PastureNet: Cross-Domain Biomass Estimation via Foundation Models

Joseph Siu
Department of Computer Science
Tsinghua University
University of Toronto
xdx25@mails.tsinghua.edu.cn

Ren Fu
Department of Computer Science
Tsinghua University
flr25@mails.tsinghua.edu.cn

Rubin Camillo Wallner
School of Economics and Management
Tsinghua University
University of Munich
wlb25@mails.tsinghua.edu.cn

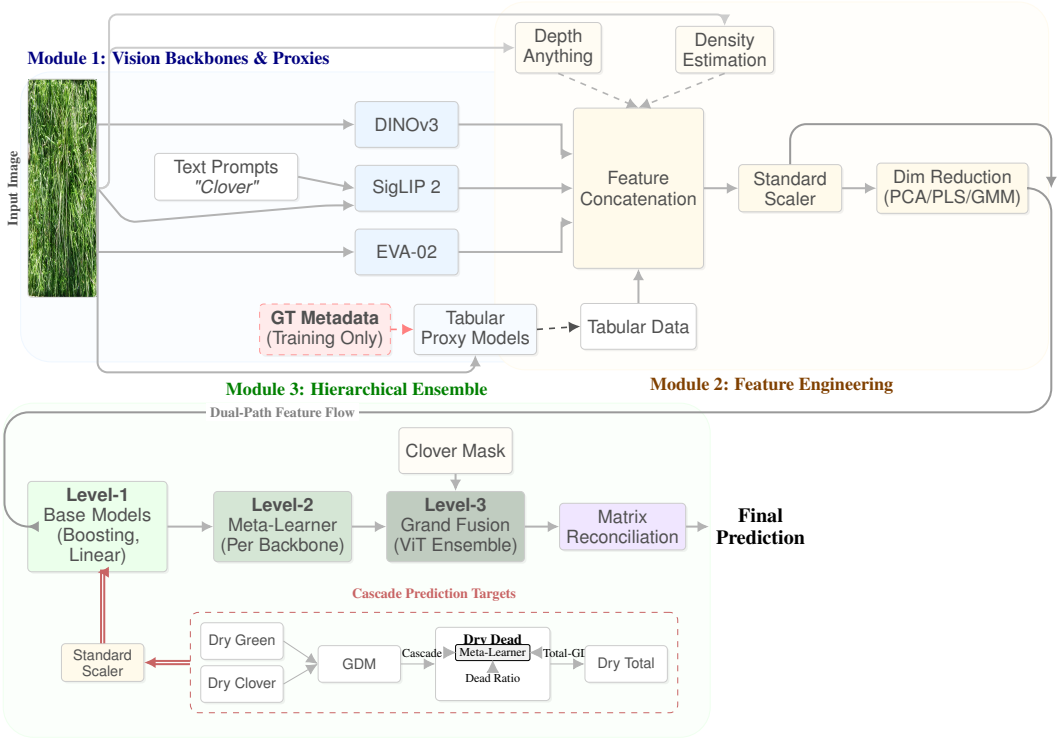


Figure 1: The PastureNet architecture. A hierarchical framework for biomass estimation without inference-time metadata. **Module 1** extracts multi-modal features using three diverse Foundation Models (DINOv3, SigLIP 2, EVA-02). **Module 2** generates "Imputed Metadata" via auxiliary proxy models and fuses them with semantic scores. **Module 3** employs a multi-level stacking strategy, culminating in Matrix Reconciliation to ensure biological consistency across prediction targets.

Preprint. Under review.

Abstract

Accurate pasture biomass estimation is critical for precision grazing management yet remains challenged by the trade-off between the scalability of remote sensing and the reliability of manual sampling. To address this, we introduce PastureNet, a novel hierarchical ensemble framework that estimates biomass directly from high-resolution RGB images. Unlike traditional approaches, PastureNet synergizes diverse inductive biases by integrating three state-of-the-art Vision Transformers: DINOv3 (object-centric), SigLIP 2 (semantic-aligned), and EVA-02 (texture-sensitive). A key innovation is the integration of Zero-shot Semantic Concept Scores to inject explicit ecological domain knowledge (e.g., clover presence) into the regression pipeline, alongside a Matrix Reconciliation post-processing step that ensures biological consistency across biomass components. Evaluated on a heterogeneous Australian dataset, our method achieves a Weighted $\mathbf{R^2}$ of 0.70, significantly outperforming CNN baselines (0.47) and demonstrating robust generalization without requiring physical metadata at inference time.

1 Introduction

Accurate estimation of pasture biomass is a critical determinant of productivity and sustainability in livestock agriculture. The ability to reliably answer the seemingly simple question: "Is there enough grass?" directly influences grazing management, animal nutrition, and long-term land health. Underestimation leads to underutilization of valuable forage resources, while overestimation risks overgrazing, soil degradation, and economic loss. Despite its fundamental importance, biomass quantification remains a persistent operational challenge, trapped between the precision of manual, destructive methods and the scalability, but often limited accuracy, of indirect sensing technologies.

Traditional ground-based techniques, such as the "clip and weigh" method, provide reliable localized measurements but are labor-intensive, destructive, and impractical for frequent, large-scale monitoring. Instrument-based approaches like rising plate meters or capacitance probes improve throughput but are sensitive to sward structure, moisture, and operator technique, introducing variability. Remote sensing, particularly using vegetation indices (e.g., NDVI) derived from satellite or drone imagery, enables broad-area coverage but struggles with site-specific calibration, cannot differentiate between botanically distinct species with similar spectral signatures, and is often confounded by soil exposure and senescent material. This leaves a clear technological gap: a method that is as reliable as direct measurement, as scalable as remote sensing, and robust to the natural heterogeneity of pastures.

The convergence of field imaging and modern machine learning offers a promising path forward. Prior work has primarily focused on convolutional neural networks (CNNs) for this vision-based regression task. However, the recent ascendancy of Vision Transformers (ViTs) presents a compelling alternative, offering superior capabilities in capturing long-range dependencies and complex spatial patterns within an image. While multi-modal approaches that combine images with tabular metadata (e.g., species mix, location) exist, they often rely on simplistic fusion methods or become impractical in field settings where detailed metadata is unavailable at inference time.

