Maize ear sensing for on-farm yield predictions

Pedro Cisdeli[1] Gustavo Nocera Santiago[1] German Mandrini[1]
Ignacio Ciampitti[1]
[1]Purdue University - Dept. of Agronomy
915 Mitch Daniels Blvd, West Lafayette, IN 47907
{pmagalh,gnoceras,mandrini,iciampit}@purdue.edu

Abstract

In maize (Zea mays L.), early yield prediction is a concept of great interest to breeders and agronomists. Limited studies have tried to leverage field-scale imaging data to perform such crop yield predictions. For this purpose, this study aims to provide a complete data pipeline that uses single-view depth and RGB image data to extract morphological traits of maize ears (length, width, and volume), to forecast yield in a non-destructive approach. A depth camera from a smartphone device was used to acquire imagery data, and those images were used to train the YOLOv12n-seg model for segmenting the maize ears. The segmentation masks were then used to cut out the background from the point clouds. These point clouds then underwent a further filtering process to remove the remaining noise. The length and width of the ear were extracted using a geometric computational approach, while the volume was predicted using the deep learning model developed for this project, the Ear Volume Network (EVNet). Lastly, a yield prediction was obtained by using the ear morphological traits as input in the gradient-boosting decision trees. The segmentation task performed well under adverse environmental conditions such as occlusion, noise and variable lighting. The EVNet accurately predicted ear volumes under ideal scenarios (RMSE = 28.91 ml). Likewise, the yield forecast step demonstrated high accuracy in both ideal (RMSE = 13.90 g) and real (RMSE = 24.12 g) scenarios. The results of this study highlight the potential of using new technologies to predict yield under field conditions.

1. Introduction

Maize (Zea mays L.) is a widely used crop for numerous industries that maintain food security [20]. Early predictions of yield components can help farmers and breeders assess field quality and make data-assisted management decisions [3]. Despite the recognized importance of early yield forecasting, the existing literature lacks robust methods that leverage on-field image data to extract morphological traits directly related to yield components. For other cereal crops, such as sorghum, this gap has been explored by Santiago et al. [19]. In contrast, most existing approaches for maize focus on grain classification and quality assessment [25]. Consequently, these methods typically do not utilize on-field data, often requiring manual labor such as husk removal, threshing maize ears, or setting up controlled imaging environments [14]. A more closely related approach is presented by Gao et al. [9], who leveraged a pose estimation method that can estimate maize ears dimensions with on-field images. However, their method significantly differs from this study, as they primarily focused on accurately predicting maize ear position and pose, with ear dimensions indirectly derived from 3D bounding boxes.

To the best of our knowledge, no study or tool explores ear volume estimation, a trait directly related to yield components, and did not use these morphological characteristics for yield forecasting or other practical applications. Therefore, the objective of this study is to develop a robust pipeline that utilizes on-field data to extract maize ear morphological traits, particularly length, width, and volume, enabling precise plant-level yield forecasts.

2. Materials and Methods

During the 2024 growing season, multiple locations around Manhattan, Kansas, were visited during the R4 to R6 stages to capture a diverse range of maize ears images. These locations included demonstration plots, experimental fields, and commercial production fields. The experimental fields included plots with varying nitrogen fertilizer application rates.

At these sites, we measured the ears' maximum length using a measuring tape and maximum width using a digital caliper. Additionally, we collected single-view field depth measurements of 120 maize ears using an Intel RealSense Depth Camera D435i. This advanced camera system provided both RGB images and depth outputs, allowing for a detailed assessment of ear morphology. To illustrate this process, we present example outputs from this camera system, including the original RGB image (Figure 1a) and the depth output for that same image (Figure 1b). Lastly, the maize ears were extracted so that their volume could be measured in the laboratory using a water displacement method.

In addition to the depth imaging, we gathered 947 RGB images of maize ears using standard smartphone cameras, such as the example shown in Figure 1c. This smartphone image dataset is publicly available in the CornDepth repository [2], and was purely used to facilitate ear segmentation, therefore no other morphological trait was collected.

The deep learning models developed in this study were trained on the following GPUS: NVIDIA GeForce RTX 4070 laptop and NVIDIA RTX 1000 Ada laptop.

2.1. Overview

The overall pipeline consisted of collecting single-view depth data of maize ears directly in the field. Maize ears were first segmented from the images by extracting regions of interest using a trained segmentation model. The segmented areas were then used to crop corresponding point clouds obtained from the depth camera. These segmented point clouds provided the basis for extracting morphological traits, specifically ear length, and width. A neural network model, detailed later in section 2.5, estimated ear volume by taking the segmented point cloud, ear dimensions (length and width), and phenological stage as inputs. Finally, length, width, and volume were utilized to forecast yield per plant. This entire pipeline was programmed and executed in Python 3.11. [24]

2.2. Image Segmentation of Maize Ears

We utilized the CornDepth dataset [2] with an addition of images taken from the RGB module of the depth camera to train the YOLOv12n-seg [22] model. Furthermore, the labeling process was done through Roboflow [5], while training was performed locally on multiple devices, as described in section 2.

