

Grain Count and Size Estimation from Field Spike Images: Post-Threshing Supervision and Subspace Regression

Kotarou Saku¹, Souta Miyamoto¹, Shun Sakuma², and Masashi Nishiyama¹

¹ Graduate School of Engineering, Tottori University, Tottori, Japan

² Faculty of Agriculture, Tottori University, Tottori, Japan

Abstract. The external appearance of intact wheat and barley spikes offers indirect visual cues to the count and size of their enclosed grains. We design and construct a dataset pairing field spike images with grain counts and mean grain areas measured after removing grains from the same spikes. This post-threshing supervision enables us to evaluate compact representations of frozen foundation-model features against measured grain traits without grain-level image annotations. Image regression improves grain-count prediction over label-only predictors, whereas gains for grain size are limited. Subspaces fitted using repeated-view correspondences, which identify images of the same physical spike, retain comparable accuracy to full-feature regression with fewer trainable regression parameters.

Keywords: Grain count · Grain size · Field spike images · Post-threshing measurements · Subspace regression

1 Introduction

Learning to predict grain count and size from field spike images requires pairing photographs of intact spikes with measurements of their enclosed grains. We design and construct a dataset that provides these pairs for wheat and barley. A *spike* is the grain-bearing structure at the top of a stem. Its smaller units, called *spikelets*, contain flowers in which grains can develop [16]. Protective coverings and overlapping structures prevent the camera from directly observing all grains. Our targets are the total grain count per spike and mean grain area, measured from grains removed from that spike. The dataset supports direct regression from frozen foundation-model features and an evaluation of whether low-dimensional subspaces retain information needed to predict these measurements.

Closely related methods use *RGB images*, ordinary color photographs with red, green, and blue channels. Xu et al. [21] segment grain regions and estimate a spike’s total count by doubling a count from one side or summing counts from opposite sides; Zang et al. [22] detect grain hulls, the protective coverings around grains. They use grain-level image annotations: masks, boxes, or center points marking individual visible grains or related structures. These are spatial labels in the input image, rather than a measured count or mean grain area for the

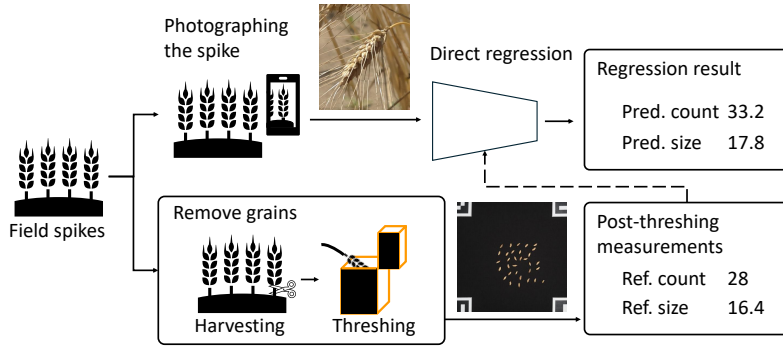


Fig. 1: Removing the grains from the photographed spike provides labels for total count and mean grain area in mm^2 , including grains hidden in the field spike image. Grain area is measured from the outlines of the removed grains. These post-threshing measurements provide post-threshing supervision through the dashed arrow; prediction uses only the field spike image.

entire spike. Occlusion makes such labels difficult to obtain, and Zang et al. [22] report confusion with coverings or structures containing no grains. Both studies report manual reference counts, but neither jointly predicts total count and mean grain area using measurements obtained after removing the grains from the same photographed spike. Measuring grain size from exposed-grain images [18, 19] or X-ray scans of harvested spikes [9] requires either direct access to the grains or penetrating imaging. Predicting size from an exterior RGB view poses a different challenge: visible pixel area depends on viewpoint and scale and does not directly reveal the sizes of hidden grains.

To address this gap, we design and construct a dataset pairing field spike images with *post-threshing measurements*: total grain count and mean grain area in mm^2 measured after removing grains from the same photographed spike. This removal process, called *threshing*, exposes hidden grains, producing one count–area pair shared by all views. We call the known grouping of images by physical spike *repeated-view correspondences*. These correspondences identify the same object across views, without specifying pixel matches. We call learning from these labels post-threshing supervision, in contrast to locating visible structures with grain-level image annotations. As Figure 1 shows, prediction requires only one field spike image of an intact spike. We compare a full-feature baseline and established subspace methods with Pooled scatter, a scatter-difference adaptation of the maximum margin criterion [13] that penalizes within-spike variation. Our three contributions are as follows:

- **Dataset design and construction with post-threshing measurements.** We design and construct a dataset linking field spike images to the count and mean grain area measured after threshing the same spike. These labels include hidden grains and support single-image prediction without grain-level image annotations. Across five crop groups, views are grouped by physical

spike for evaluation. We empirically demonstrate image-based grain-count prediction beyond global-mean and crop-group-mean label-only predictors; gains for grain size are limited.

