nevertheless, strong zero-shot generalization does not necessarily guarantee optimal performance in a specific domain. When the target distribution deviates substantially from the pretraining data, as in complex agricultural canopies under highly variable illumination, direct deployment may remain suboptimal; meanwhile, naïve full fine-tuning of modern large backbones is often prohibitively expensive.

To bridge this gap, parameter-efficient fine-tuning offers a practical trade-off between task adaptation performance and training cost. Existing transfer strategies include adapters, bias-only tuning, prompt-based tuning, partial fine-tuning, and low-rank adaptation. Among them, LoRA [17] injects small trainable low-rank matrices into frozen pretrained weights, thereby greatly reducing the number of trainable parameters and enabling efficient domain adaptation under limited data and computational budgets. This study does not claim that LoRA is universally superior to all parameter-efficient transfer methods; rather, it evaluates LoRA as a simple and effective adaptation mechanism for agricultural monocular depth estimation on WE3DS. This efficiency is particularly valuable in agricultural applications, where high-quality depth annotations are costly to acquire and deployment often requires real-time inference on edge devices. In addition to training efficiency, inference efficiency is equally important: agricultural robots typically operate under strict latency and power constraints, and even when the backbone is reused, an overly heavy decoder can still become a deployment bottleneck.

Accordingly, in this paper, we propose an efficient adaptation framework for agricultural monocular depth estimation. Building upon the V2 version of a foundation depth model, we perform parameter-efficient domain adaptation and further redesign the depth decoder for lightweight real-world deployment. The proposed decoder reduces the number of parameters in the original decoder by approximately 95% while preserving estimation quality, thereby substantially improving inference speed and on-board deployment practicality.

The main contributions of this work are summarized as follows:

(1) Parameter-efficient domain adaptation of a foundation depth model for agriculture. We propose a parameter-efficient fine-tuning strategy that adapts a foundation depth model to complex agricultural environments with lower data and training overhead.
(2) Lightweight decoding for deployment-oriented depth estimation. We design a more compact decoder that significantly reduces parameter count (approximately 95%) and improves inference efficiency while maintaining strong depth prediction performance.
(3) Comprehensive evaluation in a representative agricultural benchmark. We validate the proposed method on the WE3DS datasets and compare it with several zero-shot

and target-domain adaptation baselines. The results demonstrate strong performance on this benchmark while also highlighting the need for future cross-dataset validation before making broader claims across all agricultural environments.

The remainder of this paper is organized as follows. Section 2 presents the proposed AgriLoRA-DA framework, including the dataset, baseline model, and methodological details. Section 3 reports the experimental settings and results. Section 4 discusses the findings and limitations. Section 5 concludes the paper.

2. Materials and Methods

2.1. Dataset

As shown in Figure 1, we adopt the publicly available WE3DS [18] dataset as the target-domain benchmark for monocular depth estimation in agricultural scenes. The dataset contains 2568 RGB-D samples, including 1540 training images and 1028 testing images, together with refined depth annotations and valid-pixel masks. The metadata indicate that image acquisition covered multiple field dates in 2020 and 2021 under sunny, cloudy, and mixed weather conditions, with light, medium, and strong wind levels. Recorded crop heights range from 766 mm to 1018 mm, indicating that the dataset includes substantial variation in canopy structure and growth stage. Owing to the intrinsic complexity of agricultural imagery, WE3DS captures several challenges commonly encountered in real-world applications, including severe leaf overlap, thin structural boundaries, self-occlusion, locally high-reflectance regions, and illumination fluctuations. To ensure strict reproducibility and enable fair comparison with previous studies, we follow the official data split provided by the dataset. Because this study uses the released WE3DS benchmark, camera calibration, RGB-D alignment, refined depth generation, and valid-mask construction are inherited from the official dataset protocol [18]. No additional physical camera calibration was performed in this work. During training and evaluation, we directly use the provided RGB images, re-fined depth maps, and valid masks, and all metric computations are restricted to valid pixels.

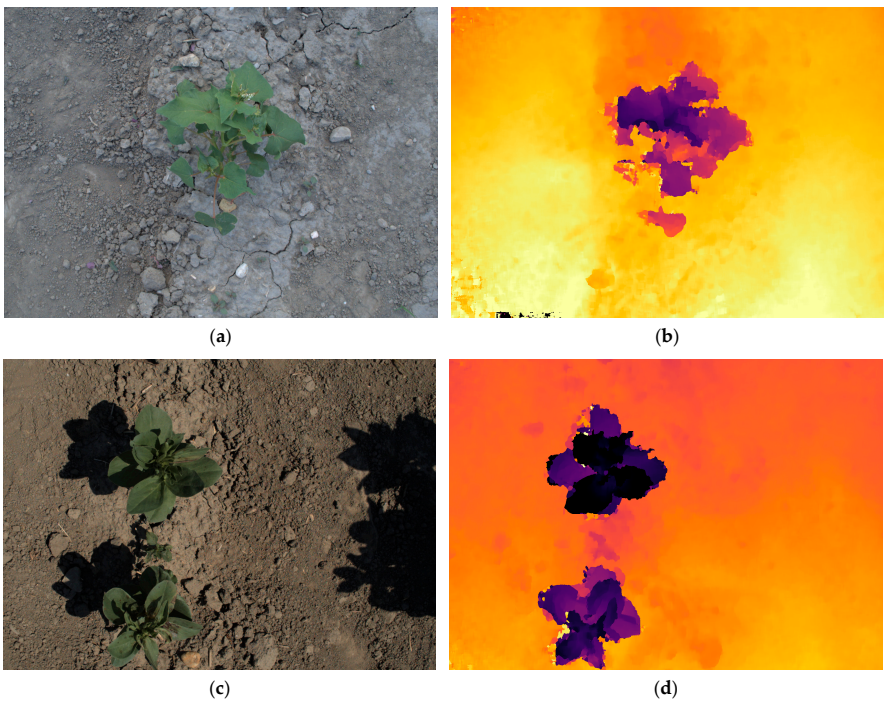

Figure 1. Visualization of the WE3DS dataset. Panels (**a**,**c**) depict two representative RGB images, whereas panels (**b**,**d**) show the corresponding refined depth annotations, respectively.

In practical image acquisition, the influence of environmental conditions can be reduced by maintaining stable camera mounting, using consistent exposure settings where possible, avoiding severe motion blur, periodically checking camera calibration, and collecting representative samples under different weather and illumination conditions for target-domain adaptation. In the present experiments, robustness to such variations is supported by the WE3DS metadata diversity, valid-mask filtering, ImageNet normalization, and supervised adaptation on the target-domain training split.

In our implementation, the depth_refined annotations in WE3DS are used as supervisory signals during training. Both loss computation and quantitative evaluation are restricted to the valid-mask regions only, thereby excluding unreliable or undefined depth values from optimization and assessment. Before being fed into the network, the input RGB images are resized to 518 × 518, with spatial dimensions constrained to be divisible by 14 so as to match the patch tokenization scheme of the ViT encoder. During preprocessing, all images are further normalized using the ImageNet mean and standard deviation, ensuring consistency with the pretraining distribution of the backbone model.

2.2. Baseline Model: Depth-Anything-V2

To establish a representative and transferable baseline for MDE in agricultural environments, this study adopts Depth-Anything-V2 (DA-V2) as the starting point. As shown in Figure 2, DA-V2 follows an encoder–decoder paradigm, where a DINOv2-based [19] vision transformer serves as the encoder and a DPT-style depth head is used to reconstruct dense geometric predictions from multi-level token representations. Under the ViT-S configuration used in this work, intermediate features are extracted from the outputs of the 2nd, 5th, 8th, and 11th transformer blocks, providing a hierarchy of semantic and structural cues at different depths of the encoder. In implementation, the model computes the patch-grid resolution from the input image and retrieves four intermediate token groups together with the class token when needed.