The segmentation masks from this step were used to trim the point cloud acquired from the depth camera, resulting in a single-view 2.5D reconstruction of the maize ear.

2.3. Point Cloud Outlier Removal

We tested two methods for filtering out the background noise. Both groups include the entire set of point clouds, with Group 2 undergoing an additional filtering step. The first method (Group 1) involved performing a Statistical Outlier Removal followed by a Radius Outlier Removal. The second method (Group 2) contains an additional Small Cluster Removal step to clean the remaining noise. Each of these techniques was implemented as follows:

Let (1) represent the original maize ear point cloud.

$$P = \{p_1, p_2, \dots, p_N\},\ p_i \in \mathbb{R}^3 \tag{1}$$

Statistical Outlier Removal (SOR)

For each point $p_i \in P$, we compute the average distance d_i to its k nearest neighbors:

$$d_i = \frac{1}{k} \sum_{j \in \mathrm{NN}(p_i)} \|p_i - p_j\|_2,$$

where $\mathrm{NN}(p_i)$ denotes the set of the k nearest neighbors of p_i. The global mean μ and standard deviation σ of $\{d_i\}_{i=1}^N$ are computed. A point p_i is retained as an inlier if:

$$d_i \leq \mu + \alpha\sigma,$$

where $k = 20$ and $\alpha = 2$ in our implementation.

Radius Outlier Removal (ROR)

For each point p_i, we define the neighborhood:

$$N_r(p_i) = \{p_j \in P \mid \|p_i - p_j\|_2 \leq r\},$$

with $r = 0.05$. A point p_i is retained if the number of neighbors satisfies:

$$|N_r(p_i)| \geq m,$$

with $m = 16$.

The intermediate cleaned point cloud P' after applying both SOR and ROR is defined as:

$$P' = \{p_i \in P \mid d_i \leq \mu + 2\sigma \ \wedge\ |N_r(p_i)| \geq 16\}.$$

Group 1: Without Small Cluster Removal

The final cleaned point cloud in this case is simply:

$$P_{\text{clean}}^{(1)} = P'.$$

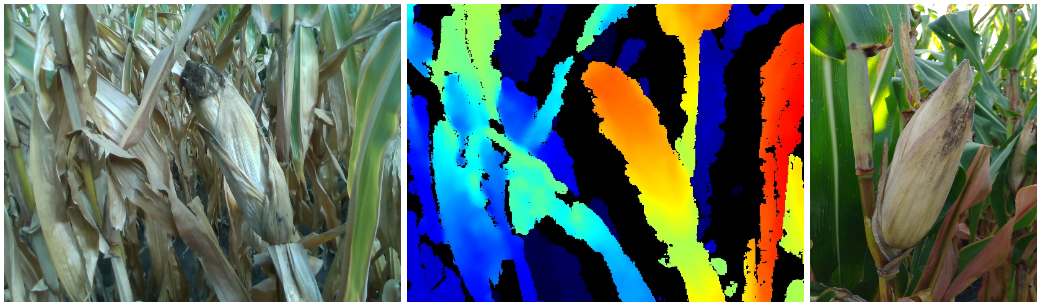
Group 2: With Small Cluster Removal (SCR)

In this approach, P' is further partitioned into clusters using DBSCAN [6] with parameters $\varepsilon = 0.02$ and minimum points $m_{\text{DB}} = 30$:

$$P' = \left(\bigcup_{i=1}^{K} C_i \right) \cup N,$$

where C_i (for $i = 1, \dots, K$) are the resulting clusters, and N represents noise points (not assigned to any cluster). The largest cluster is selected as the final cleaned point cloud:

$$P_{\text{clean}}^{(2)} = \arg\max_{C_i} |C_i|.$$



(a) Example of an image from D435i's RGB module. (b) Example of a Depth Map from D435i. Colors represent proximity. (c) Example of a smartphone RGB image.

Figure 1. On-field data of maize ears obtained via a D435i depth camera and a regular smartphone camera.

Summary of Outlier Removal Methods

The final cleaned point clouds can be summarized as follows:

- **Without Small Cluster Removal:**

$$P^{(1)}_{\text{clean}} = \{p_i \in P \mid d_i \le \mu + 2\sigma \ \wedge\ |N_r(p_i)| \ge 16\}.$$

- **With Small Cluster Removal:**

$$P^{(2)}_{\text{clean}} = \arg\max_{C_i} |C_i|,$$

$$\text{where} \quad P' = \{p_i \in P \mid d_i \le \mu + 2\sigma \ \wedge\ |N_r(p_i)| \ge 16\}.$$

2.4. Geometric Extraction of Maize Ear Length and Width

Using the same definition of a point cloud P (1), the algorithm to extract the maize ear dimensions consists of the following.