- **A controlled baseline without grain-level image annotations.** Twelve combinations of frozen representations and regression heads evaluate direct prediction of count and mean grain area. The models learn from post-threshing measurements, without locating individual grains in the field spike image. The image regressor with the lowest grain-count error provides the reference for studying feature compression under the same prediction task.
- **Reference results for subspace regression.** We compare a full-feature baseline, established subspace methods, and Pooled scatter, which suppresses within-spike variation. With a common deep-learning-based regression architecture, the two methods using repeated-view correspondences retain comparable accuracy to 4,096-dimensional full features using 16 dimensions and about one-sixteenth as many regression parameters. These results compare accuracy and regression complexity for predicting grain count and size.

2 Related Work

2.1 Image-based grain counting and size measurement

Segmentation-based counting. Xu et al. [21] and Geng et al. [7] segment grain regions and count them after image processing; Geng et al. mainly photograph spikes removed from plants in a laboratory. Touching and occluded regions complicate grain-level image annotations, and total-count estimates use doubling from one side or summation from opposite sides. Our regressor learns total count and mean grain area directly from one field spike image using post-threshing measurements.

Point-based counting. WSG-P2PNet [6] predicts grain-center locations and counts from smartphone field spike images using point-based grain-level image annotations. Protective coverings are thin leaf-like structures that surround grains and obscure their boundaries. These coverings and viewing angle complicate annotation and counting; the authors report manually adjusting spike orientation. Our post-threshing measurements supervise total count and mean grain area in mm^2 without predicting grain-center locations or converting a visible-side count into a total count.

Foundation-model counting. AgriCountDINO [8] combines frozen DINOv3 spatial features with exemplar boxes to count and localize plants and plant organs, using point annotations during training. Its size cues describe exemplar-box dimensions rather than measured grain area. We instead predict post-threshing count and mean grain area from one spike image, without individual-grain locations or exemplar inputs, and examine compact subspaces of the global image representation.

Grain-hull detection. Zang et al. [22] train YOLOv8 variants and Faster R-CNN with boxes around grain coverings as grain-level image annotations. Occlusion and coverings that contain no grain can make detected-object counts differ from actual grain counts. Our regressor uses post-threshing supervision to predict total count and mean grain area from a field spike image. These prior studies also report manual counts; our distinction is the supervision and prediction task, not the absence of reference counts.

Grain-size measurement. SmartGrain [18] and GrainScan [19] measure grain dimensions and silhouette area from images of exposed grains. X-ray tomography measures grain number and three-dimensional size within harvested spikes [9]. These inputs reveal the grains directly or through penetrating imaging. Our model instead predicts mean grain area in mm^2 from the exterior RGB image, supervised by post-threshing measurements.

2.2 Subspace representations for regression

Subspace methods replace high-dimensional features with fewer linear combinations. Principal component analysis (PCA) retains input variance [11], whereas partial least squares (PLS) [20] and supervised PCA [1] use responses to select predictive directions. In field spike images, viewpoint and background can change features without changing the measured targets, so high variance alone need not identify useful regression information.

Sufficient dimension reduction seeks to preserve response information through conditional means [14], conditional covariances [3], or extensions for multivariate responses [12] and auxiliary information [10]. Discriminant methods offer another route: Fisher analysis relates between-class to within-class scatter, while the maximum margin criterion uses their difference [13]. Using repeated-view correspondences to treat physical spikes as classes, we adapt the maximum margin criterion using trace-normalized between- and within-spike scatter. A weighted within-spike penalty discourages variation across repeated views. We empirically assess whether the resulting subspaces retain information for predicting grain measurements.

3 Dataset and Post-Threshing Measurements

3.1 Sampling and field acquisition

Figure 2 shows the appearance variation and the measured count and mean grain area associated with each example. The dataset contains 100 physical spikes from five *crop groups*: bread wheat, durum wheat, einkorn wheat, six-row barley, and two-row barley. Each group contributes 20 *cultivars* (cultivated plant varieties), with one spike from one plant per cultivar. These categories differ in spike appearance but do not all have the same taxonomic rank. Twenty views per spike give $100 \times 20 = 2,000$ images.

	Bread wheat	Durum wheat	Einkorn wheat	Six-row barley	Two-row barley
					
Grain count (grains)	18	23	21	42	27
Grain size (mm ²)	16.1	18.8	12.8	24.4	26.1
					
Grain count (grains)	34	17	24	18	25
Grain size (mm ²)	15.3	19.3	13.2	17.9	25.9

Fig. 2: Field spike images from the five crop groups, with two examples per group. Rows show the post-threshing measurements of each photographed spike: total grain count and grain size (mean grain area in mm²).