In this work, we introduce **PastureNet**, a novel, practically-oriented framework that leverages the strengths of modern vision architectures and ensemble learning to achieve robust biomass estimation using only images. Our approach specifically targets the integration of **complementary visual cues**—including depth information, 3D structure, and species-specific density patterns—through a sophisticated multi-stage architecture that does not require auxiliary metadata at test time.

Beyond structural features, we introduce a novel mechanism to capture semantic ecological context by exploiting the multi-modal nature of SigLIP 2. Unlike standard ViTs that produce abstract embeddings, SigLIP 2 is trained to align visual features with text descriptions. We leverage this by querying the image encoder with a curated set of natural language prompts tailored to pasture analysis, such as "white clover flowers", "dry brown dead grass", and "lush green vibrant pasture". The dot-product similarity between the image embedding and these text embeddings yields explicit "Semantic Concept Scores." These scores act as soft, zero-shot classifiers, allowing the model to explicitly differentiate between biomass-contributing components (e.g., green leaf area) and confounding factors (e.g., dead matter or weeds) without requiring pixel-level annotation.

Our core innovation is a specialized two-stage ensemble that generates and fuses diverse predictive signals from pasture imagery. In the first stage, we create a heterogeneous set of base predictors:

1. **Vision Transformer (ViT)** models that extract global contextual patterns from RGB images,
2. **Depth-aware features** using specialized models (e.g., Depth Anything V2) to capture 3D canopy structure,
3. **Density estimators** that predict biomass distribution from localized patch features, and
4. **Tabular proxy models** that learn to predict metadata values (e.g., NDVI, species class) from images using Gradient Boosting regressors.
5. **Zero-shot semantic extractors** that leverage SigLIP 2's vision-language alignment to quantify ecological attributes. By querying images with domain-specific text prompts (e.g., "lush green pasture", "white clover"), we generate explicit semantic probability scores and biomass ratios as interpretable inputs for the ensemble.

Crucially, these tabular proxy models are trained exclusively on image features and only use metadata during training, creating a distilled representation that requires no external data at inference. In the second stage, a meta-learner (a lightweight neural network) learns optimal fusion weights for these diverse predictions, effectively distilling both direct visual evidence and implicitly learned ecological context into a single robust model.

We evaluate PastureNet on a professionally curated dataset of Australian pastureland, containing images paired with biomass measurements and auxiliary metadata. Our experiments demonstrate that the proposed multi-feature ensemble significantly outperforms both individual Vision Transformers and previous CNN-based benchmarks. The resulting system provides a scalable, accurate, and deployable tool for precision pasture management, translating visual data into immediate, actionable insights for farmers.

Our contributions are threefold:

1. **The PastureNet Architecture:** A novel, vision-only ensemble model that generates and integrates complementary predictions from multiple feature sources—including ViT representations, semantic concept scores, depth features, density estimates, and tabular proxy predictions—via a meta-learner, eliminating the inference-time need for external metadata.
2. **A Training Strategy for Feature Distillation:** A methodology that trains auxiliary models to predict tabular metadata from images during training, then uses these predictions as additional features for the ensemble, effectively distilling contextual knowledge into the fusion process without requiring metadata at deployment.
3. **Empirical Validation:** Comprehensive experiments on a real-world pasture dataset demonstrating state-of-the-art performance for an image-based model, highlighting the practical advantage of our multi-feature ensemble approach for sustainable agricultural management.

2 Related Work

Our work is situated at the intersection of computer vision for agriculture, multi-modal learning, and advanced ensemble techniques. This section reviews the relevant literature across these domains.

2.1 Computer Vision for Biomass Estimation

A primary line of research applies deep learning directly to images for biomass prediction. Convolutional Neural Networks (CNNs), such as ResNet and VGG architectures, have become the de facto standard[Li et al., 2024], extracting spatial-hierarchical features from pasture and crop imagery. These models learn to correlate visual textures, color, and canopy structure with biomass mass. More recently, Vision Transformers[Dosovitskiy et al., 2021] (ViTs) have emerged as a powerful alternative. By leveraging self-attention mechanisms, ViTs model long-range dependencies within an image, potentially capturing global canopy patterns that CNNs might miss due to their localized receptive fields. Studies in related agricultural tasks [Mehdipour et al., 2026] have shown ViTs can match or exceed CNN performance, particularly when data is abundant. Our model leverages the global context awareness of ViTs to form a visual feature extractor.

Building on this foundation, recent large-scale pre-trained vision models have demonstrated remarkable transfer learning capabilities. Models such as DINOv3 [Siméoni et al., 2025], EVA-02 [Fang et al., 2023], and SigLIP [Zhai et al., 2023] represent the state-of-the-art in visual representation learning, trained on billions of images with advanced self-supervised objectives. In agricultural applications, these models have shown promise for fine-grained recognition tasks [Mehdipour et al., 2026], but their application to regression tasks like biomass estimation remains underexplored, particularly in ensemble configurations.

2.2 Multi-Modal and Ensemble Learning

Acknowledging the limitations of unimodal data, recent work explores multi-modal fusion. A common but simplistic approach is *early fusion* or *feature concatenation*, where image-derived features and tabular data (e.g., soil moisture, NDVI) are merged into a single vector for a final regressor. This method often fails to model complex, non-linear interactions between fundamentally different data types. More sophisticated strategies include *late fusion* (averaging predictions from separate models) or cross-modal attention mechanisms. Our approach is most closely aligned with *model stacking* or *super learning*, an ensemble technique where predictions from diverse base learners ("level-0" models) are used as input features for a second-stage "meta-learner". This framework is theoretically optimal for combining predictive power, but its application in agriculture remains underexplored.

Recent advances in hierarchical ensemble methods [Odegua, 2019] demonstrate that two-level stacking architectures can significantly outperform simple averaging or voting schemes.