Principal Component Analysis (PCA): Compute the centroid of P:

$$\bar{p} = \frac{1}{N}\sum_{i=1}^{N} p_i.$$

Center the point cloud by defining

$$p'_i = p_i - \bar{p}, \quad i = 1, \ldots, N.$$

The covariance matrix C is given by

$$C = \frac{1}{N-1}\sum_{i=1}^{N} p'_i\,(p'_i)^T.$$

Perform an eigen-decomposition of C to obtain eigenvalues $\lambda_1 \ge \lambda_2 \ge \lambda_3$ and corresponding eigenvectors v_1, v_2, v_3. In our implementation, we define:

$$e_Z = v_1, \quad \text{(largest principal axis)}$$
$$e_X = v_2, \quad \text{(second principal axis)}$$
$$e_Y = v_3, \quad \text{(third principal axis)}$$

Rotation to Principal Axes: We construct the rotation matrix R whose columns are the reordered eigenvectors:

$$R = \begin{bmatrix} e_X & e_Y & e_Z \end{bmatrix}.$$

The rotated point cloud is then

$$P_R = \{R^T(p_i - \bar{p}) \mid p_i \in P\}.$$

In this rotated coordinate system, the new z-axis aligns with the largest principal component.

Maize Ear Length: The ear length is defined as the extent along the z-axis:

$$L = \max_{p \in P_R} z(p) - \min_{p \in P_R} z(p),$$

where $z(p)$ denotes the z-coordinate of the point $p \in P_R$.

Maize Ear Width: To compute width, first consider the xy-projection of P_R:

$$Q = \{(x, y) \mid (x, y, z) \in P_R\}.$$

Outliers are removed by retaining only points within the 5^{th} and 95^{th} percentiles for each coordinate:

$$Q' = \{(x, y) \in Q \mid x_{5\%} \le x \le x_{95\%} \wedge y_{5\%} \le y \le y_{95\%}\},$$

where $x_{5\%}$, $x_{95\%}$, $y_{5\%}$, and $y_{95\%}$ are the corresponding percentile values.

For each angle $\theta \in \{0°, 5°, 10°, \ldots, 175°\}$, define the 2D rotation matrix:

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Rotate the filtered points:

$$Q'_\theta = \{R_\theta \begin{pmatrix} x \\ y \end{pmatrix} \mid (x, y) \in Q'\}.$$

The width along angle θ is computed as:

$$w(\theta) = \max_{q \in Q'_\theta} q_1 - \min_{q \in Q'_\theta} q_1,$$

where q_1 denotes the first coordinate of the rotated point. The width W is then defined as

$$W = \max_{\theta \in \{0^\circ, 5^\circ, \ldots, 175^\circ\}} w(\theta).$$

Unit Conversion: Finally, depending on the desired measurement units, the computed length L and width W may be scaled accordingly.

2.5. Ear Volume Network: Predicting Ear Volume with Deep Learning

The Ear Volume Network (EVNet) is a deep learning architecture designed to predict maize ear volume. It receives two numeric values—the ear's length and width—along with a point cloud representation of the ear. Note that since the point cloud is captured from a single viewpoint, it is effectively a 2.5D representation that captures only the visible surfaces. Each point in the cloud is represented as a 4-dimensional feature vector $[x, y, z,\ stage_{\text{encoded}}]$, where the spatial coordinates (x, y, z) capture geometric information and the categorical variable is the phenological stage (e.g., "R4", "R6"), which was converted to a numeric value.

2.5.1. Point Cloud Sampling Strategy

Inspired by [18], we employed a Farthest Point Sampling (FPS) method to downsample the original point cloud and accelerate computations. Given the point cloud P (1), the goal is to select a subset $S \subset P$ of k points that are maximally dispersed throughout P. The FPS algorithm proceeds as follows:

1. **Initialization:**
   - Randomly select an initial point p_{i_0} from P and add it to the set S.
   - Initialize a distance array $\{d_i\}_{i=1}^N$ with a large constant value for each point in P.
2. **Iterative Selection:** For each subsequent selection until $|S| = k$:
   - For every point $p_i \in P$, compute the Euclidean distance $\|p_i - p_j\|_2$ from the most recently selected point $p_j \in S$.
   - Update the distance d_i as:

     $$d_i = \min\left(d_i, \|p_i - p_j\|_2\right).$$

   - Select the point p_{i^*} with the maximum distance:

     $$p_{i^*} = \arg\max_{p_i \in P} d_i.$$

   - Add p_{i^*} to S.
3. **Termination:** The algorithm terminates once k points have been selected. The set S then represents the downsampled point cloud P_{FPS}.