Images were acquired from 12 May to 6 June 2025 at an experimental field in Japan, when the plants were ready for harvest. We used a Samsung Galaxy S25 between 09:00 and 12:00 under sunny or cloudy conditions. The field spike images have $3,768 \times 8,160$ pixels and span viewpoints within a 180-degree range. Calibrated camera poses are unavailable. Public release of the dataset is currently under consideration.

Post-threshing measurements provide one count–area label pair for each physical spike. Repeated-view correspondences group the twenty views of each spike, all sharing that pair even though different grains or coverings are visible. In contrast, grain-level image annotations describe the positions of individual visible structures separately in each field spike image. The dataset therefore links changing visual observations to fixed, physically measured outputs.

A *representative spike* is the prominent foreground spike selected as the target object, usually near the image center. It must be large enough to show its structure, with all its spikelets inside the frame. The input image is taken before threshing, so grains can remain hidden inside the spike’s coverings. Threshing is used only to obtain training and evaluation labels; prediction from a new image leaves the spike intact.

Table 1: Mean $\pm$ population standard deviation of measured per-spike targets by crop group. Size is mean grain area. Each physical spike contributes one observation.

Crop group	Spikes	Grain count (grains)	Grain size (mm ²)
Bread wheat	20	33.10 $\pm$ 10.11	14.26 $\pm$ 2.26
Durum wheat	20	25.30 $\pm$ 7.36	16.44 $\pm$ 2.32
Einkorn wheat	20	23.20 $\pm$ 9.10	12.41 $\pm$ 3.17
Six-row barley	20	55.70 $\pm$ 16.08	16.98 $\pm$ 3.47
Two-row barley	20	26.90 $\pm$ 6.96	21.99 $\pm$ 3.01
All spikes	100	32.84 $\pm$ 15.84	16.42 $\pm$ 4.33

3.2 Grain measurements and regression targets

After threshing each photographed spike, we imaged its recovered grains from above with a SONY DSC-RX0M2 inside an ArUco marker frame. Calibration and planar rectification yielded $3,000 \times 3,000$ -pixel images. SmartGrain [18] jointly counts grains and extracts their silhouettes; known marker distances convert silhouette areas to mm². This calibration applies to the grain photograph, not the field spike image input. The post-threshing measurements are total grain count n_s and mean grain silhouette area a_s for spike s , forming the target pair $\mathbf{y}_s = (n_s, a_s)^\top$.

Table 1 reports means and population standard deviations over 20 physical spikes per group and 100 overall. Six-row barley has the largest mean count, while two-row barley has the largest mean grain area. All views are linked to the measurements of their physical spike and kept together when splitting the data.

3.3 Evaluation protocol

We use 20-fold cross-validation, holding out five physical spikes per fold, one from each crop group. Each fold contains 95 training spikes (1,900 source images) and five validation spikes (100 source images). Every spike is held out once, and all views and augmentations of that spike stay in the same partition. This prevents the same physical object from contributing images to both training and validation. The protocol evaluates new spikes within the five known crop groups.

We report mean absolute error (MAE), averaging the absolute difference between each validation image’s prediction and its spike’s post-threshing measurement. Errors are reported in grains and mm²; views are scored individually rather than averaged before evaluation.

4 Baseline Regression Methods and Results

4.1 Common prediction task

Direct regression matches our supervision: each image has two measured outputs, total grain count and mean grain area in mm². We use a frozen foundation

model as the image encoder, retaining its pretrained visual representations while learning only the regression head from paired images and measurements. This provides a common task for comparing encoders and regression heads.

The encoder converts each image into a feature vector. A trainable multilayer perceptron (MLP) or feature adapter [5] maps this vector to separate predictions of total grain count and mean grain area, using the post-threshing targets from Section 3.2.

4.2 Feature extractors and regression configurations

We use DINOv3 [17], SteerViT [15], and TIPsv2 [2] as frozen foundation models for visual feature extraction. DINOv3 provides a strong general-purpose baseline learned by visual self-supervision. SteerViT lets a text prompt guide feature extraction toward the spike. TIPsv2 strengthens alignment between local image regions and text during pretraining, making it a complementary choice for assessing local spatial and semantic cues in grain regression. We use the TIPsv2 image encoder without a text prompt.

These encoders represent an image using feature vectors called *tokens*. A *class-token (CLS) feature* $\mathbf{c}$ summarizes the whole image, whereas *patch-token features* describe small image regions. We compare the CLS feature with the average of patch features to examine how image summarization affects prediction. Their CLS dimensions are 4,096, 768, and 3,072, respectively. SteerViT is conditioned on “the most prominent set of spikelets”.