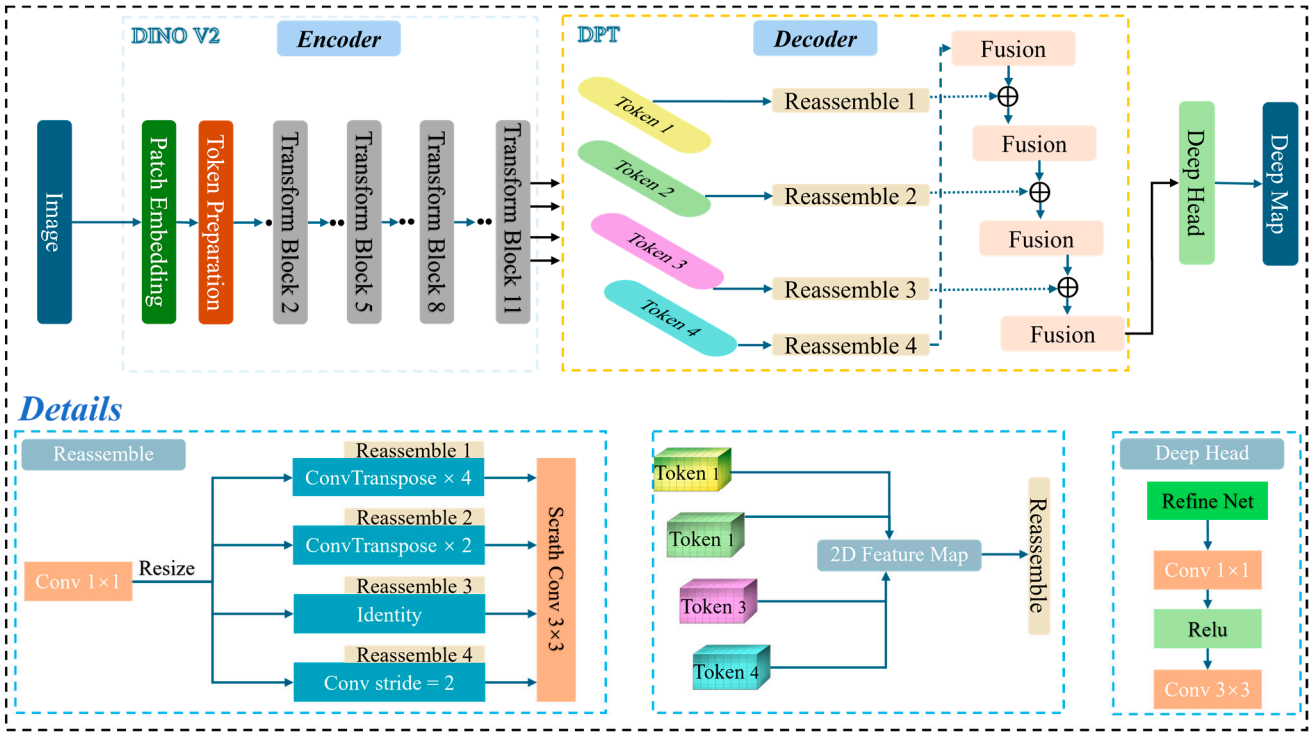

Figure 2. Schematic illustration of the Depth-Anything-V2 architecture. When ViT-S is adopted as the encoder, the intermediate feature tokens, namely Token 1, Token 2, Token 3, and Token 4, are extracted from the outputs of the 2nd, 5th, 8th, and 11th Transformer blocks, respectively.

The original DPT decoder first converts each token sequence into a spatial feature map by rearranging the token dimension according to the patch grid. These features are then projected by 1×1 convolutions and explicitly aligned across scales through a multi-branch reassembly design, including $\times 4$ transposed convolution, $\times 2$ transposed convolution, identity mapping, and stride-2 convolution. After scale-specific reassembly, the decoder applies a scratch network followed by four progressive fusion blocks to aggregate information in a coarse-to-fine manner. Finally, the fused feature is refined by the output head and upsampled to produce a single-channel dense depth map. This design preserves both high-level semantics and fine-grained spatial structures, making DA-V2 a strong and reliable baseline for dense geometric perception.

The appeal of this baseline lies in its effective combination of foundation-model priors and structured dense prediction. The DINOv2 encoder contributes robust global contextual modeling learned from large-scale pretraining, while the DPT decoder restores the spatial organization required for depth estimation through hierarchical feature reconstruction. Such a combination is particularly relevant to agricultural vision, where depth cues are often distributed across both broad scene layout and subtle local structures, such as crop stems, irrigation pipes, trellis wires, and partially occluded plant organs. Nevertheless, despite its strong representational capability, a substantial portion of the model complexity is concentrated in the decoder, especially in the multi-branch reassembly and progressive fusion stages. This motivates a more deployment-oriented redesign in the subsequent section.

2.3. AgriLoRA-DA

Building upon the baseline described above, we propose AgriLoRA-DA, a domain-oriented and efficiency-aware adaptation framework for monocular depth estimation in agricultural scenarios. As illustrated in Figure 3, the core idea is not to uniformly modify the entire network, but to intervene only where adaptation and efficiency matter most. Specifically, the encoder retains the strong representational backbone of DINOv2, while the decoder is redesigned to reduce inference overhead. In this way, the proposed framework balances three practical requirements that frequently arise in agricultural deployment: reliable transfer from large-scale visual pretraining, efficient adaptation to domain-specific data, and lightweight dense prediction under resource constraints.

The rationale behind this design is twofold. On the one hand, agricultural scenes exhibit substantial domain-specific characteristics, including repetitive crop textures, irregular plant geometry, variable illumination, severe self-occlusion, and cluttered backgrounds. A practical agricultural depth model therefore requires task-aware adaptation beyond generic visual representations. On the other hand, the decoder of the original DA-V2 baseline contributes non-negligible computational and structural overhead due to its explicit multi-scale reassembly and refinement pipeline. Consequently, AgriLoRA-DA introduces a two-branch improvement strategy at the framework level: the encoder is kept frozen and adapted in a parameter-efficient manner, while the decoder is replaced with a lightweight alternative tailored for dense depth reconstruction. The former preserves the transferable priors of the pretrained vision transformer, whereas the latter improves inference efficiency and deployment flexibility. The detailed formulation of the encoder-side adaptation will be presented in Section 2.5.