2.5.2. Network Architecture and Training

The proposed neural network architecture, **EVNet** balances training and inference performance with simplicity and accuracy. The design integrates Dynamic EdgeConv [26] layers with global pooling and numeric feature fusion to predict maize ear volume from point cloud data.

Each point sampled via FPS (section 2.5.1) is represented by a four-dimensional feature vector comprising its normalized spatial coordinates $[x, y, z]$ and the phenological stage encoded numerically $(stage_{\text{encoded}})$.

Dynamic EdgeConv layers [26] are employed to extract hierarchical local geometric features from these point representations. The first EdgeConv layer dynamically constructs a k-nearest neighbor graph (with $k = 20$) and maps the initial 4-dimensional input into an intermediate 64-dimensional feature space. Subsequently, a second EdgeConv layer captures more abstract local geometric relationships, transforming these intermediate features from 64 to 128 dimensions.

The output from the second EdgeConv layer is then aggregated using global mean pooling to derive a single global descriptor for each maize ear's point cloud. This global descriptor is concatenated with additional numeric inputs, specifically the maize ear's length and maximum width, forming a comprehensive feature representation.

Finally, the combined features undergo regression via a Multi-Layer Perceptron (MLP) comprising fully connected layers, batch normalization, ReLU activations, and dropout layers to mitigate overfitting. The final output of the network is a scalar prediction of the maize ear volume.

Training details: The training conditions for EVNet were 400 epochs using the Adam optimizer with a learning rate of 10^{-3}. The loss function employed was mean squared error (MSE). Models were trained using batch sizes of 8 on graphics cards described in Section 2.

Two versions of the model were selected for this study. The first, **EVNetIdeal**, was trained under ideal data conditions using ground truth ear length and width measurements. This approach isolates the performance of the network architecture by eliminating error propagation from preceding methods. In contrast, **EVNetReal** was trained using ear dimensions derived from the geometric extraction process described in Section 2.4, thereby assessing the performance of the complete pipeline in real-world scenarios.

2.6. Forecasting Grain Yield per Plant

We evaluated two gradient-boosting decision tree models (GBDT), XGBoost [1] and CatBoost [4], for predicting maize ear grain yield. The input features included numeric variables—ear length, width, and volume—as well as a categorical variable representing the phenological stage. The model output was the total grain dry weight per ear, serving as an estimate of grain yield per plant.

This forecasting step was also evaluated using both the EVNetIdeal and EVNetReal scenarios described in Section 2.5.2.

3. Results

3.1. Performance of Maize Ear Segmentation

The YOLOv12n-seg model achieved high segmentation accuracy under diverse environmental conditions, as displayed in Figure 2. Additionally, Table 1 summarizes the precision, recall, and mean average precision (mAP) obtained at epoch 250 for bounding boxes (BBox) and segmentation masks. The segmentation method yielded a precision and recall of 95.83% and a mAP@0.5 of 98.60%, while the masks outperformed the bounding boxes in mAP over IoU thresholds from 0.5 to 0.95. Lastly, the segmentation step averaged 1.112 seconds per image.

Method	Precision	Recall	mAP@0.5	mAP@0.5:0.95
BBox	0.9583	0.9582	0.9860	0.7720
Mask	0.9583	0.9582	0.9860	0.7981

Table 1. Final YOLOv12n-seg performance metrics at epoch 250.

3.2. Comparing Point Cloud Outlier Removal Groups

The two outlier removal groups were compared based on visual outcomes, computational performance and Root Mean Squared Error (RMSE) between the predicted and ground truth volumes using **EVNetIdeal**. As shown in Figure 3, the Group 2 method, which included the SCR step, effectively eliminated visually notable outliers, excluding points from the background and overall noise, resulting in cleaner point clouds. However, this additional step did not improve EVNet predictions; it made them worse. As presented in Table 2, the point clouds from Group 1 outperformed the ones from Group 2 by approximately 1.97 milliliters (ml) in RMSE when passed as input for inferences with **EVNetIdeal**.

	Group 1	Group 2
RMSE (ml)	28.9067	30.8700

Table 2. Comparison of Root Mean Squared Error (RMSE) by Group of outlier removal methods, calculated by ground truth and the volumes predicted by **EVNetIdeal**.

3.3. Accuracy of Length and Width Estimations

To evaluate the performance of the dimensions estimation using the geometric approach described in Section 2.4, we defined three metrics: RMSE, Mean Absolute Percentage Error (MAPE), and Coefficient of Determination R^2. As shown in Table 3, the MAPE for length and width measurements are 9.37% and 14.57%, respectively. The RMSE values are 2.9659 cm for length and 0.7843 cm for width. The coefficient of determination R^2 is 0.1968 for length and 0.4300 for width.