For the CLS feature $\mathbf{c}$ and P patch-token features $\mathbf{p}_j$, we use $\mathbf{x}_{\text{CLS}} = \mathbf{c}$ or $\mathbf{x}_{\text{patch}} = P^{-1} \sum_j \mathbf{p}_j$. Patch tokens represent encoder-defined image regions. A *configuration* combines this representation with a regression module, giving $4 \times 3 = 12$ conditions across the three encoders:

- **A: CLS–MLP.** The CLS feature enters an MLP for direct nonlinear regression.
- **B: CLS–Adapter.** An adapter [5] transforms the CLS feature before linear regression.
- **C: Patch-mean–MLP.** Averaged patch tokens enter the MLP.
- **D: Patch-mean–Adapter.** An adapter transforms averaged patch tokens before linear regression.

For input width p , the MLP in A/C shares two biased linear layers, $p \rightarrow \lfloor p/4 \rfloor \rightarrow \lfloor p/16 \rfloor$, each followed by ReLU, then branches into two biased scalar output layers. The adapter in B/D uses bias-free layers $p \rightarrow \lfloor p/4 \rfloor \rightarrow p$ with an intermediate ReLU. Its output $A(\mathbf{x})$ is mixed as $0.8\mathbf{x} + 0.2A(\mathbf{x})$, followed by two biased $p \rightarrow 1$ layers. Both regressors omit dropout; output weights start at zero and output biases use default initialization.

4.3 Training conditions and reported statistics

The DINOv3 full-feature regression uses $100 \times 20 \times 10 = 20,000$ CLS vectors of width 4,096, from 100 spikes, 20 views per spike, and ten augmentation

draws per image. Features are extracted after the encoder’s final LayerNorm in evaluation mode. Each crop is resized to a square. Nine augmentation draws include rotation sampled uniformly within $\pm 15^\circ$; one omits rotation. All ten use horizontal flipping with probability 0.5 and brightness scaling in $[0.85, 1.15]$. RGB-channel normalization uses means $(0.485, 0.456, 0.406)$ and standard deviations $(0.229, 0.224, 0.225)$. Each fold uses all ten draws for training (19,000 vectors) and only the draw without rotation for validation (100 vectors); validation still includes flipping and brightness adjustment.

Each target is standardized using the mean and sample standard deviation of the 19,000 training labels, adding 10^{-8} to the denominator. We minimize equally weighted count and area Huber losses with $\delta = 1$. AdamW uses learning rate 10^{-4} , weight decay 10^{-6} , batch size 64, and 500 epochs; seed 1 is set at startup. We select the epoch with the lowest validation Huber loss, choosing the earlier epoch in a tie, and restore predictions to physical units for MAE.

4.4 Baseline prediction accuracy

Table 2(a) reports experiments to select a reference encoder and regression configuration for predicting grain count and mean grain area from field spike images. We average absolute errors over all 2,000 validation images and report their population standard deviation, using 2,000 in the variance denominator. DINOv3 has the lowest count MAE in every configuration and the lowest size MAE in A–C; SteerViT has the lowest size MAE in D.

To test whether images provide information beyond dataset or crop-group averages, we compare them with two label-only predictors in Table 2(b). Each fold uses the training mean over all 95 spikes (Global mean) or the 19 spikes in the validation spike’s true group (Crop-group mean). The crop-group predictor uses the validation spike’s true crop group, which is not provided to image models. It provides a label-only reference under perfect group identification and does not define a bound on attainable MAE. We report the mean and population standard deviation of absolute errors over 100 held-out spikes, equivalent to all 2,000 views because each spike receives one prediction. Group means reduce count MAE from 11.66 to 8.32 grains and size MAE from 3.36 to 2.28 mm². DINOv3 CLS–MLP reaches 6.12 grains and 2.24 mm², indicating a clearer image-based benefit for count than size. Image models also use validation-based epoch selection, unlike the label-only predictors.

Figure 3 shows six selected examples from configuration A (CLS–MLP), with bread wheat above two-row barley in each panel. The examples contrast low-error cases with cases showing larger discrepancies between predictions and post-threshing measurements. They illustrate prediction behavior, not the frequency of each error category.

5 Subspace Regression Methods and Results

Views of one spike share measured targets but differ in viewpoint, illumination, and background, as illustrated in Figure 4. We use these repeated-view corre-

Table 2: Baseline errors (MAE $\pm$ population standard deviation): (a) image regressors over 2,000 images; (b) label-only references over 100 spikes. Bold marks the best image-regression result for each target. Size is mean grain area.