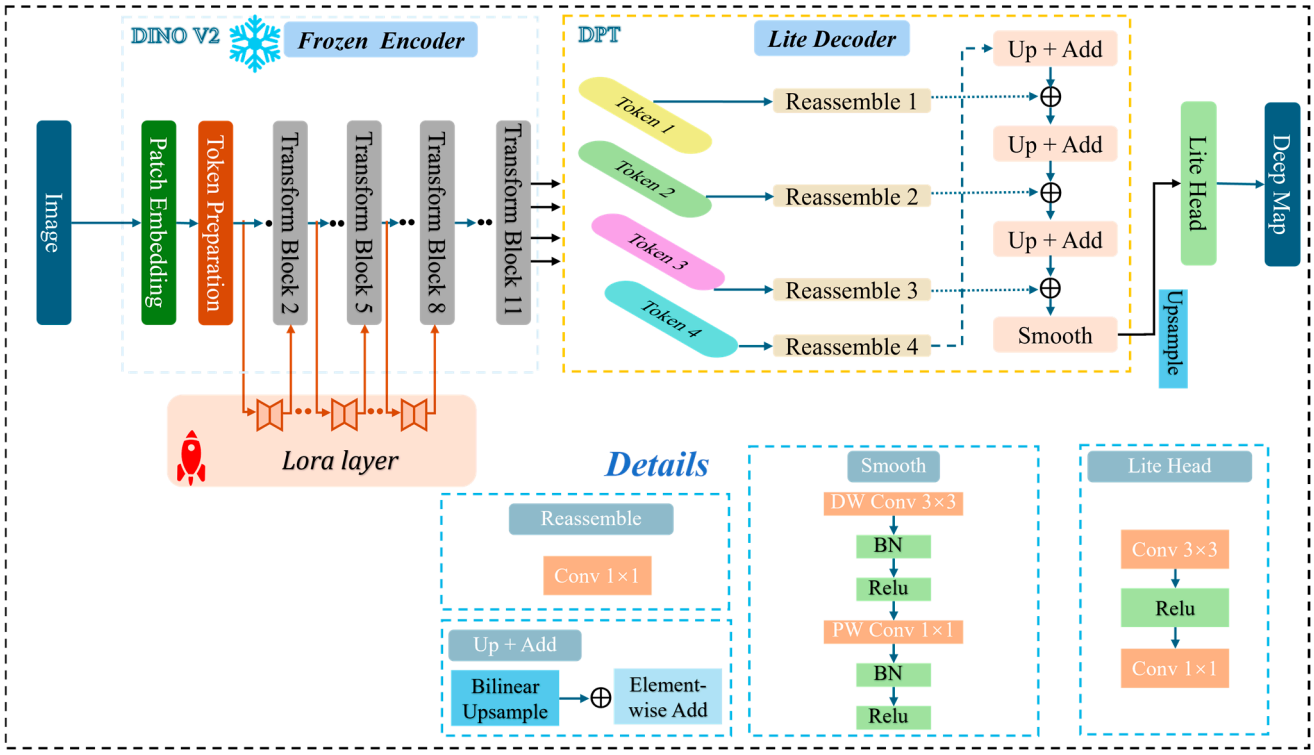

Figure 3. Schematic illustration of the AgriLoRA-DA architecture. The proposed framework freezes the DINOv2 encoder and introduces LoRA-based adaptation in selected transformer blocks, while replacing the original DPT decoder with a lightweight Lite-FPNHead for efficient dense depth prediction in agricultural scenes.

2.4. Lite-FPNHead

To reduce the decoding burden of the original DPT head while retaining effective multi-scale depth reconstruction, we design a lightweight decoder termed Lite-FPNHead, as shown in Figure 4. The redesign starts from a simple observation: although the original DPT decoder is effective, its complexity is heavily tied to explicit multi-branch reassembly and progressive fusion blocks. In practice, these components introduce substantial parameter and computation overhead, which is undesirable for agricultural deployment scenarios requiring efficiency, robustness, and ease of adaptation. Therefore, instead of reproducing the full DPT decoding pathway, we reformulate the decoder as a compact feature pyramid that preserves the essential cross-scale aggregation mechanism while removing structurally expensive operations.

The first step of Lite-FPNHead is to convert transformer tokens into 2D feature maps. For each intermediate layer, the token sequence is first rearranged by permuting the channel dimension and then reshaped according to the patch grid, yielding a spatial feature representation. Compared with the original DPT decoder, which immediately couples token-to-map conversion with scale-specific reassembly branches, the proposed Lite-FPNHead keeps this stage minimal and regularized. After token-to-map transformation, each feature map is projected by a 1×1 convolution into a unified channel dimension. This channel unification is crucial because it enables subsequent cross-scale aggregation through simple addition rather than heavy block-wise fusion.

The second step is a top-down feature pyramid aggregation. Let the projected features be denoted as {layer1, layer2, layer3, layer4}, where deeper layers contain stronger semantic

abstraction and shallower layers preserve finer spatial detail. The decoder first takes the deepest feature as p_4, and then progressively constructs:

$$p3 = layer3 + Up(p4), p2 = layer2 + Up(p3), p1 = layer1 + Up(p2),$$

where $Up(\cdot)$ denotes bilinear upsampling to the spatial size of the current layer. In contrast to the original DPT decoder, which relies on feature fusion blocks after explicit scale reconstruction, this design performs cross-scale interaction through bilinear upsampling and element-wise addition. Such a formulation preserves the top-down semantic injection characteristic of feature pyramids while substantially simplifying the decoder pipeline.

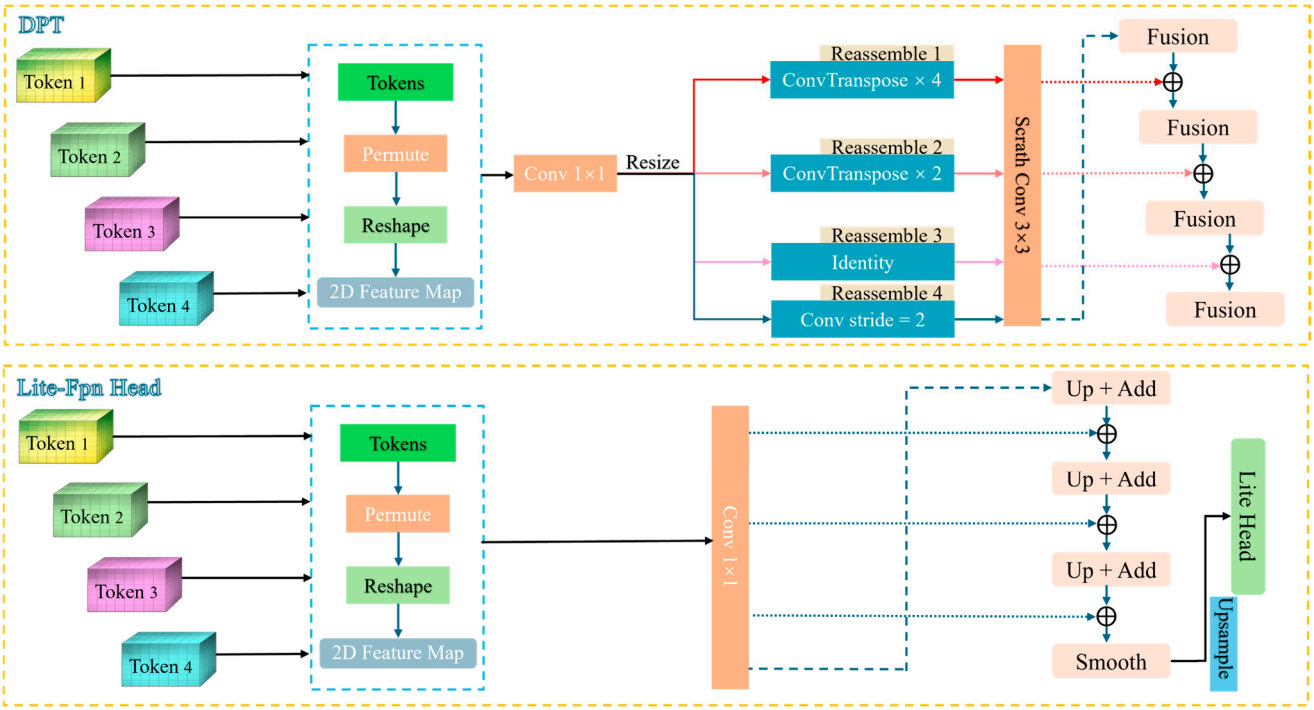

Figure 4. Comparison between the DPT and Lite-FPNHead architectures.

After hierarchical aggregation, the fused high-resolution feature p1 is refined by a lightweight smooth module composed of a depthwise 3×3 convolution, normalization, Rectified Linear Unit (ReLU) activation, and a pointwise 1×1 convolution followed again by normalization and ReLU. This module serves as a compact local refinement stage that enhances the consistency of the fused representation at low computational cost. The refined feature is then bilinearly upsampled to the image-scale patch resolution and passed to a compact output head consisting of a 3×3 convolution, ReLU, and a final 1×1 convolution to predict the depth map. Compared with the original DPT output head and its preceding refinement blocks, the proposed design is considerably shorter and structurally cleaner, yet remains fully compatible with the multi-level features extracted by the transformer encoder.

The proposed Lite-FPNHead offers three advantages. First, it replaces heterogeneous scale reconstruction operators with a unified projection-and-aggregation pipeline, which simplifies optimization and reduces implementation complexity. Second, it retains multi-scale interaction through a principled top-down pathway, ensuring that deep semantic cues can still guide shallow spatial features. Third, the use of depthwise separable smoothing and a compact output head makes the decoder better suited to real-world agricultural systems with limited computation budgets. Taken together, these properties make LiteF-PNHead not merely a simplified substitute for the original DPT decoder, but a task-driven

redesign that better aligns dense depth estimation with the deployment constraints of agricultural vision.