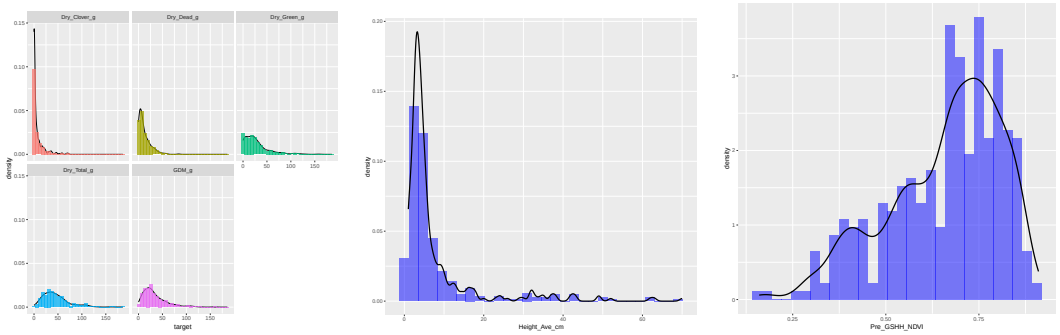
2.3 Regression Techniques and Meta-Learning

The tabular component of our problem draws from classical and modern regression. Linear models with regularization—Ridge regression (L2 penalty) and Lasso regression (L1 penalty)—are foundational for handling correlated features and performing automatic feature selection, respectively. Gradient Boosting Machines (GBMs), such as XGBoost [Chen and Guestrin, 2016] and LightGBM [Ke et al., 2017], construct powerful non-linear models by sequentially correcting the errors of previous trees and are often state-of-the-art for tabular data. In our ensemble, these algorithms constitute the suite of base learners trained on the tabular metadata. The concept of a *meta-learner* (or *blender*) that learns to optimally combine these base predictions is core to stacked generalization. Our model investigates the capacity of powerful meta-learners to capture complex weighting schemes conditioned on the learned behaviors of the image and tabular base models.

Parameter-efficient fine-tuning techniques, particularly Low-Rank Adaptation (LoRA) [Hu et al., 2021], have emerged as a crucial development for adapting large pre-trained models to specialized domains with limited data. While widely used in natural language processing, their application to vision tasks in agriculture [Liu et al., 2025] is more recent and provides a pathway to leverage billion-parameter models without prohibitive computational costs.

2.4 Positioning of Our Contribution

Prior research has established the value of CNNs/ViTs for image-based estimation and of ensemble methods like boosting for tabular data. However, a significant gap exists in developing integrated models that: (1) jointly leverage multiple state-of-the-art vision transformers (EVA-02, DINOv3, SigLIP) within a single framework, (2) employ a two-level hierarchical meta-learner architecture to exploit the complementary strengths of different model families and regression algorithms, (3) incorporate domain-specific strategies such as clover detection masking and dual-path dead biomass estimation, and (4) utilize parameter-efficient adaptation techniques for practical deployment. **PastureNet** addresses this comprehensive gap by proposing a coherent architecture that unifies these advanced components into a single, deployable system for pasture biomass estimation.



(a) Biomass Components Density (b) Pasture Height Distribution (c) NDVI Distribution

Figure 2: Distribution analysis of the dataset. (a) shows the long-tail distribution of biomass components (Dry, Green, Dead). (b) and (c) illustrate the distributions of auxiliary tabular features (Height and NDVI), which are utilized in our imputation module.

3 Dataset and Exploratory Analysis

3.1 Data Characteristics and Distributions

The dataset comprises a professionally curated collection of Australian pasture images paired with ground-truth biomass measurements. As illustrated in Figure 2a, the target variable (Total Biomass) and its sub-components (Green, Dead, Clover) exhibit a pronounced long-tail distribution. Specifically, the Dry_Clover_g component is highly sparse, with a significant portion of samples containing zero values, while Dry_Green_g constitutes the majority of the biomass but retains a non-Gaussian skew.

This distributional heterogeneity poses a challenge for standard regression objectives, motivating our use of robust loss functions. Furthermore, auxiliary physical attributes—specifically Pasture Height and NDVI (Normalized Difference Vegetation Index)—were analyzed. As shown in Figure 2b and Figure 2c, these attributes follow distinct distributions that physically constrain the biomass potential. The preservation of these statistical properties during the training of our Tabular Proxy models is critical for accurate imputation.

3.2 Correlation Analysis and Physical Proxies

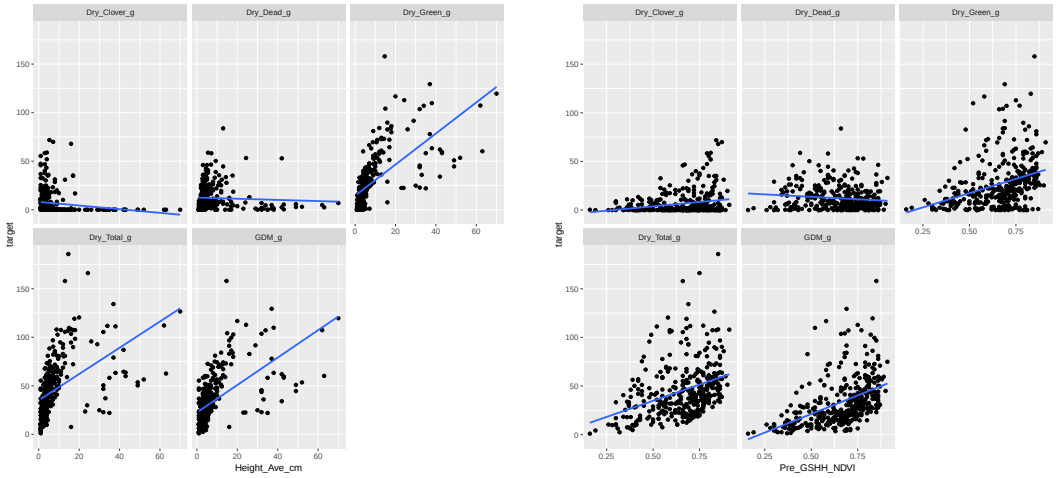
To validate the efficacy of integrating tabular features, we analyzed the linear relationships between physical proxies and the target biomass. Figure 3 presents the scatter plots of Biomass versus Pasture Height and Biomass versus NDVI.