(a) Image-based regression				
Target	Configuration	DINOv3	SteerViT	TIPsv2
Grain count (grains)	A: CLS-MLP	6.12 $\pm$ 5.37	6.87 $\pm$ 5.91	8.43 $\pm$ 7.30
	B: CLS-Adapter	6.35 $\pm$ 5.62	7.10 $\pm$ 5.76	8.43 $\pm$ 7.31
	C: Patch-mean-MLP	6.38 $\pm$ 5.44	6.92 $\pm$ 5.79	8.11 $\pm$ 6.98
	D: Patch-mean-Adapter	6.57 $\pm$ 5.49	6.89 $\pm$ 5.84	8.12 $\pm$ 6.87
Grain size (mm ²)	A: CLS-MLP	2.24 $\pm$ 2.03	2.30 $\pm$ 2.05	2.58 $\pm$ 2.34
	B: CLS-Adapter	2.16 $\pm$ 2.00	2.28 $\pm$ 2.08	2.58 $\pm$ 2.34
	C: Patch-mean-MLP	2.24 $\pm$ 2.05	2.35 $\pm$ 2.14	2.51 $\pm$ 2.25
	D: Patch-mean-Adapter	2.30 $\pm$ 2.02	2.28 $\pm$ 2.11	2.53 $\pm$ 2.28
(b) Label-only prediction errors				
Method	Count MAE (grains)		Size MAE (mm ²)	
Global mean (labels only)	11.66 $\pm$ 10.92		3.36 $\pm$ 2.75	
Crop-group mean (labels + true group)	8.32 $\pm$ 7.20		2.28 $\pm$ 2.01	

spondences to estimate within-spike variation and seek compact subspaces that suppress it while preserving differences between spikes. This could reduce regression parameters without losing predictive information. Because view changes can also reveal useful grain cues, and between-spike differences can reflect background, the benefit requires empirical evaluation.

5.1 Shared processing

We compare established subspace methods (Configurations F–H) and a subspace method adapted in this study (Configuration I) against the full-feature baseline (Configuration E). All use frozen DINOv3 CLS features of width $D = 4,096$ and the same data partitions. Let r denote the regression input dimension: 4,096 for E and 16 for F–I. The common MLP has shared layers $r \rightarrow 1024 \rightarrow 256$, ReLU after each hidden layer, and separate scalar outputs for grain count and size. Training follows Section 4.3.

For F–I, let $\mathbf{x}_i \in \mathbb{R}^D$ be the frozen CLS feature of input image i , and let $\boldsymbol{\mu}_{\mathcal{T}}$ be the training-feature mean. The matrix $\mathbf{B}_r \in \mathbb{R}^{D \times r}$ contains the r projection directions fitted from training data. Its transpose maps the centered feature to $\mathbf{z}_i \in \mathbb{R}^r$, which the MLP g_ψ , with trainable parameters ψ , uses to predict the count-size pair $\hat{\mathbf{y}}_i$:

$$\mathbf{z}_i = \mathbf{B}_r^\top (\mathbf{x}_i - \boldsymbol{\mu}_{\mathcal{T}}), \quad \hat{\mathbf{y}}_i = g_\psi(\mathbf{z}_i). \quad (1)$$

(a) Low error			(b) Near-average error			(c) High error		
Bread wheat			Bread wheat			Bread wheat		
								
	Count	Size		Count	Size		Count	Size
Measured	18	16.1	Measured	26	15.8	Measured	21	10.3
Predicted	17.9	16.4	Predicted	31.9	13.8	Predicted	33.1	16.6
Absolute error	0.1	0.3	Absolute error	5.9	2.0	Absolute error	12.1	6.3
Two-row barley			Two-row barley			Two-row barley		
								
	Count	Size		Count	Size		Count	Size
Measured	20	22.2	Measured	28	23.7	Measured	25	25.9
Predicted	20.7	21.4	Predicted	31.9	20.8	Predicted	38.4	20.3
Absolute error	0.7	0.8	Absolute error	3.9	2.8	Absolute error	13.4	5.5

Fig. 3: Prediction examples from configuration A (CLS-MLP), with two examples in each panel. Measured: post-threshing value. Predicted: estimate from one image. Count: grains. Size: mean grain area (mm^2).

Figure 5 illustrates the subspace pipeline used by Configurations F–I. Full (E) uses the original feature without centering or projection. Both pipelines predict total grain count and mean grain area in mm^2 from one field spike image.