2.5. Integration of LoRA into Attention Mechanisms

The attention mechanism is a cornerstone of modern vision and multimodal models. However, its large number of parameters and high computational cost often become bottlenecks for model transfer and downstream task adaptation. Traditional fine-tuning re-quires updating the entire attention weight matrix, which is computationally expensive and prone to overfitting when data are limited. To address this challenge, we incorporate the low-rank adaptation method (LoRA) into the attention module. The basic idea of LoRA is to approximate the update ΔW of the original weights by multiplying two low-rank matrices, thereby significantly reducing the number of trainable parameters while adaptation capability. Its approximation can be expressed as Equations (1) and (2):

$$\hat{W} = W + \Delta W \tag{1}$$

$$\Delta W \approx sBA, s = \frac{\alpha}{r} \tag{2}$$

In:

- $W \in R^{d \times k}$ are the pre-trained weights;
- $\Delta W \in R^{d \times k}$ is the weight updated for a specific task;
- $B \in R^{d \times r}$, $A \in R^{r \times k}$ and $r \ll \min(d,k)$.
- s is a commonly used scaling factor (decoupling the magnitude of low-rank updates from the rank).

Since the updates of pre-trained weights are stable, a low-rank decomposition technique is proposed to achieve this update. Therefore, to achieve low-rank adaptation, we first freeze the image encoder to keep W fixed, training only ΔW for backpropagation. Because the feature extraction capability of the image encoder comes from its multi-head self-attention mechanism, to improve performance for specific tasks, we add a bypass structure to its projection layer. As shown in Equation (3), this bypass structure consists of three linear layers, and the input features $X \in R^{N \times d_{\text{in}}}$ are obtained after passing through this weight matrix:

$$\begin{aligned} Q' &= X(W_Q + s_Q B_Q A_Q) = Q + s_Q(XB_Q)A_Q \\ K' &= X(W_K + s_K B_K A_K) = K + s_K(XB_K)A_K \\ V' &= X(W_V + s_V B_V A_V) = V + s_V(XB_V)A_V \end{aligned} \tag{3}$$

In AgriLoRA-DA, the schematic diagram of LoRA in Attention is shown in Figure 5 below:

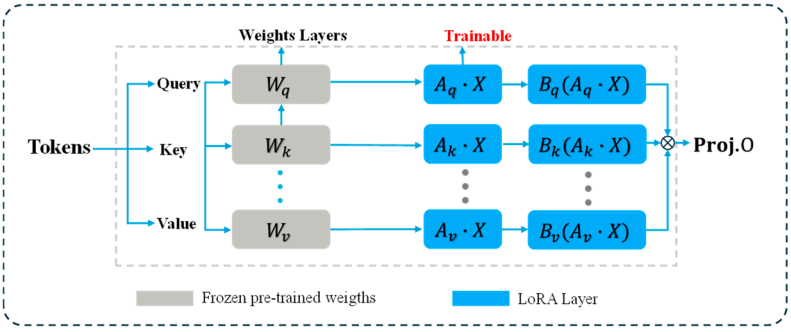

Figure 5. LoRA in attention.

As can be seen, LoRA is equivalent to adding a bottleneck bypass after the original linear layer, first mapping the input X to the low-rank subspace $XB \in R^{N\times r}$, and then pulling it back to the head dimension via $A \in R^{r\times d_h}$. Due to the core multi-head attention mechanism in ViT-Block, which uses cosine similarity for calculation, the original multi-head attention mechanism can be expressed as Equation (4):

$$\mathrm{Att}(Q,K,V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right)V \tag{4}$$

Therefore, the attention scoring matrix ratio after the LoRA linear layer becomes Equation (5):

$$S' = \frac{Q'K'^{\top}}{\sqrt{d_h}} = \frac{(Q+\Delta Q)(K+\Delta K)^{\top}}{\sqrt{d_h}} = \frac{QK^{\top}}{\sqrt{d_h}} + \frac{Q(\Delta K)^{\top} + (\Delta Q)K^{\top} + (\Delta Q)(\Delta K)^{\top}}{\sqrt{d_h}} \tag{5}$$

When r is small but the scaling factor s is reasonable, $(\Delta Q)(\Delta K)^{\top}$ is often a second-order small quantity, which can be ignored during reanalysis. Therefore, the attention mechanism after LoRA fine-tuning is defined as shown in Equation (6):

$$S' \approx \frac{QK^{\top}}{\sqrt{d_h}} + \frac{QA_K(XB_K)^{\top} + (XB_Q)A_QK^{\top}}{\sqrt{d_h}} \tag{6}$$

This demonstrates that LoRA fine-tunes the attention score through a learnable low-rank correction term, thereby changing "who to focus on" with almost no increase in computation/memory. Furthermore, by adding all three projection layers to a trainable LoRA low-rank matrix, the number of new trainable parameters per head is significantly less than that of full fine-tuning ($r(d_{\mathrm{in}} + d_h) \times 3 \ll 3d_{\mathrm{in}}d_h$). Meanwhile, the complexity of forward propagation remains almost unchanged (the main time consumption is still in $QK^{\top}$ and Softmax), because the multiplication of XB and A is low-rank.

2.6. Performance Metrics

To comprehensively evaluate the performance of the proposed model on the monocular depth estimation task, this study adopts five widely used quantitative metrics, including Absolute Relative Error (AbsRel), Squared Relative Error (SqRel), Root Mean Squared Error (RMSE), $log10$ error, and threshold accuracy (δk). Let d_i and $\hat{d}_i$ denote the ground-truth depth and the predicted depth at the i-th valid pixel, respectively, and let N denote the total number of valid pixels used for evaluation. These metrics are defined as follows:

$$\mathrm{AbsRel} = \frac{1}{N}\sum_i \frac{\left|d_i - \hat{d}_i\right|}{d_i} \tag{7}$$

$$\mathrm{SqRel} = \frac{1}{N}\sum_i \frac{\left(d_i - \hat{d}_i\right)^2}{d_i} \tag{8}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_i \left(d_i - \hat{d}_i\right)^2} \tag{9}$$

$$log10 = \frac{1}{N}\sum_i \left|log_{10}(d_i) - log_{10}\left(\hat{d}_i\right)\right| \tag{10}$$

$$\delta_k : \max\left(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}\right) < 1.25^k,\ k \in \{1,2,3\} \tag{11}$$

Among these metrics, AbsRel and SqRel measure the discrepancy between the predicted depth and the ground-truth depth from the perspectives of absolute relative deviation and squared relative deviation, respectively. RMSE characterizes the overall magnitude of prediction errors in the absolute depth space and is more sensitive to large deviations. The *log*10 metric evaluates the difference between prediction and ground truth in the logarithmic space, thereby better reflecting the proportional error characteristics of depth estimation. In contrast, threshold accuracy (δ_k) measures the proportion of pixels whose predicted-to-ground-truth depth ratio falls within a specified threshold range, and is therefore used to characterize the prediction accuracy of the model. Specifically, δ_1, δ_2, and δ_3 correspond to thresholds of 1.25, 1.25^2 and 1.25^3, respectively.

It should be noted that lower values of AbsRel, SqRel, RMSE, and *log*10 indicate that the predicted depth is closer to the ground truth, whereas higher values of threshold accuracy (δ_k) indicate better prediction performance. Since these metrics jointly describe model behavior from the perspectives of relative error, absolute error, and threshold-based accuracy, they provide a comprehensive and complementary evaluation of the effectiveness and robustness of monocular depth estimation methods.

3. Results

3.1. Experimental Environment

All experiments were conducted on a unified hardware and software platform, as summarized in Table 1.

Table 1. Experimental platform.