We observe a strong positive correlation, particularly between Pasture Height and Total Biomass ($r > 0.6$). However, the relationship exhibits significant heteroscedasticity, where variance increases with biomass mass. This confirms that while Height and NDVI are strong predictors, they are insufficient as standalone regressors. This observation supports our PastureNet architecture, which utilizes the Vision Transformer backbone to capture complex textural features that explain the variance not captured by simple linear proxies, while using the Tabular Proxy module to anchor predictions within physically plausible ranges.

3.3 Categorical Heterogeneity and Species Bias

A critical challenge in pasture estimation is the biological diversity of the sward. Figure 4 highlights the significant class imbalance and domain shifts present in the data.

First, Figure 4a reveals that the dataset is dominated by specific grass mixes (e.g., Ryegrass), while legume-heavy samples (e.g., White Clover) are underrepresented. Second, and crucially, Figure 4b demonstrates that species composition fundamentally alters the biomass-to-visual relationship; samples containing clover exhibit distinct biomass distributions compared to pure grass samples. This empirical discrepancy provides the direct motivation for our Clover Existence Masking strategy, which conditionally activates specific regression heads when clover is detected. Finally, Figure



(a) Biomass vs. Height

(b) Biomass vs. NDVI

Figure 3: Correlation analysis between total biomass and physical proxies. The positive correlation observed in both (a) Height and (b) NDVI validates our strategy of training auxiliary tabular proxy models to assist the vision backbone.

4c shows biomass variations across different Australian states (NSW, WA, Vic), underscoring the necessity for the domain adaptation techniques applied in our preprocessing pipeline.

4 Methodology

The overall architecture of PastureNet is illustrated in Figure 1. Our approach leverages a hierarchical ensemble of diverse vision transformers to extract multi-modal features, which are then fused and refined through a multi-stage regression pipeline.

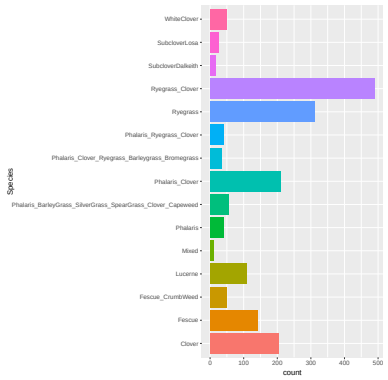
4.1 Vision Backbones and Pre-training

4.1.1 Backbones

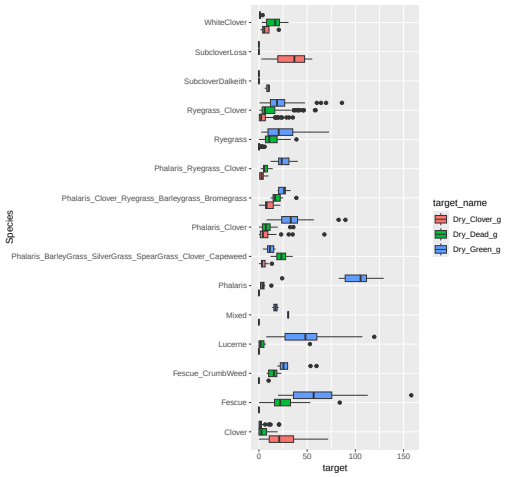
To capture robust visual representations of pasture biomass under varying field conditions, we leverage a diverse set of state-of-the-art Vision Transformers (ViTs). Each backbone contributes a unique inductive bias derived from its specific pre-training objective:

DINOv3 We employ DINOv3 [Siméoni et al., 2025], a self-supervised model that extends the distillation-based training of its predecessors. DINOv3 utilizes a student-teacher architecture where the student network learns to predict the output of a dynamically updated teacher network. Unlike standard contrastive methods that rely on negative pairs, DINOv3 focuses on regularization and specific architectural biases to learn "object-centric" representations. By enforcing consistency between different augmented views of the same image, it captures high-level structural information (e.g., separating sward canopy from soil background) without requiring semantic labels, making it highly effective for extracting dense visual descriptors.

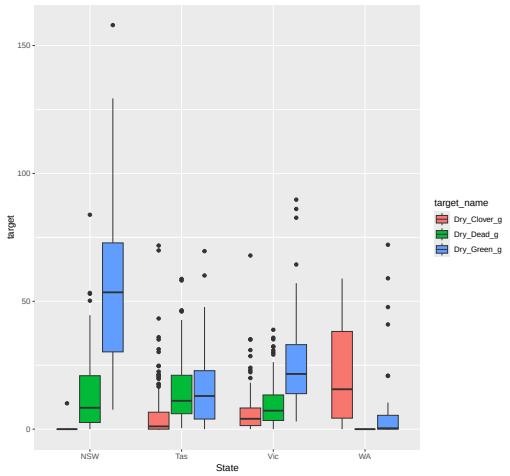
SigLIP 2 We incorporate SigLIP 2 [Tschannen et al., 2025], which optimizes the alignment between image and text modalities using a Sigmoid Loss for Language-Image Pre-training. Unlike the Softmax loss used in standard CLIP [Radford et al., 2021] which requires global normalization across a batch, SigLIP treats the image-text matching problem as a binary classification task for every pair. This allows for scaling to larger batch sizes and more efficient training. Crucially, SigLIP 2 aligns visual features with semantic concepts in a shared embedding space. This alignment enables our model to leverage the "semantic concept scores" described in Section 4.2.1, effectively grounding visual patterns (e.g., "green leaf") to human-interpretable linguistic attributes.



(a) Species Sample Count (Class Imbalance)



(b) Biomass Distribution by Species



(c) Biomass Distribution by State

Figure 4: Categorical analysis highlighting data heterogeneity. (a) reveals significant class imbalance, particularly for Clover-mixes. (b) shows how specific species compositions (e.g., Ryegrass vs. Clover) dictate biomass ranges, motivating our *Clover Existence Masking*. (c) displays regional variations across states (NSW, Vic, Tas, WA).