Metric	Length	Width
MAPE (%)	9.37	14.57
RMSE (cm)	2.9659	0.7843
R^2	0.1968	0.4300

Table 3. Performance metrics for maize ear dimensions estimation using the geometric approach.

3.4. EVNet Volume Prediction Performance

Figures 4 and 5 illustrate the volume prediction performance of our models. The **EVNetIdeal** model, trained using ground truth maize ear measurements, achieved an R^2 of 0.90 (RMSE = 25.12 ml, MAPE = 5.65%) on the combined training and validation datasets, and an R^2 of 0.88 (RMSE = 28.91 ml, MAPE = 6.90%) on the independent test dataset. In contrast, the **EVNetReal** model, which utilized ear dimensions from the geometric extraction process, showed lower performance metrics. On the combined training and validation datasets, it achieved an R^2 of 0.82 (RMSE = 35.79 ml, MAPE = 8.06%), while performance declined substantially on the independent test dataset with an R^2 of 0.48 (RMSE = 75.15 ml, MAPE = 18.33%).

3.5. Accuracy Grain Yield Forecast per Plant

As described in Section 2.6, we evaluated yield forecasting under two scenarios: an ideal scenario using ground truth dimensions and volume predictions from **EVNetIdeal**, and a real-world scenario using geometrically derived dimensions and volume predictions from **EVNetReal**.

Figure 6 presents the grain yield forecasting results per plant. The best-performing model under the ideal scenario was XGBoost. Its yield predictions strongly correlated with actual measured grain yields ($R^2 = 0.96$), achieving an RMSE of 13.90 grams (g) and a MAPE of 10.07%. For the real scenario, however, CatBoost resulted in superior metrics. The model predicted grain yield with an $R^2 = 0.89$, an RMSE of 24.12 g, and a MAPE of 14.85%.

4. Discussion

4.1. Segmentation Performance

Analyzing field data is challenging due to noise, occlusion, variable lighting, and different poses of the regions of interest [13]. Despite these challenges, our segmentation method

(a) Double detection and segmentation.
(b) Dry husk, typical from later development stages.
(c) Crop canopy casting shadow upon the depth camera.
(d) Sun directly affecting the depth camera.

Figure 2. Empirical results: YOLOv12n-seg masks under different environment conditions.

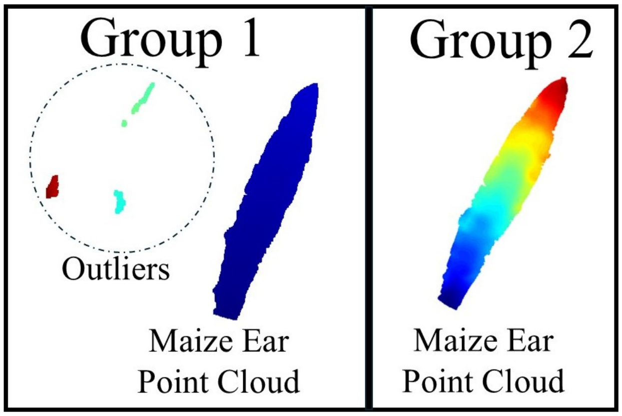


Figure 3. Comparison of the same maize ear point cloud processed with two different outlier removal methods. Group 1 (left) shows the point cloud processed with Statistical and Radius Outlier Removal methods, which failed to eliminate all noise and left remaining outliers behind (circled). Group 2 (right) shows the same maize ear processed with the same methods from Group 1 followed by an additional Small Cluster Removal step, resulting in complete outlier removal. Color map represents proximity to the camera but is not relevant for evaluating the outlier removal methods in this context.

performed well across various field conditions (Figure 2). The YOLOv12-seg model achieved high precision and recall and a strong mean Average Precision as shown in Table 1. Additionally, the segmentation masks slightly outperformed bounding box detections at higher IoU thresholds, providing better ear shape details.

Unlike previous methods that depend on controlled setups for improved image quality and noise reduction [10][29], our approach works directly with raw field images. The inference speed of 1.112 seconds per image is practical for field applications, and with further optimization using lighter models [12], real-time processing could be achievable.

4.2. Point Cloud Processing

The slight underperformance of 1.97 ml in RMSE of Group 2 compared to Group 1 was a surprise since performing all of the outlier filtering steps (SOR, ROR, and SCR) yielded the best visual representation of the maize ear. This unexpected result can be explained by modern deep neural network architectures being robust to noise [17]. Aggressive processing methods have also been shown to cause information loss [23], and DBSCAN parameters can significantly impact clustering outcomes [8]. Therefore, while background noise was removed, the additional filtering steps could have also affected the shape of the edges that represented closer dimensions to ground truth. In the future, we would like to test and investigate this result further. Nonetheless, avoiding the additional filtering step saved computational resources.