Category	Item	Configuration
Operating system	OS	Microsoft Windows 11 Professional
Programming language	Python	3.12.3
Deep learning framework	PyTorch	2.3.1
GPU acceleration	CUDA	12.1
Backend library	CuDNN	8907
Hardware	GPU	NVIDIA GeForce RTX 4070
Hardware	GPU memory	12 GB VRAM

This configuration was sufficient to support the complete workflow of this study, including zero-shot inference, full fine-tuning, parameter-efficient fine-tuning, and result visualization.

To ensure fair comparability across different models, all methods were evaluated under a unified protocol on the same WE3DS test set, with identical input resolution and fully shared evaluation logic. For both the zero-shot experiments and the subsequent fine-tuning experiments, the same valid-pixel mask constraint and median scale alignment strategy were adopted, thereby avoiding potential biases introduced by inconsistencies in the evaluation protocol.

The hyperparameter values set before the experiment are shown in Table 2.

Table 2. Hyperparameter values of AgriLoRA-DA.

Hyperparameter	Value
epoch	120
patience	10
batch_size	8
img_size	518
Lr 0	1×10^{-3}
LoRA rank	2/4/6/8
alpha	2/4/6/8
targets	Attention (Q/K/V/O)

3.2. Zero-Shot Generalization on WE3DS

Under the zero-shot setting, substantial performance differences were observed among general-purpose depth models on WE3DS. As shown in Table 3. Among all compared methods, DA-v3 achieved the best overall performance, reaching 0.0438 AbsRel, 24.303 SqRel, 400.684 RMSE, 0.0189 *log*10, and 0.9979 δ_1. This result suggests that a stronger foundation depth model provides better cross-domain geometric generalization, even without target-domain adaptation. However, despite its clear advantage over the other zero-shot baselines, the remaining error is still considerable for WE3DS agricultural scenes, indicating that open-domain pretraining alone is insufficient to fully bridge the domain gap on this benchmark.

Table 3. Quantitative zero-shot comparison on WE3DS.

Method	AbsRel	SqRel	RMSE	log10	delta1
DA-v1 [20]	0.1757	610.175	1694.386	0.0741	0.7316
DA-v2 [21]	0.1967	1000.560	2008.809	0.0799	0.7063
DA-v3 [22,23]	0.0438	24.303	400.684	0.0189	0.9979
DPT [24]	0.2934	1524.738	2822.468	0.1347	0.5025
NeWCRFs [25,26]	0.1476	373.016	1614.451	0.0602	0.7978

The qualitative results in Figure 6 reveal a clear visual domain gap between agricultural scenes and the source-domain distributions of the pretrained models. Although DA-v3 preserves the global depth layout more faithfully than the other methods, noticeable errors remain around thin plant structures, object boundaries, and regions with self-occlusion. DPT and NeWCRFs exhibit more severe scale distortion and structural inconsistency, especially in scenes containing repetitive soil textures, sparse vegetation, and weak geometric cues. These observations are consistent with the quantitative results and collectively indicate that target-domain adaptation is necessary for accurate depth estimation in WE3DS.

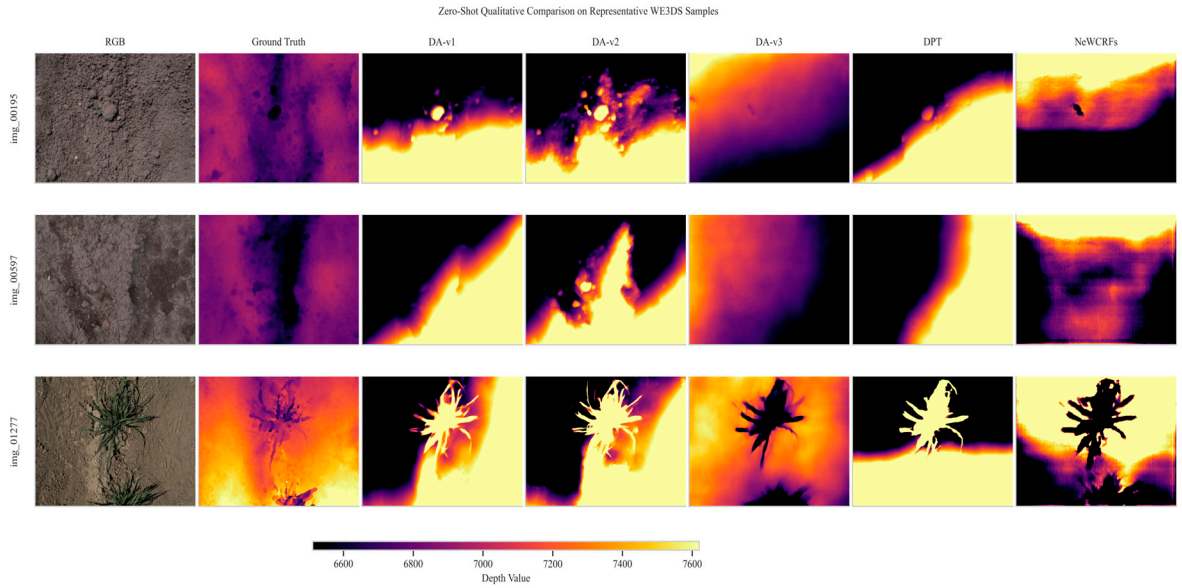

Figure 6. Qualitative zero-shot depth comparison of DA-v1, DA-v2, DA-v3, DPT, and NeWCRFs on representative WE3DS samples.

3.3. Comparison with Full Fine-Tuning and Efficient Adaptation

After target-domain adaptation, all major error metrics were reduced substantially compared with the baseline model. As shown in Table 4, introducing Lite-FPNHead

reduced AbsRel from 0.0168 to 0.0140 and RMSE from 195.109 to 140.100, demonstrating that a lightweight decoder is better suited to the multi-scale and fine-grained structures that frequently appear in WE3DS scenes. A further improvement was obtained by AgriLoRA-DA (QKV-r4), which achieved the best overall performance with 0.0133 AbsRel, 3.518 SqRel, 132.264 RMSE, 0.0057 $log10$, and 0.9990 δ_1. At the same time, it required only 0.19 M trainable parameters, highlighting the practical value of parameter-efficient adaptation under limited training and deployment resources.

Figure 7 summarizes the multi-dimensional trade-off between reconstruction accuracy and computational efficiency. The full fine-tuning baseline remains relatively expensive across parameter count, model size, and FLOPs, while Lite-FPNHead shifts the trade-off toward a more efficient operating point. AgriLoRA-DA (QKV-r4) occupies the most favorable overall region, combining the lowest reconstruction errors with the lowest adaptation cost.