5.2 Configurations E–I

- **E: Full.** The original 4,096-dimensional features enter the MLP directly, without mean subtraction or projection.
- **F: PCA.** PCA [11] retains directions of greatest variance in centered training features, without targets or repeated-view correspondences. We form $\mathbf{B}_r$ from the $r = 16$ leading right singular vectors of the centered training features, using an exact thin singular value decomposition (SVD) in double precision. The resulting projection is used without whitening before the shared MLP.
- **G: PLS.** PLS [20] selects directions related to the measured outputs. We use its multiple-response form, called PLS2, to fit count and mean grain area jointly from centered training features and standardized targets. We retain $r = 16$ components and apply thin QR decomposition to their input-space directions. QR denotes a factorization into a matrix $\mathbf{Q}$ with orthonormal columns and an upper-triangular matrix $\mathbf{R}$; we use $\mathbf{Q}$ as $\mathbf{B}_r$. The centered features are projected onto this basis before the shared MLP.
- **H: Regularized Fisher.** We apply regularized linear discriminant analysis (LDA) [4] using repeated-view correspondences: each physical spike is a

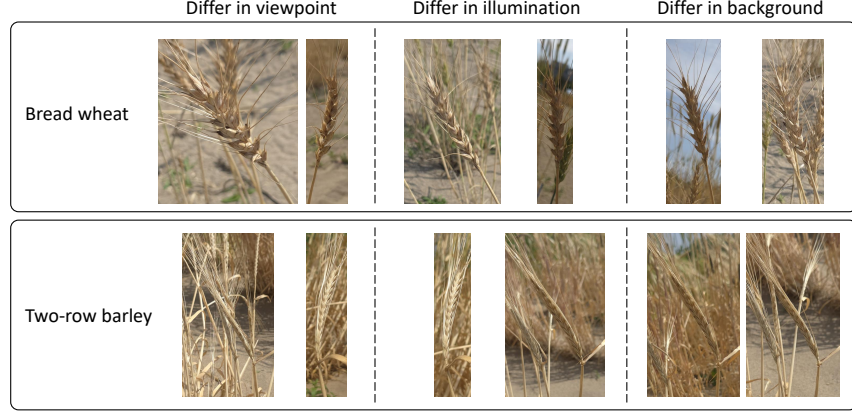


Fig. 4: Appearance variation in pairs of field spike images of the same physical spike. Rows show bread wheat and two-row barley; columns emphasize viewpoint, illumination, and background changes, which may co-occur. Each pair shares the same post-threshing count and mean grain area.

class, and its views and augmentations are samples. Here $\mathbf{S}_B$ describes differences between spike means, while $\mathbf{S}_W$ describes variation among views and augmentations of the same spike. Fisher directions favor a large ratio

$$R_{\text{Fisher}}(\mathbf{a}) = \frac{\mathbf{a}^\top \mathbf{S}_B \mathbf{a}}{\mathbf{a}^\top (\mathbf{S}_W + \alpha \tau \mathbf{I}_q) \mathbf{a}}, \quad \tau = \frac{\text{tr}(\mathbf{S}_B + \mathbf{S}_W)}{q}, \quad (2)$$

where q is the numerical rank of the centered training features, $\mathbf{a}$ is a direction in that span, and $\mathbf{I}_q$ is the identity matrix. The parameter α controls the ridge term $\alpha \tau \mathbf{I}_q$, which stabilizes the ratio’s denominator. We retain the leading $r = 16$ generalized eigen-directions with $\alpha = 1$. To obtain $\mathbf{B}_r$, we map these directions from the training span back to the original feature space and orthonormalize them by thin QR.

- **I: Pooled scatter.** We adapt the scatter-difference criterion [13] using repeated-view correspondences. Before forming $\mathbf{S}_B$, we subtract the mean of the spike means weighted by their numbers of training samples, which is equivalent in exact arithmetic to the global centering used in H. Unlike Fisher’s ratio in Eq. (2), this configuration subtracts normalized within-spike scatter from normalized between-spike scatter:

$$\mathbf{J} = T(\mathbf{S}_B) - \lambda T(\mathbf{S}_W), \quad T(\mathbf{A}) = \frac{\mathbf{A}}{\text{tr}(\mathbf{A})}. \quad (3)$$

Trace normalization puts the terms on a common scale, and λ controls the amount of within-spike variation subtracted. We use $\lambda = 10$ and retain the top $r = 16$ eigenvectors of $\mathbf{J}$ in descending algebraic eigenvalue order. We form $\mathbf{B}_r$ by mapping these eigenvectors back to the original feature space

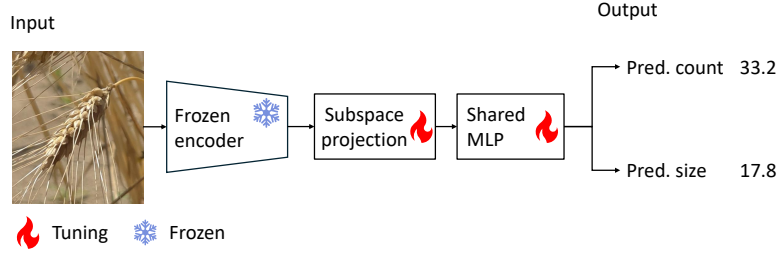


Fig. 5: Frozen CLS features are centered and projected before an MLP predicts grain count and size. Red symbols mark components fitted from training data: the projection is estimated first and stays fixed during MLP training. Post-threshing measurements supervise regression. Prediction uses one field spike image; numbers illustrate possible outputs.