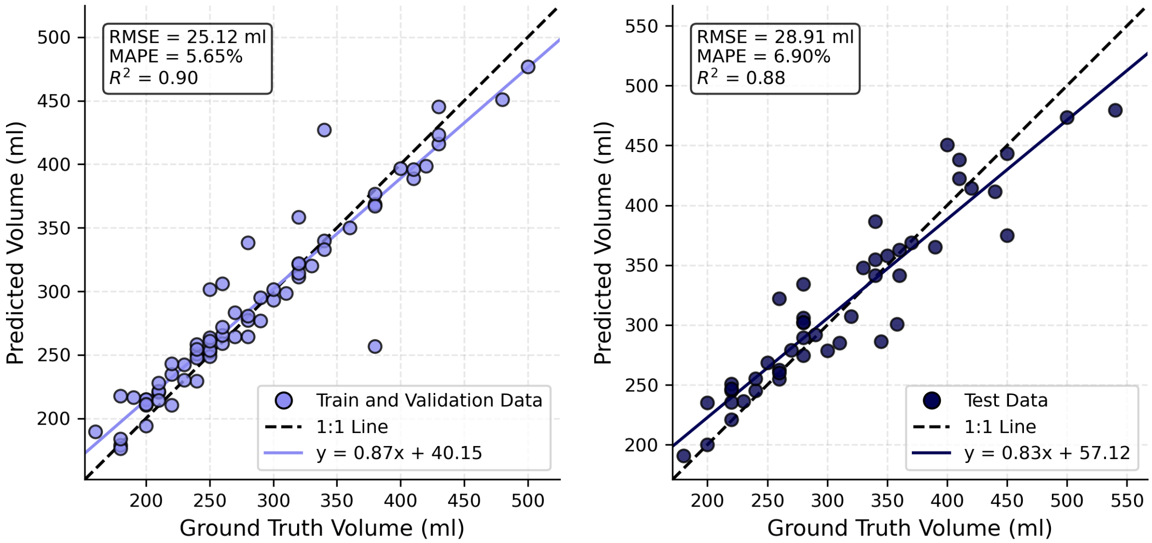
4.3. Dimension Estimation Challenges

Table 3 highlights the underlying limitations of our current approach. The low coefficient of determination values are an indication of this, despite acceptable RMSE values for length and width.

The most evident cause can be traced to the work with high-noise environments [13], such as uncontrolled field data. This factor caused issues to the point cloud structure that directly affected its shape. Additionally, single-view approaches are a staple in 3D reconstruction studies, but they are often accompanied by adversities [16] like shape ambiguity and, in our case, wrong or incomplete geometry information. Furthermore, the higher accuracy in width estimation suggests that principal component analysis effectively identifies the ear's main axis, but struggles with endpoint determination. Lastly, maintaining the depth camera within its precision range is particularly challenging inside dense maize fields.

To improve this step, specialized feature extraction methods [28] [15] could be adopted, as they offer robust strategies for addressing high noise levels and shape ambiguities in point cloud data.

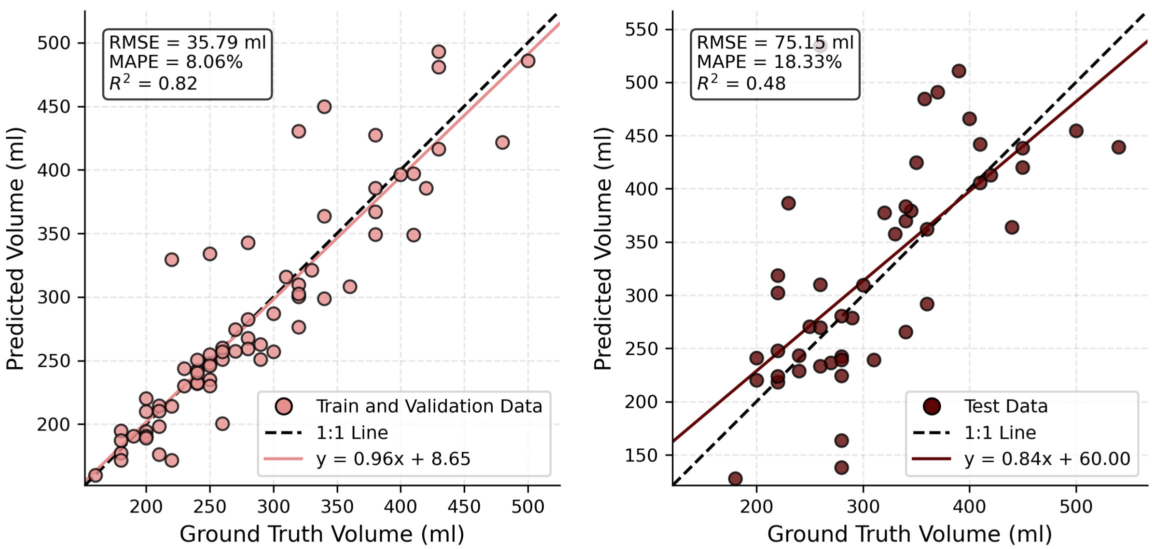
Regardless of these limitations and the propagation of



(a) Model performance on training and validation data.