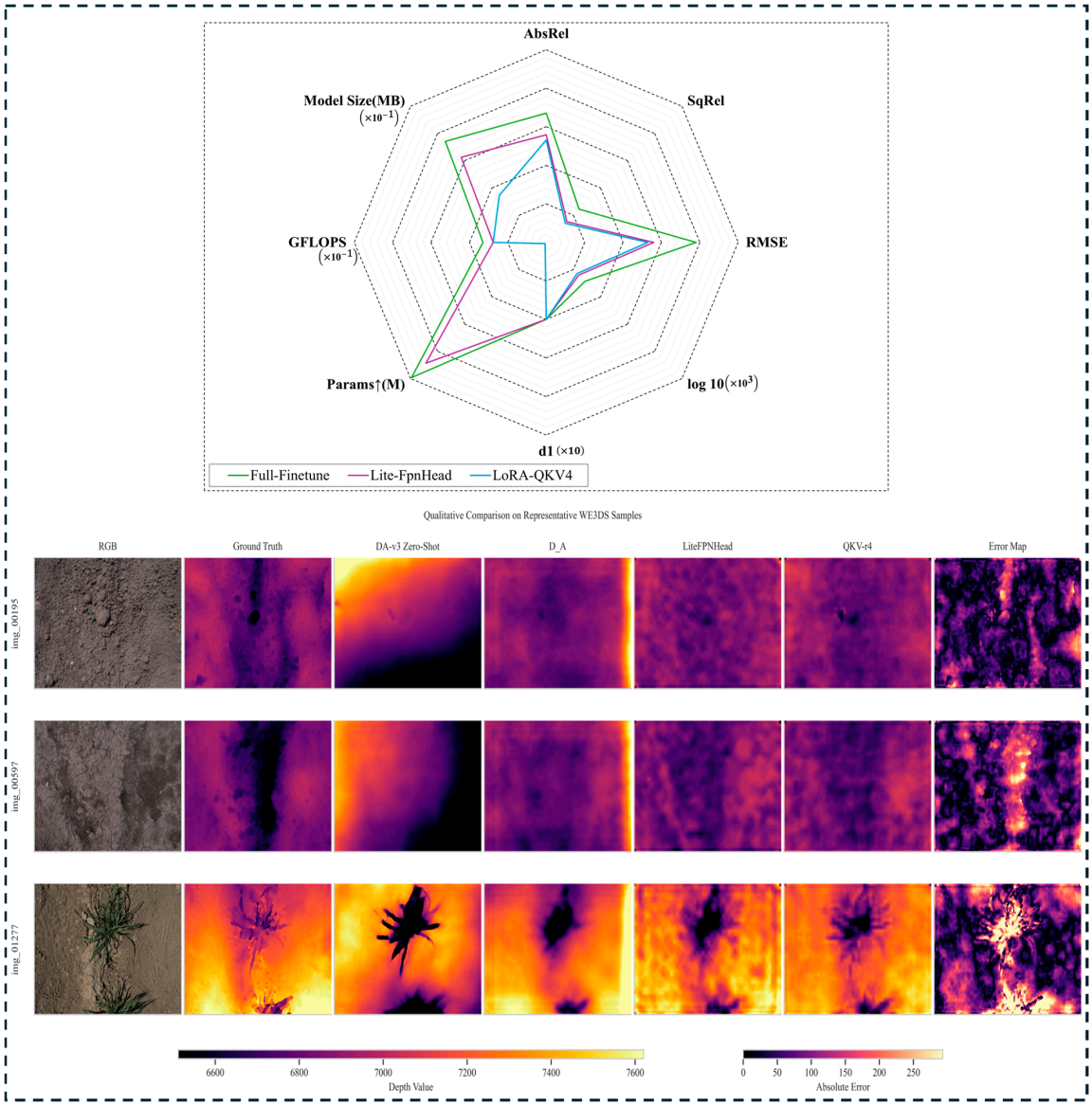

Figure 7. Main adaptation comparison on WE3DS. The figure summarizes the accuracy-efficiency trade-off and qualitative depth-estimation results of DA-v3 zero-shot, the full fine-tuning baseline, Lite-FPNHead, and AgriLoRA-DA (QKV-r4).

Table 4. Accuracy-efficiency comparison of the full fine-tuning baseline, Lite-FPNHead, and AgriLoRA-DA (QKV-r4) on WE3DS after target-domain adaptation. For the two fully fine-tuned baselines, parameter counts refer to optimized model parameters; for AgriLoRA-DA, the parameter count refers to trainable adaptation parameters.

Method	AbsRel	SqRel	RMSE	log10	δ_1	Params↑ (M)	GFLOPs	Model Size (MB)
Depth-Anything-V2	0.0168	6.080	195.109	0.0072	0.9990	24.79	82.4	185.6
Lite-FPNHead	0.0140	3.797	140.100	0.0060	0.9989	22.18	68.9	156.3
AgriLoRA-DA (QKV-r4)	0.0133	3.518	132.264	0.0057	0.9990	0.19	68.9	86.4

The qualitative comparison in Figure 7 provides further evidence of the benefit of target-domain adaptation. Relative to the zero-shot DA-v3 model, all adapted variants produce depth maps with more consistent global structure and clearer local boundaries. Compared with the full fine-tuning baseline, Lite-FPNHead suppresses part of the structural noise and yields cleaner transitions near object contours. AgriLoRA-DA (QKV-r4) produces the most faithful reconstructions overall, particularly in regions containing thin leaves, irregular plant edges, and cluttered soil backgrounds.

Figure 8 highlights the structural efficiency advantage of Lite-FPNHead. Compared with the original decoder, the number of decoder parameters is reduced by 95.54%, and decoder GFLOPs are reduced by 56.99%. These architectural savings translate into practical runtime benefits, with latency reduced by 55.17% and decoder-side GPU memory reduced by 35.78%. Therefore, the improvement brought by Lite-FPNHead should not be interpreted as a simple accuracy-efficiency compromise; rather, it functions as a better decoder design that improves both computational efficiency and adaptation quality.

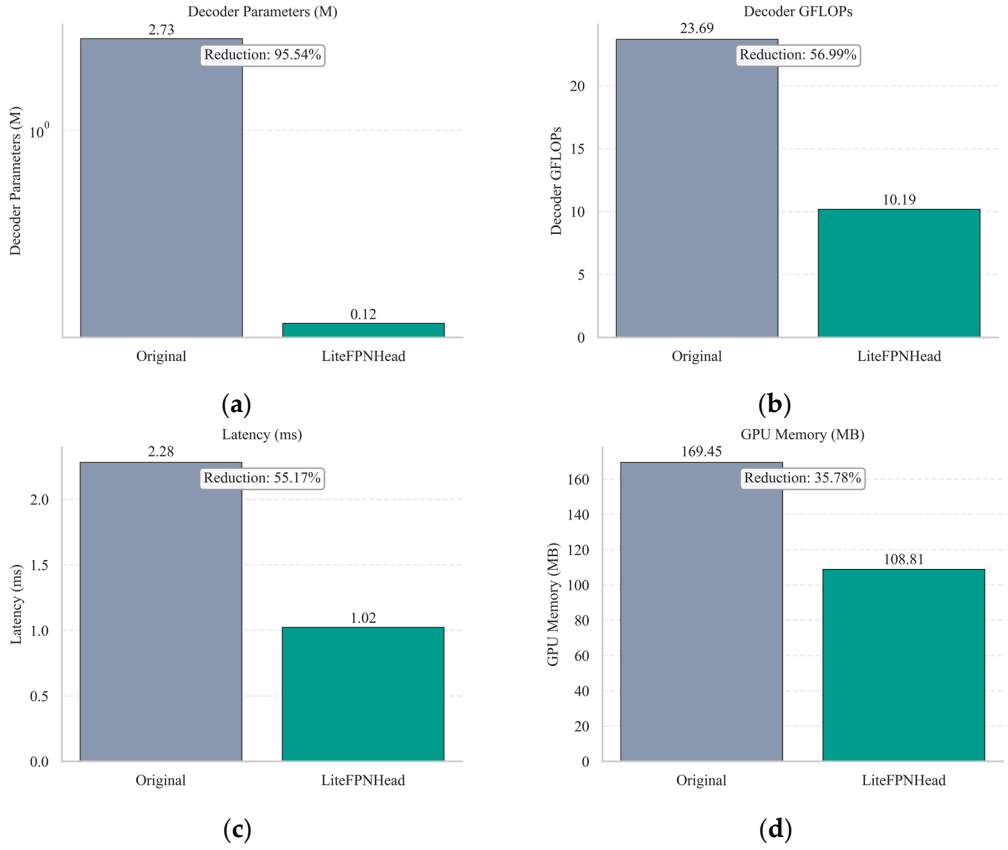

Figure 8. Efficiency comparison between the original decoder and LiteFPNHead in terms of (**a**) decoder parameters, (**b**) decoder GFLOPs, (**c**) latency, and (**d**) GPU memory.