Table 3: Regression results: (a) a full-feature baseline, (b) established subspace methods, and (c) a subspace method adapted in this study. All configurations use the same MLP architecture. Entries are $\text{MAE} \pm \text{population standard deviation}$ across 2,000 image-level absolute errors; r is regression input dimension. Size is mean grain area.

Configuration	r	Count (grains)	Size (mm^2)
(a) Full-feature baseline			
E: Full	4,096	6.12 ± 5.37	2.24 ± 2.03
(b) Established subspace methods			
F: PCA	16	7.32 ± 6.28	2.31 ± 2.20
G: PLS	16	6.66 ± 5.72	2.33 ± 2.10
H: Regularized Fisher	16	6.09 ± 5.44	2.21 ± 2.06
(c) Subspace method adapted in this study			
I: Pooled scatter	16	6.01 ± 5.28	2.24 ± 2.10

and orthonormalizing them by thin QR. Near-zero within-spike eigenvalues are discarded for stability; zero scatter gives zero penalty.

5.3 Results and discussion

Table 3 tests whether compact representations preserve full-feature accuracy and how methods using repeated-view correspondences compare with PCA and PLS. Entries give the mean and standard deviation of absolute errors over 2,000 validation images, rather than variation between fold means. All four subspace methods use 16 regression inputs. The reported Regularized Fisher and Pooled scatter settings were selected retrospectively by count MAE from the explored settings. Epochs are selected by validation loss.

Table 4: Grain-count results for I: Pooled scatter by crop group. Relative error is MAE divided by the measured group mean.

Crop group	Mean (grains)	MAE (grains)	Relative (%)
Bread wheat	33.10	6.41	19.36
Durum wheat	25.30	4.29	16.96
Einkorn wheat	23.20	6.05	26.08
Six-row barley	55.70	8.98	16.13
Two-row barley	26.90	4.30	15.98
All spikes	32.84	6.01	18.29

Table 5: Grain-size results for I: Pooled scatter by crop group. Size is mean grain area in mm^2 ; relative error is MAE divided by the measured group mean.

Crop group	Mean (mm^2)	MAE (mm^2)	Relative (%)
Bread wheat	14.26	1.89	13.28
Durum wheat	16.44	1.90	11.53
Einkorn wheat	12.41	2.08	16.79
Six-row barley	16.98	2.93	17.24
Two-row barley	21.99	2.40	10.93
All spikes	16.42	2.24	13.65

In Table 3(b), PCA (F) preserves feature variance without targets or repeated-view correspondences. PLS (G) uses targets without explicitly modeling repeated-view correspondences and improves count prediction over PCA, but not size prediction. Regularized Fisher (H) has lower count and size errors than both.

Pooled scatter (I) in Table 3(c) has the lowest count MAE, 6.01 grains, and a size MAE of 2.24 mm^2 . With 16 inputs, it retains full-feature accuracy and improves both errors over PCA and PLS. The small differences between H and I do not establish superiority of one scatter objective. Their lower errors than PCA and PLS suggest that using repeated-view correspondences helps retain predictive information that variance or target correlation alone does not prioritize.

Compared with the crop-group-mean label-only predictor in Table 2(b), Pooled scatter (I) reduces count MAE from 8.32 to 6.01 grains, but size only from 2.28 to 2.24 mm^2 . Image-based gains are therefore clearer for count. Within-group size prediction remains unestablished, and validation-based selection on a single-year dataset limits generalization claims.

Using 16 instead of 4,096 MLP inputs reduces regression parameters from 4,458,242 to 280,322, about one-sixteenth. Preprojecting one fold’s 20,000 vectors at 32-bit precision would require 1.28 MB rather than 327.68 MB, excluding the basis, mean, and metadata. Our implementation retains full features and projects during forward passes; encoder computation is unchanged.

Tables 4 and 5 compare crop-group accuracy for Pooled scatter (I), which has the lowest overall count MAE. Each group has 20 spikes and 400 validation

images. Relative error is MAE divided by the measured group mean in Table 1, not the mean of per-spike percentage errors. Prediction difficulty varies across crop groups. Six-row barley has the largest absolute errors for both traits, while einkorn wheat has the largest count error relative to its group mean. Both groups also have high relative size errors. Overall accuracy therefore conceals group-specific weaknesses. Future subspaces and regression objectives should address these differences while preserving useful variation within each group.

6 Conclusions

We design and construct a dataset pairing field spike images with post-threshing count and mean grain area. Image regression lowers count MAE to 6.12 grains from 11.66 and 8.32 for global-mean and crop-group-mean label-only predictors; size gains are small. At 16 dimensions, Regularized Fisher and Pooled scatter, both fitted using repeated-view correspondences, retain comparable full-feature accuracy with about one-sixteenth as many regression parameters.