EVA-02 We utilize EVA-02 [Fang et al., 2023], a model trained via Masked Image Modeling (MIM). EVA-02 reconstructs masked patches of an input image, conditioned on visible patches. Specifically, it uses the features from a CLIP-aligned teacher model as the reconstruction target rather than raw pixels. This approach forces the model to learn fine-grained contextual details and texture dependencies to "fill in the blanks," which is particularly advantageous for biomass estimation where the density and texture of the grass sward are key predictors.

4.1.2 Preprocessing

We explore a suite of preprocessing techniques to normalize environmental variability.

- **Gray World Algorithm**[Buchsbaum, 1980]: Assumes that the average reflectance of a scene is achromatic (gray). It adjusts the image channels to balance the average color, correcting for color casts caused by varying ambient light (e.g., overcast vs. sunny).
- **Percentile-based White Patch**[Land, 1977]: A robust variation of the White Patch algorithm. Instead of using the single brightest pixel (which may be noise), it assumes the 99th percentile of pixel intensity represents the "true white" of the scene. It scales the channels to map this value to white, effectively normalizing the dynamic range.

- **CLAHE (Contrast Limited Adaptive Histogram Equalization)**[Zuiderveld, 1994]: Enhances local contrast by equalizing the histogram in small grid regions. This is crucial for revealing texture details in shadowed or over-exposed areas of the dense grass canopy.
- **FDA (Fourier Domain Adaptation)**[Yang and Soatto, 2020]: Aligns the frequency content of the input image with a reference target style (e.g., the mean style of the training set) in the Fourier domain, reducing the "style gap" between images taken under different conditions.

High-Resolution Input Strategy Unlike standard approaches that resize images to 224×224, we utilize a significantly higher input resolution of 1792×896 (2:1 aspect ratio, maintaining the input ratio). This design choice is critical for biomass estimation, as distinguishing intermixed species (e.g., clover amidst grass) and estimating sward density relies on fine-grained textural cues that are lost at lower resolutions. To accommodate this, we leverage the advanced resolution-handling capabilities of our selected backbones:

- **SigLIP 2 (NaFlex Integration)**: We utilize the *NaFlex* (Native Flexible resolution) variant of SigLIP 2 [Tschannen et al., 2025]. Unlike traditional models that require square resizing or padding, NaFlex is trained to handle variable aspect ratios and resolutions natively. It processes the 1792×896 input by preserving the original patch sequence layout, ensuring that the learned semantic alignment remains robust without geometric distortion.
- **DINOv3 and EVA-02 (Positional Interpolation)**: Both DINOv3 and EVA-02 are Vision Transformers designed with scalability in mind. To adapt them to our 1792×896 input, we employ Positional Embedding Interpolation. Since these models are pre-trained at lower resolutions (e.g., 518^2 or 384^2), we mathematically interpolate their learned positional encodings (using bicubic interpolation) to match the larger grid size. This allows the self-attention mechanism to operate globally over the high-density feature map, effectively capturing long-range dependencies across the entire pasture scene while retaining local high-frequency details.

Given the substantial memory footprint of this high-resolution strategy, we implement Gradient Accumulation and Dynamic Batch Sizing in our training pipeline to maintain stable optimization on hardware with limited VRAM.

4.1.3 Domain Adaptation

Pasture appearance varies significantly across regions due to soil types, species composition, and climate. To bridge the domain gap between European and Australian pastures, we implement a Transfer Learning strategy.

We utilize the *Irish Grass Clover Dataset* [Albert et al., 2024] as a source domain for intermediate pre-training. Although geographically distinct, the Irish dataset shares fundamental biological characteristics with our target domain (e.g., the visual distinction between ryegrass and clover). By first fine-tuning our backbones on this larger dataset, the models learn generalized features of sward structure and legume detection. These weights serve as a warm-start initialization for the final fine-tuning on the smaller Australian dataset, accelerating convergence and improving robustness to species heterogeneity.

Figure (a) shows an example from the *Irish Grass Clover Dataset*, and Figure (b) shows an example from the Australian pastureland dataset.

4.2 Feature Engineering and Augmentation

To synthesize the diverse information captured by our backbone models, we employ a multi-modal fusion strategy followed by robust dimensionality reduction.

4.2.1 Multi-modal Fusion

We construct a comprehensive *Augmented Feature Vector* for each sample by concatenating three distinct information sources:

- **Visual Embeddings:** The high-dimensional latent representations extracted from the penultimate layers of DINOv3, SigLIP 2, and EVA-02 backbones.

(a) Source Domain (Irish Dataset)

(b) Target Domain (Australian Dataset)

Figure 5: Domain Adaptation Visualization. The models are pre-trained on the lush, clover-rich Irish pastures (a) to learn general vegetative features before being adapted to the more arid and diverse Australian pastures (b).

- **Semantic Scores:** Leveraging the vision-language alignment of SigLIP 2, we compute dot-product similarity scores between image embeddings and a set of domain-specific prompts (e.g., "lush green pasture", "flowering clover") [Tschannen et al., 2025]. These scores act as explicit semantic descriptors, quantifying the presence of specific biological attributes.
- **Imputed Tabular Data:** Since external metadata (e.g., NDVI, Pasture Height) is unavailable at test time, we train auxiliary Boosting regressors (CatBoost) to predict these values solely from visual features. These models achieved strong predictive performance ($R^2 \approx 0.82$ for Pasture Height and $R^2 \approx 0.74$ for NDVI). These imputed physical proxies are then fed back into the main pipeline as auxiliary features, grounding the model in physical reality.

4.2.2 Dimensionality Reduction

The raw concatenated feature vector contains thousands of dimensions, posing a risk of overfitting for downstream regressors, especially the linear models. To mitigate this, we apply a suite of unsupervised reduction techniques: Principal Component Analysis (PCA) [Pearson, 1901] to preserve global variance, Partial Least Squares (PLS) [Wold et al., 2001] to align features with target variance, and Gaussian Mixture Models (GMM) [Reynolds, 2009] to capture cluster-based densities. This compresses the feature space to a compact representation ($D \approx 64 - 128$) while retaining critical signal.