3.4. LoRA Configuration Ablation and Training Dynamics

Figure 9 presents a compact ablation analysis of the LoRA configuration by jointly comparing insertion targets, rank settings, efficiency indicators, and validation dynamics. Among the tested QV, QKV, and QKVO insertion strategies with ranks of 2, 4, 6, and 8, QKV-r4 achieves the best overall accuracy, with 0.0133 AbsRel, 3.518 SqRel, 132.264 RMSE, and 0.0057 *log*10. This result indicates that adapting the query, key, and value projections provides sufficient flexibility to capture the agricultural domain shift while avoiding unnecessary parameter expansion. In contrast, further extending LoRA to the output projection does not lead to consistent improvement, suggesting that the dominant adaptation benefit mainly comes from the attention interaction components. The rank comparison also shows that increasing the rank beyond 4 does not consistently improve depth-estimation accuracy, although it increases the number of trainable parameters and memory cost. The validation curves further confirm that QKV-r4 provides stable convergence and a favorable balance between final accuracy and adaptation efficiency. Therefore, QKV-r4 is selected as the final LoRA configuration used in the proposed AgriLoRA-DA framework.

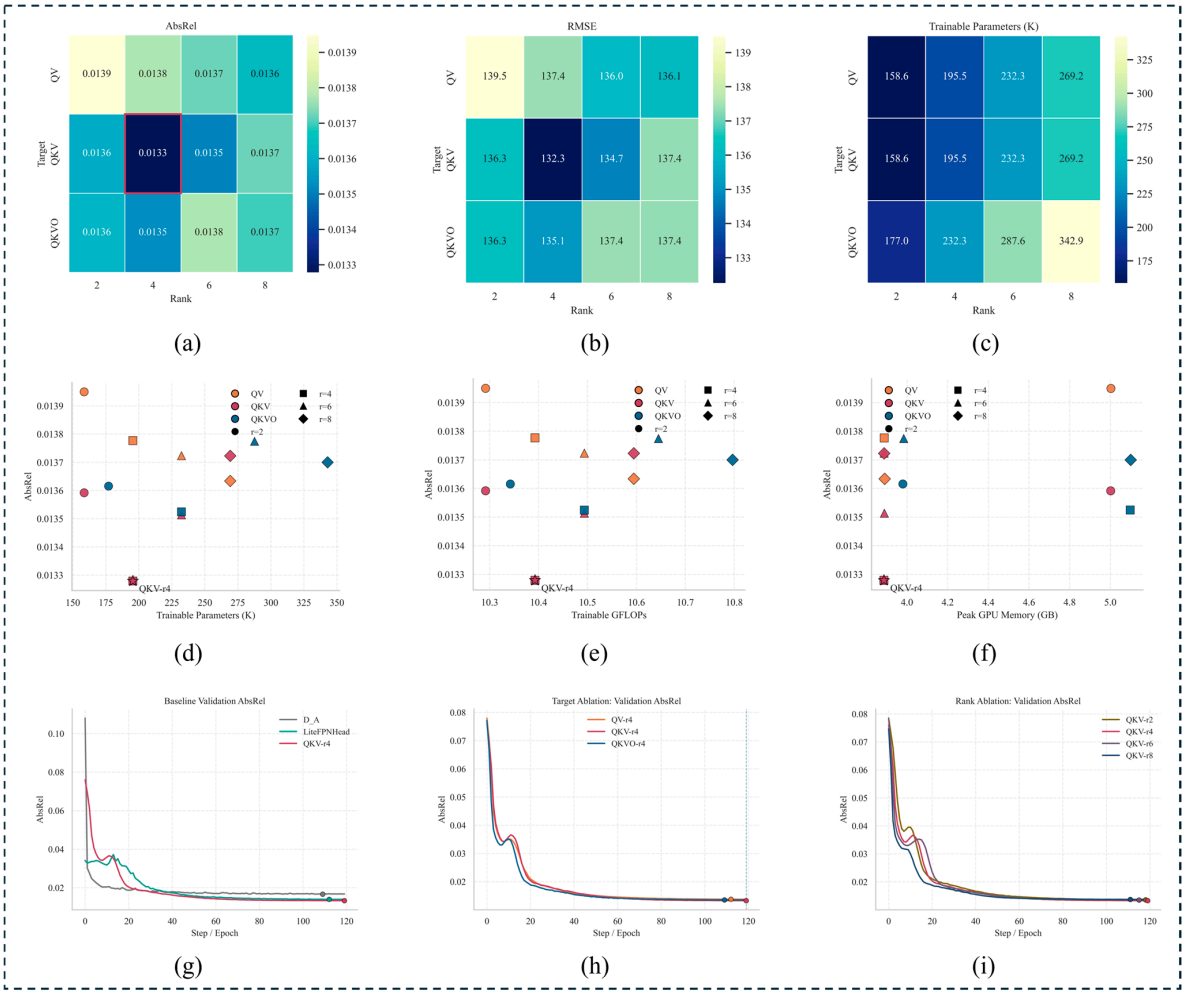

Figure 9. LoRA target/rank ablation and training dynamics. (**a**) AbsRel heatmap under different LoRA target/rank settings; (**b**) RMSE heatmap under different LoRA target/rank settings; (**c**) trainable-parameter heatmap under different LoRA target/rank settings; (**d**) AbsRel versus trainable parameters; (**e**) AbsRel versus trainable GFLOPs; (**f**) AbsRel versus peak GPU memory; (**g**) validation AbsRel curves for the baseline comparison; (**h**) validation AbsRel curves for LoRA target-module ablation; (**i**) validation AbsRel curves for LoRA rank ablation. The red box/marker highlights the selected QKV-r4 configuration, which achieves the best overall balance between accuracy and adaptation efficiency.

4. Discussion

The results indicate that strong open-domain pretraining alone is insufficient for reliable depth estimation on WE3DS agricultural scenes. Although DA-v3 clearly outperforms the other zero-shot baselines on WE3DS, noticeable errors remain around thin plant structures, occluded regions, and cluttered soil backgrounds, confirming the existence of a substantial agricultural domain gap. This observation is important because it shows that even recent foundation depth models cannot be directly assumed to generalize optimally to crop canopies and field imagery without target-domain adaptation.

The proposed framework addresses this issue from two complementary perspectives. On the encoder side, LoRA enables effective target-domain adaptation at very low training cost by updating only a small set of low-rank parameters instead of the full backbone. On the decoder side, Lite-FPNHead reduces structural complexity while preserving multi-scale aggregation, and the experimental results show that this redesign improves not only efficiency but also reconstruction quality in WE3DS scenes. A likely explanation is that the original DPT decoder is unnecessarily complex for this relatively compact agricultural benchmark, whereas the proposed top-down fusion pathway acts as an architectural regularizer and better matches plant structures such as leaves, stems, and narrow object boundaries. By removing heavy reassembly and fusion stages, Lite-FPNHead may reduce overfitting while retaining the multi-scale cues needed for dense depth recovery.

The ablation study further shows that adaptation quality is sensitive to where LoRA is inserted, but not necessarily improved by simply increasing rank. Among the tested settings, QKV-r4 provides the most favorable balance between accuracy and efficiency, whereas extending LoRA to the output projection or using larger ranks introduces additional cost without consistent gains. This suggests that, for agricultural monocular depth estimation, careful adaptation design is more important than increasing adaptation capacity. More broadly, the present results should be interpreted as evidence that LoRA is a practical parameter-efficient option for this task, rather than as a claim that it is universally superior to adapters, bias-only tuning, prompt-based methods, or partial fine-tuning in all agricultural settings.

From an application perspective, AgriLoRA-DA is most directly relevant to agricultural robots and unmanned plat-forms that require low-cost geometric perception for inter-row navigation, obstacle avoidance, canopy structure estimation, safe-distance maintenance, and field monitoring. Compared with LiDAR, stereo, or active RGB-D sensing, a monocular RGB camera can reduce hardware cost, power consumption, and mechanical complexity. The economic value of the proposed framework therefore lies not only in higher depth accuracy on WE3DS, but also in reducing the cost of adapting a pretrained model to local crop conditions: only a small number of trainable parameters must be stored and updated, making farm- or season-specific model adaptation more feasible.