Future work will develop subspaces and regressors tailored to crop-group difficulties, especially within-group grain-size variation. We will also evaluate generalization independently and compare annotation schemes on matched spikes for accuracy and annotation effort.

References

1. Barshan, E., Ghodsi, A., Azimifar, Z., Zolghadri Jahromi, M.: Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition* **44**(7), 1357–1371 (2011)
2. Cao, B., Chen, K., Maninis, K.K., Chen, K., Karpur, A., Xia, Y., Dua, S., Dabral, T., Han, G., Han, B., Ainslie, J., Bewley, A., Jacob, M., Wagner, R., Ramos, W., Choromanski, K., Seyedhosseini, M., Zhou, H., Araujo, A.: TIPSV2: Advancing vision-language pretraining with enhanced patch-text alignment. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 29325–29335 (2026)
3. Cook, R.D.: SAVE: A method for dimension reduction and graphics in regression. *Communications in Statistics—Theory and Methods* **29**(9–10), 2109–2121 (2000)
4. Fisher, R.A.: The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**(2), 179–188 (1936)
5. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: CLIP-Adapter: Better Vision-Language Models with Feature Adapters. *International Journal of Computer Vision* **132**(2), 581–595 (2024)
6. Geng, Q., Xu, X., Ma, X., Li, L., Xu, F., Gao, B., Ma, Y., Huang, J., Yang, J., Yao, X.: WSG-P2PNet: A deep learning framework for counting and locating wheat spike grains in the open field environment. *Computers and Electronics in Agriculture* **235**, 110314 (2025)
7. Geng, Q., Zhang, H., Gao, M., Qiao, H., Xu, X., Ma, X.: A rapid, low-cost wheat spike grain segmentation and counting system based on deep learning and image processing. *European Journal of Agronomy* **156**, 127158 (2024)

8. Guo, S., Li, X., Arsova, B., Scharr, H., Salvi, S.: AgriCountDINO: Parameter-efficient exemplar-guided counting and localization in agriculture. arXiv preprint arXiv:2609.29460 (2026)
9. Hughes, A., Askew, K., Scotson, C.P., Williams, K., Sauze, C., Corke, F., Doonan, J.H., Nibau, C.: Non-destructive, high-content analysis of wheat grain traits using X-ray micro computed tomography. *Plant Methods* **13**(1), 76 (2017)
10. Hung, H., Liu, C.Y., Lu, H.H.S.: Sufficient dimension reduction with additional information. *Biostatistics* **17**(3), 405–421 (2016)
11. Jolliffe, I.T., Cadima, J.: Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**(2065), 20150202 (2016)
12. Li, B., Wen, S., Zhu, L.: On a projective resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association* **103**(483), 1177–1186 (2008)
13. Li, H., Jiang, T., Zhang, K.: Efficient and robust feature extraction by maximum margin criterion. In: *Advances in Neural Information Processing Systems*. vol. 16 (2003)
14. Li, K.C.: Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**(414), 316–327 (1991)
15. Ruthardt, J., Gaur, M., Ramanan, D., Tapaswi, M., Asano, Y.M.: Steerable visual representations. In: *Computer Vision – ECCV 2026*. *Lecture Notes in Computer Science*, vol. 17071, pp. 357–374. Springer (2026)
16. Sakuma, S., Schnurbusch, T.: Of floral fortune: tinkering with the grain yield potential of cereal crops. *New Phytologist* **225**(5), 1873–1882 (2020)
17. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3. *Transactions on Machine Learning Research* (2026)
18. Tanabata, T., Shibaya, T., Hori, K., Ebana, K., Yano, M.: SmartGrain: High-throughput phenotyping software for measuring seed shape through image analysis. *Plant Physiology* **160**(4), 1871–1880 (2012)
19. Whan, A.P., Smith, A.B., Cavanagh, C.R., Ral, J.P.F., Shaw, L.M., Howitt, C.A., Bischof, L.: GrainScan: a low cost, fast method for grain size and colour measurements. *Plant Methods* **10**(1), 23 (2014)
20. Wold, S., Sjöström, M., Eriksson, L.: PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58**(2), 109–130 (2001)
21. Xu, X., Geng, Q., Gao, F., Xiong, D., Qiao, H., Ma, X.: Segmentation and counting of wheat spike grains based on deep learning and textural feature. *Plant Methods* **19**(1), 77 (2023)
22. Zang, H., Wang, Y., Wang, S., Ren, S., Yang, Y., Zhang, J., Zhao, Q.: Study on automatic detection of wheat spike grain number based on deep learning. *Frontiers in Plant Science* **17**, 1724501 (2026)