Supervised Dimensionality Reduction (PLS) While Principal Component Analysis (PCA) performs unsupervised variance maximization, we strictly employ Partial Least Squares (PLS) as a supervised reduction technique for our linear models. Unlike PCA, PLS constructs latent components by maximizing the covariance between the input features X and the target biomass y [Wold et al., 2001]. This ensures that the retained dimensions are not merely high-variance (which could be noise) but are physically predictive of the biomass, effectively filtering out visual artifacts irrelevant to the regression task.

Feature Standardization Prior to dimensionality reduction, it is critical to address the scale disparity between different feature sources (e.g., high-magnitude DINOv3 vectors vs. bounded SigLIP probability scores). We apply Z-score normalization (StandardScaler) to all input features:

$$x'_j = \frac{x_j - \mu_j}{\sigma_j} \tag{1}$$

This transformation ensures that each feature contributes equally to the variance maximization in PCA and covariance maximization in PLS, preventing features with larger raw magnitudes from dominating the subspace projection.

Differentiated Reduction Strategy We implement a model-specific feature engineering strategy based on the inductive biases of our Level-1 learners.

- **For Linear Models (Ridge, Lasso):** We strictly utilize the compressed features (PCA/PLS/GMM) derived above. Since linear models are highly sensitive to multicollinearity and the "curse of dimensionality," working in this orthogonal, lower-dimensional space ($D \approx 128$) significantly improves stability and generalization.
- **For Boosting Models (CatBoost, XGBoost):** We bypass the aggressive dimensionality reduction and feed the standardized full-dimensional features (or a larger subset) directly. Gradient Boosting Machines are inherently capable of feature selection and handling non-linear interactions; supplying them with the raw, high-dimensional representations preserves fine-grained textural information that might be lost during PCA compression.

4.3 Hierarchical Ensemble Strategy

We implement a three-tier stacking architecture designed to maximize diversity and generalization.

4.3.1 Level-1: Heterogeneous Base Learners

To extract maximal information from the augmented feature space, we train two distinct families of regressors for every target variable. This heterogeneity is crucial as it captures complementary patterns in the data.

Gradient Boosting Machines (GBMs) We employ CatBoost [Prokhorenkova et al., 2018], XGBoost [Chen and Guestrin, 2016], and LightGBM [Ke et al., 2017]. These models excel at capturing non-linear interactions between visual features and imputed tabular metadata. Specifically, CatBoost's ordered boosting handles the distributional shifts in our augmented features effectively, while XGBoost and LightGBM provide robust handling of potential outliers through histogram-based learning.

Regularized Linear Models To complement the high-variance nature of boosting trees, we incorporate Lasso [Tibshirani, 1996], Ridge [Hoerl and Kennard, 1970], and ElasticNet [Zou and Hastie, 2005]. These models impose linearity constraints (L_1 and L_2 regularization), preventing the ensemble from overfitting to noise in the high-dimensional embedding space. By capturing global linear trends that tree-based models might approximate as step functions, they act as a stabilizer for the ensemble.

4.3.2 Level-2: Intra-Model Stacking

For each individual vision backbone (e.g., SigLIP 2), predictions from the six Level-1 regressors are aggregated using a Ridge Regression Meta-Learner. Unlike standard regression, we configure this meta-learner with strict physical and statistical constraints to function as a *robust expert weighting system*:

- `alpha=1.0`: Applies standard L2 regularization. Since base learners (e.g., XGBoost, CatBoost) are often highly correlated (multicollinearity), standard linear regression would produce unstable, high-variance weights. L2 regularization suppresses this instability, forcing the meta-learner to distribute "trust" more evenly among similar experts.
- `positive=True`: Enforces non-negative weights ($w_i \geq 0$). This ensures that every base model acts as a constructive signal. It prevents the ensemble from learning subtractive correction terms, which preserves the physical interpretability of the weights as "confidence scores" for each model.
- `fit_intercept=False`: Disables the bias term, forcing the regression line to pass through the origin. This is crucial for physical consistency: if all base models predict zero biomass,

the ensemble must logically predict zero. This constraint prevents the meta-learner from learning an arbitrary global offset, thereby strictly limiting its role to finding the optimal weighted combination of the provided expert opinions.

4.3.3 Level-3: Grand Fusion (ViT Ensemble)

The final prediction is generated by a top-level meta-learner that fuses the outputs of the three independent backbone streams (DINOv3, SigLIP 2, EVA-02). We maintain the same constrained configuration (`positive=True`, `fit_intercept=False`). By forcing a bias-free, non-negative weighted combination, Level-3 stacking effectively reduces variance by averaging out the distinct inductive biases of the object-centric (DINOv3), semantic-aligned (SigLIP 2), and texture-sensitive (EVA-02) streams without introducing systematic drift.

4.4 Training Strategy

Hybrid Target Transformation To address the long-tail distribution of biomass while preserving optimization stability, we implement a model-specific target transformation strategy:

- **For Linear Models:** We apply a logarithmic transformation $y' = \log(1 + y)$ to mitigate heteroscedasticity, ensuring that the residual errors are normally distributed—a strict requirement for OLS-based objectives.
- **For Boosting Models:** We utilize Max-Scaling normalization $y' = y/y_{max}$ to map targets to the $[0, 1]$ interval. Unlike the log transform, this preserves the original distributional shape, allowing gradient boosting trees (e.g., CatBoost) to learn the raw density nuances without the bias introduced by exponential restoration.

Our training pipeline is designed to address the specific biological constraints and data irregularities identified in the EDA (Section 3).

Robust Loss Functions Standard Mean Squared Error (MSE) is used for the majority of our targets (e.g., Green, Clover) where the signal-to-noise ratio is high. However, for the Dry Dead component, which exhibits extreme variance and significant labeling noise (outliers), we switch to the Huber Loss:

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \leq \delta \\ \delta(|y - f(x)| - \frac{1}{2}\delta) & \text{otherwise} \end{cases} \tag{2}$$

By setting $\delta = 1.0$, the loss transitions from quadratic to linear for large errors, preventing the model from over-fitting to noisy outliers while maintaining differentiability near zero.