Compared with prior agricultural depth-estimation studies that often rely on crop- or task-specific data acquisition, self-supervised video pipelines, or multi-stage perception systems [7–10], the present work focuses on adapting a pre-trained foundation depth model with a compact number of trainable parameters. This makes the proposed approach complementary to existing agricultural depth pipelines: it does not replace task-specific sensing designs when specialized hardware is available, but it provides a practical alternative when low-cost cameras and efficient local adaptation are preferred.

Several limitations should also be acknowledged. First, the evaluation is conducted on WE3DS only, and broader cross-dataset validation would further strengthen the generality of the conclusions. Accordingly, claims in this manuscript are restricted to WE3DS and to crop-row or canopy scenes with similar visual characteristics. The method is expected to be applicable to herbaceous field crops and row-crop scenarios, such as cereal crops, maize,

cotton, soy-bean, and vegetable canopies, after target-domain adaptation; however, use in orchards, vineyards, greenhouses with different camera settings, or highly different weather and season conditions should be validated separately. Second, the current study focuses on supervised adaptation and metric depth prediction, whereas practical agricultural deployment may also require stronger robustness to seasonal variation, dust, motion blur, and extreme illumination. Third, although the efficiency results are promising, real-time performance on embedded robotic hardware should be investigated more directly in future work.

5. Conclusions

In this paper, AgriLoRA-DA was proposed for agricultural monocular depth estimation by combining LoRA-based parameter-efficient adaptation with a lightweight LiteFPNHead built on top of Depth-Anything-V2. The experimental results on WE3DS demonstrated that, while foundation models such as DA-v3 provide strong zero-shot performance, target-domain adaptation remains necessary for accurate depth prediction on the WE3DS agricultural benchmark. After adaptation, the proposed method achieved the best overall performance, with 0.0133 AbsRel and 132.264 RMSE, while introducing only 0.19 M trainable parameters, thereby offering a favorable balance between accuracy and adaptation efficiency.

Overall, the study shows that agricultural depth estimation can benefit substantially from a joint design of efficient encoder adaptation and lightweight decoder reconstruction. The proposed framework is particularly suitable for low-cost monocular perception in crop-row robotic systems, where reducing sensing and adaptation costs is important for practical deployment. Future work will extend this framework to broader agricultural datasets, more diverse crop species and seasons, and deployment on edge robotic platforms for real-world field applications.

Author Contributions: Conceptualization, Y.M. and L.C.; methodology, Y.M. and L.C.; software, Y.M., W.Z. and L.C.; validation, Y.M. and L.C.; formal analysis, Y.M., W.Z. and L.C.; investigation, Y.M. and L.C.; resources, Y.M., W.Z. and L.C.; data curation, Y.M., W.Z. and L.C.; writing—original draft preparation, Y.M. and L.C.; writing—review and editing, L.C.; visualization, Y.M. and W.Z.; supervision, L.C.; project administration, L.C.; funding acquisition, L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Regional Innovation Guidance Plan of the Science and Technology Bureau of Xinjiang Production and Construction Corps (2023AB040, 2021BB012 and 20250196).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data utilized in this study can be obtained upon request from the corresponding author.

Acknowledgments: The authors would like to thank the research team members for their contributions to this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bai, Y.; Zhang, B.; Xu, N.; Zhou, J.; Shi, J.; Diao, Z. Vision-based navigation and guidance for agricultural autonomous vehicles and robots: A review. *Comput. Electron. Agric.* **2023**, *205*, 107584. [CrossRef]
2. Xiang, L.; Wang, D. A review of three-dimensional vision techniques in food and agriculture applications. *Smart Agric. Technol.* **2023**, *5*, 100259. [CrossRef]
3. Ehret, T. Monocular Depth Estimation: A Review of the 2022 State of the Art. *Image Process. Line* **2023**, *13*, 38–56. [CrossRef]

4. Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; Koltun, V. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1623–1637. [CrossRef] [PubMed]
5. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision Transformers for Dense Prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 12179–12188.
6. Bhat, S.F.; Alhashim, I.; Wonka, P. AdaBins: Depth Estimation Using Adaptive Bins. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2021; pp. 4009–4018. [CrossRef]
7. Zheng, Z.; Williams, H.; MacDonald, B.A. OrchardDepth: Precise Metric Depth Estimation of Orchard Scene from Monocular Camera Images. *arXiv* **2025**, arXiv:2502.14279. [CrossRef]
8. Cui, X.-Z.; Feng, Q.; Wang, S.-Z.; Zhang, J.-H. Monocular Depth Estimation with Self-Supervised Learning for Vineyard Unmanned Agricultural Vehicle. *Sensors* **2022**, *22*, 721. [CrossRef] [PubMed]
9. Coll-Ribes, M.; Torres-Rodriguez, I.J.; Grau, A.; Guerra, E.; Sanfeliu, A. Accurate detection and depth estimation of table grapes and peduncles for robot harvesting, combining monocular depth estimation and CNN methods. *Comput. Electron. Agric.* **2023**, *215*, 108362. [CrossRef]
10. Zhao, X.; Lyu, X.; Li, X. MDE-AgriVLN: Agricultural Vision-and-Language Navigation with Monocular Depth Estimation. *arXiv* **2025**, arXiv:2512.03958. [CrossRef]
11. Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; Zhao, H. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. *arXiv* **2024**, arXiv:2401.10891. [CrossRef]
12. Yang, L.; Kang, B.; Huang, Z.; Zhao, Z.; Xu, X.; Feng, J.; Zhao, H. Depth Anything V2. *arXiv* **2024**, arXiv:2406.09414. [CrossRef]
13. Lin, H.; Chen, S.; Liew, J.; Chen, D.Y.; Li, Z.; Shi, G.; Feng, J.; Kang, B. Depth Anything 3: Recovering the Visual Space from Any Views. *arXiv* **2025**, arXiv:2511.10647. [CrossRef]
14. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment Anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 4015–4026. [CrossRef]
15. Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. SAM 2: Segment Anything in Images and Videos. *arXiv* **2024**, arXiv:2408.00714. [CrossRef]
16. Carion, N.; Gustafson, L.; Hu, Y.-T.; Debnath, S.; Hu, R.; Suris, D.; Ryali, C.; Alwala, K.V.; Khedr, H.; Huang, A.; et al. SAM 3: Segment Anything with Concepts. *arXiv* **2025**, arXiv:2511.16719. [CrossRef]
17. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685. [CrossRef]
18. Kitzler, F.; Barta, N.; Neugschwandtner, R.W.; Gronauer, A.; Motsch, V. WE3DS: An RGB-D Image Dataset for Semantic Segmentation in Agriculture. *Sensors* **2023**, *23*, 2713. [CrossRef] [PubMed]
19. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv* **2023**, arXiv:2304.07193. [CrossRef]
20. LiheYoung. Depth Anything GitHub Repository. Available online: https://github.com/LiheYoung/Depth-Anything (accessed on 9 May 2026).
21. DepthAnything. Depth Anything V2 GitHub Repository. Available online: https://github.com/DepthAnything/Depth-Anything-V2 (accessed on 9 May 2026).
22. ByteDance-Seed. Depth Anything 3 GitHub Repository. Available online: https://github.com/ByteDance-Seed/Depth-Anything-3 (accessed on 9 May 2026).
23. Depth Anything. DA3-SMALL Hugging Face Model. Available online: https://huggingface.co/depth-anything/DA3-SMALL (accessed on 9 May 2026).
24. Intel ISL. DPT GitHub Repository. Available online: https://github.com/isl-org/DPT (accessed on 9 May 2026).
25. Alibaba. NeWCRFs GitHub Repository. Available online: https://github.com/aliyun/NeWCRFs (accessed on 9 May 2026).
26. Yuan, W.; Gu, X.; Dai, Z.; Zhu, S.; Tan, P. Neural Window Fully-connected CRFs for Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022*; IEEE: New York, NY, USA, 2022; pp. 3916–3925.