(b) Model performance on test data.

Figure 4. EVNetIdeal volume prediction performance.



(a) Model performance on training and validation data.

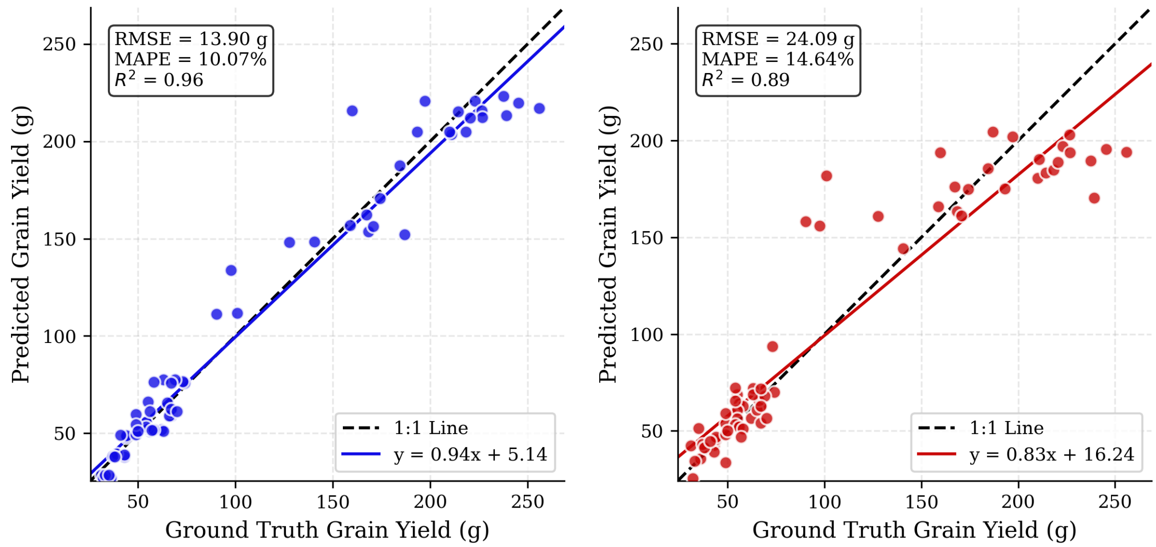
(b) Model performance on test data.

Figure 5. EVNetReal volume prediction performance.

errors derived from this step in the pipeline, the results of both **EVNet** models presented in Figures 4b, 5b and the yield forecasting process (Figure 6b) demonstrate that the method is adequate for real-world applications.

4.4. Volume Prediction Performance

The **EVNet** model presented promising results on test data while being trained with a small dataset of only 120 point clouds and ear dimensions. In the ideal scenario that used the ground truth maize ear dimensions for training and testing to avoid propagating error from section 2.4, the model's performance on unseen data was notably accurate (Figure 4b). This highlights the success of the architectural design, as indicated by precise predictions on both the training and testing data, suggesting good generalization with minimal overfitting [11]. Additionally, Feng et al. [7] showed that incorporating different types of data, such as images and LiDAR, helps improve model prediction accuracy. This supports future exploration of ways to enhance the model and



(a) XGBoost Grain yield prediction results for maize ears under ideal scenario (ground truth ear dimensions and volume obtained from predictions of EVNetIdeal).

(b) CatBoost Grain yield prediction results for maize ears under real scenario (dimensions calculated with geometric methods and volume obtained from predictions of EVNetReal).

Figure 6. Maize ear yield forecasting results.

highlights the importance of integrating numerical and categorical data as inputs.

As discussed in section 4.3, the reason for the poor performance from **EVNetReal** was the underlying accuracy issue with the geometric method. Even then, as noted by [21], a larger dataset could help improve results since the model shows acceptable convergence potential, as evidenced in Figure 5b. This is supported by the metrics, which, while indicating moderate prediction accuracy, suggest that with more training data, the model could potentially achieve better generalization and reduced error rates.

4.5. Yield Forecasting Potential

Our results demonstrate the potential of accurate maize yield prediction using both ideal and real scenarios. Even with the dimensional estimation challenges, the real scenario (Figure 6b) achieved impressive accuracy with CatBoost, closely approaching the ideal scenario's (Figure 6a) performance. This highlights the robustness of our methodology, even with imperfect input data. This is particularly significant as our approach moves away from methods that are highly dependent on manual labor and intricate imaging setups [27], eliminating the need for harvesting ears, removing husks, or threshing grains to forecast yield at the plant level.