4.4.1 Clover Existence Masking

As highlighted in Section 3.3, samples containing clover exhibit distinct biomass density profiles compared to ryegrass-dominant swards. To handle this heterogeneity, we train a dedicated binary classifier to detect the presence of legumes. We purposely employ a lenient classification threshold ($\tau = 0.15$). This decision is driven by the asymmetry of prediction errors: failing to identify a clover-mix sample (False Negative) results in a significantly larger penalty on the evaluation metric than falsely flagging a pure-grass sample (False Positive). By prioritizing recall, we ensure that the specialized regression head is active for all potential legume-containing samples.

4.4.2 Sequential Target Chaining

Biomass components are biologically correlated (e.g., $Total = GDM + Dead$). Exploiting this, we adopt a *Regressor Chain* strategy where predictions from biologically simpler targets are used as input features for complex ones. The training order is:

1. **Dry Clover** and **Dry Green**: Predicted first as independent foundational components.
2. **Green Dry Matter (GDM)**: Uses predictions of Clover and Green as auxiliary inputs.
3. **Dry Dead**: Uses predictions of Green and GDM.
4. **Dry Total**: The final aggregation, using predictions of GDM and Dead.

This dependency chain allows the model to "correct" its total biomass estimation based on its understanding of the sub-components.

4.4.3 Hybrid Dead Matter Estimation

Dead material (senescent vegetation) is visually ambiguous and notoriously difficult to estimate. To stabilize predictions for `Dry_Dead_g`, we implement a specialized meta-learner that ensembles three distinct estimation paths:

1. **Direct Regression:** The output from the standard sequential chain.
2. **Ratio-based Estimation:** We train a separate model to predict the *Dead Ratio* ($R_{dead} = Dead/Total$) and derive mass via $Dead = R_{dead} \times \widehat{Total}$.
3. **Mass Balance:** A derived estimate calculated as $\widehat{Total} - \widehat{GDM}$.

A Ridge Regression meta-learner [Hoerl and Kennard, 1970] optimally combines these three views, significantly reducing variance for this challenging target.

4.5 Post-processing: Matrix Reconciliation

Since our target variables are hierarchically related (e.g., $Total = GDM + Dead$, and $GDM = Green + Clover$), independent regression models often produce "incoherent" predictions where the sum of components does not match the predicted total. We address this via **Matrix Reconciliation** [Wickramasuriya et al., 2019].

4.5.1 Principle

We formulate the reconciliation as a Weighted Least Squares (WLS) optimization problem. Let $\hat{y}$ be the vector of incoherent base forecasts for all components and totals, and S be the summing matrix defining the hierarchy. We seek reconciled forecasts $\tilde{y}$ that minimize the weighted deviation from $\hat{y}$ while satisfying linear constraints:

$$\tilde{\mathbf{y}} = \mathbf{S}(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}^{-1}\hat{\mathbf{y}} \tag{3}$$

Here, W is a diagonal weighting matrix where W_{ii} corresponds to the importance weight of each target (defined in the competition metric).

4.5.2 Why it Improves Performance

This projection step mathematically guarantees that the sub-components sum to the total. Crucially, it redistributes the error terms across the hierarchy: if the $Total$ prediction is highly confident (low variance) but the $Green$ component is noisy, the algorithm will "adjust" the $Green$ component to match the robust $Total$ signal, effectively transferring information from easy-to-predict targets to harder ones. This explicitly optimizes the Weighted R^2 metric used for evaluation.

5 Results

5.1 Experimental Setup

5.1.1 Cross-Validation Strategy

Given the temporal nature of pasture growth, selecting an appropriate validation strategy is critical to avoid performance overestimation. We evaluated three strategies:

- **K-Fold Cross-Validation:** Randomly splits the data into k folds. In our dataset, multiple images are often taken from the same paddock on the same day. Standard K-Fold fails to respect this grouping, leading to *data leakage*, where highly correlated samples (e.g., adjacent images from the same sampling event) appear in both the training and validation sets.

- **Stratified Group K-Fold:** Attempts to preserve the distribution of the target variable (Biomass) across folds while keeping groups (Sampling Date) separate. However, due to the limited size of our dataset and the continuous, long-tail distribution of biomass values, stratifying often resulted in heavily imbalanced fold sizes or forced splits that violated temporal independence.
- **Group K-Fold:** We ultimately adopted Group K-Fold, grouping samples strictly by `Sampling_Date`. This ensures that all images collected on a specific date appear exclusively in either the training or the validation set. This setup mimics the real-world inference scenario where the model must generalize to future, unseen time points, providing the most rigorous and realistic performance estimate.

5.1.2 Evaluation Metric: Weighted R^2

To provide a comprehensive assessment of model performance across all biomass components, we utilize a Weighted R^2 score. This metric aggregates the coefficient of determination (R^2) for each target variable, weighted by its biological and economic importance. Specifically, we assign the weights as follows based on the competition protocol:

- **Dry Total Biomass ($w = 0.5$)**: Assigned the highest weight as the primary indicator of pasture yield.
- **Green Dry Matter (GDM) ($w = 0.2$)**: Weighted significantly to prioritize the photosynthetically active and nutritionally valuable portion of the sward.
- **Component Targets ($w = 0.1$ each)**: Dry Green, Dry Dead, and Dry Clover are assigned equal lower weights to ensure the model maintains compositional accuracy without overfitting to noisy sub-components.

The final metric is calculated as:

$$R^2_{weighted} = \sum_{i=1}^{N} w_i \cdot R_i^2 \tag{4}$$

This configuration ensures that 70% of the evaluation focus ($0.5 + 0.2$) is placed on the aggregate productive capacity of the pasture, while still enforcing structural consistency among the sub-components.

5.2 Performance Comparison

The performance of various models and configurations is summarized in Table 1. The models are evaluated using the weighted R^2 score, which accounts for the relative importance of different biomass components.

Model Architecture	Configuration Details	Weighted R^2
Baselines (CNN)	Pure EfficientNet-B0	0.47
Baselines (ViT)	DINOv2 + Lasso Regressor	0.43
	DINOv2 + ElasticNetCV	0.40
Ensemble V1	Simple Ensemble (DINOv2, SigLIP, Eva02)	0.64
Feature Engineering	SigLIP + Semantic Concept Scores	0.66
Proposed Method	**PastureNet (DINOv3 + SigLIP 2 + Chain)**	**0.70**

Table 1: Performance comparison of different model architectures. The proposed PastureNet significantly outperforms baselines and earlier ensemble versions.

5.3 Analysis of Baseline Models

To establish a performance benchmark, we evaluated representative architectures from both Convolutional and Transformer families:

- **CNN Baselines**: The EfficientNet-B0 architecture, serving as the standard for convolutional approaches, yielded a weighted R^2 score of 0.47. This result suggests that while CNNs capture local textural patterns effectively, their limited receptive field may hinder the modeling of global dependencies required for accurate biomass estimation.
- **ViT Baselines**: We investigated the off-the-shelf performance of Vision Transformers using linear probing.
  - **DINOv2 + Lasso Regressor** achieved a weighted R^2 of 0.43.
  - **DINOv2 + ElasticNetCV** demonstrated suboptimal performance with a weighted R^2 of 0.40.

  The inferior performance of these ViT-based linear baselines compared to the CNN benchmark highlights that simple linear mapping is insufficient to exploit the high-dimensional semantic representations of Foundation Models for regression tasks, necessitating the non-linear ensemble approach proposed in our methodology.

5.4 Impact of Ensemble Strategies (Ensemble V1)

An ensemble of three Vision Transformers (ViTs) DINOv2, SigLIP, and Eva02 yielded a significantly improved weighted R^2 score of 0.64. This approach demonstrated the benefit of combining multiple ViT architectures to capture a broader range of features from the input data.

5.5 Impact of Feature Engineering

The inclusion of **semantic concept scores** alongside **SigLIP** features led to a further improvement in model performance, with a weighted R^2 score of 0.66. This highlights the value of integrating additional semantic cues related to the biomass, such as greenness, clover presence, and cover.

5.6 Performance of Proposed Method (PastureNet)

The proposed method, which combines multi-modal features and a dependency chain (stacked models), achieved the highest performance with a weighted R^2 score of 0.70. This method incorporates a hierarchical ensemble strategy, leveraging both feature engineering and advanced stacking techniques. The integration of multiple sources of information visual embeddings, semantic scores, and inter-target correlations enabled this model to produce the most accurate predictions of pasture biomass.

6 Conclusion

In this work, we presented **PastureNet**, a novel cross-domain framework for pasture biomass estimation that effectively bridges the gap between raw visual data and biological interpretability. By transcending traditional CNN-based baselines, we demonstrated that leveraging the diverse inductive biases of modern Foundation Models—specifically the object-centric representations of **DINOv3**, the vision-language alignment of **SigLIP 2**, and the texture-sensitive features of **EVA-02**—yields superior generalization performance in unstructured agricultural environments.

Our study establishes three key insights. First, the integration of **Zero-shot Semantic Concept Scores** via SigLIP 2 proved instrumental. By querying images with ecological prompts (e.g., "clover", "dead grass"), we successfully injected explicit, human-interpretable domain knowledge into the regression pipeline, significantly enhancing the model's capacity to handle species heterogeneity. Second, our **Hierarchical Stacking Strategy**, reinforced by physical constraints and **Matrix Reconciliation**, ensured that predictions remained biologically consistent, effectively resolving the incoherence often observed in multi-target regression tasks. Third, the successful adaptation from the Irish source domain to the Australian target domain validates the efficacy of our transfer learning approach for data-scarce scenarios.

Future work will focus on integrating temporal dynamics directly into the architecture using video-based foundation models to capture growth rates over time, and distilling the ensemble for lightweight edge deployment on mobile devices, paving the way for real-time, accessible farmer assistance.

References

P. Albert, M. Saadeldin, B. Narayanan, B. Mac Namee, D. Hennessy, N. E. O'Connor, and K. McGuinness. Irish grass clover dataset (vistamilk), 2024. URL https://zenodo.org/records/14191859.

G. Buchsbaum. A spatial processor model for object colour perception. *Journal of the Franklin institute*, 310(1):1–26, 1980.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM, Aug. 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023. URL https://arxiv.org/abs/2303.11331.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154, 2017.

E. H. Land. The retinex theory of color vision. *Scientific American*, 237(6):108–129, 1977.

X. Li, Z. Hayder, A. Zia, C. Cassidy, S. Liu, W. Stiller, E. Stone, W. Conaty, L. Petersson, and V. Rolland. Neff-bionet: Crop biomass prediction from point cloud to drone imagery, 2024. URL https://arxiv.org/abs/2410.23901.

J. Liu, J. Xing, G. Zhou, J. Wang, L. Sun, and X. Chen. Transfer large models to crop pest recognition—a cross-modal unified framework for parameters efficient fine-tuning. *Computers and Electronics in Agriculture*, 237:110661, 2025. ISSN 0168-1699. doi: https://doi.org/10.1016/j.compag.2025.110661. URL https://www.sciencedirect.com/science/article/pii/S0168169925007677.

S. Mehdipour, S. A. Mirroshandel, and S. A. Tabatabaei. Vision transformers in precision agriculture: A comprehensive survey. *Intelligent Systems with Applications*, 29:200617, Mar. 2026. ISSN 2667-3053. doi: 10.1016/j.iswa.2025.200617. URL http://dx.doi.org/10.1016/j.iswa.2025.200617.

R. Odegua. An empirical study of ensemble techniques (bagging, boosting and stacking). 03 2019.

K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. URL https://arxiv.org/abs/2103.00020.

D. Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741:659–663, 2009.

O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, and P. Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

S. L. Wickramasuriya, G. Athanasopoulos, and R. J. Hyndman. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, 114(526):804–819, 2019.

S. Wold, M. Sjöström, and L. Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.

Y. Yang and S. Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.

X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

K. Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics gems IV*, pages 474–485. Academic Press, 1994.