Future development will focus on achieving the same results on earlier phenological stages such as R2 and R3, expanding the dataset size to improve the pipeline generalization [21], and exploring an end-to-end solution in order to avoid many steps for yield forecasting.

5. Conclusion

This study developed a pipeline for extracting morphological traits from maize ears using on-field data to forecast yield at the plant level. Our work represents one of the first attempts to create a field-to-forecast system that processes raw field imagery without requiring controlled environments or destructive sampling. The segmentation component performed well (95.83% precision and recall, 98.60% mAP@0.5) despite challenging field conditions with variable lighting, occlusion, and noise. The dimension extraction component showed limitations (R^2 of 0.20 for length and 0.43 for width), yet downstream models compensated for these imperfections surprisingly well. Volume prediction worked effectively when accurate dimensions were provided (R^2 = 0.88, RMSE = 28.91 ml), validating our approach. Yield forecasting remained robust even with real-world data, with CatBoost achieving solid results (R^2 = 0.89, RMSE = 24.12 g) that approached ideal conditions (R^2 = 0.96, RMSE = 13.90 g). We plan to test our methods at earlier phenological stages (R2 and R3), expand our dataset size, and explore more streamlined solutions. Addressing dimension extraction limitations through specialized feature extraction methods will be a primary focus in our upcoming work.

References

[1] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, page 785–794. ACM, 2016. 4
[2] Pedro Henrique Cisdeli and Gustavo Santiago. Corndepth dataset, 2024. 2
[3] Monica F. Danilevicz, Philipp E. Bayer, Farid Boussaid, Mohammed Bennamoun, and David Edwards. Maize yield prediction at an early developmental stage using multispectral images and genotype data for preliminary hybrid selection. *Remote Sensing*, 13(19), 2021. 1
[4] Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support, 2018. 4
[5] B. Dwyer, J. Nelson, T. Hansen, et al. Roboflow (version 1.0) [software], 2024. Computer Vision Software. 2
[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page 226–231. AAAI Press, 1996. 2
[7] Di Feng, Christian Haase-Schutz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2021. 7
[8] Junhao Gan and Yufei Tao. Dbscan revisited: Mis-claim, un-fixability, and approximation. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, page 519–530, New York, NY, USA, 2015. Association for Computing Machinery. 6
[9] Yuliang Gao, Zhen Li, Qingqing Hong, Bin Li, and Lifeng Zhang. Corn pose estimation using 3d object detection and stereo images. *Computers and Electronics in Agriculture*, 231:110016, 2025. 1
[10] Juan M. Gonzalez, Nayanika Ghosh, Vincent Colantonio, Francielly de Cássia Pereira, Ricardo A. Pinto Jr., Chase Wasson, Kristen A. Leach, and Marcio F. R. Resende Jr. EarCV: An open-source, computer vision package for maize ear phenotyping. *The Plant Phenome Journal*, 5(1):e20055, 2022. 6
[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org. 7
[12] Rachel Huang, Jonathan Pedoeem, and Cuixian Chen. Yololite: A real-time object detection algorithm optimized for non-gpu computers. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2503–2510, 2018. 6
[13] Derek Kelly, Avimanyou Vatsa, Wade Mayham, Linh Ngô, Addie Thompson, and Toni Kazic. An opinion on imaging challenges in phenotyping field crops. *Machine Vision and Applications*, 27(5):681–694, 2016. 5, 6
[14] Xiaoming Li, Baisheng Dai, Hongmin Sun, and Weina Li. Corn classification system based on computer vision. *Symmetry*, 11(4), 2019. 1
[15] Zheng Liu, Xiaopeng Xin, Zheng Xu, Weijie Zhou, Chunxue Wang, Renjie Chen, and Ying He. Robust and accurate feature detection on point clouds. *Comput. Aided Des.*, 164(C), 2023. 6
[16] Pedro O. Pinheiro, Negar Rostamzadeh, and Sungjin Ahn. Domain-adaptive single-view 3d reconstruction. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7637–7646, 2019. 6
[17] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016. 6
[18] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *CoRR*, abs/1706.02413, 2017. 4
[19] Gustavo N. Santiago, Pedro H. Cisdeli Magalhaes, Ana J. P. Carcedo, Lucia Marziotte, Laura Mayor, and Ignacio A. Ciampitti. Deep learning methods using imagery from a smartphone for recognizing sorghum panicles and counting grains at a plant level. *Plant Phenomics*, 6:0234, 2024. 1
[20] Bekele Shiferaw, Boddupalli M. Prasanna, Jonathan Hellin, and Marianne Bänziger. Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. *Food Security*, 3(3):307–327, 2011. 1
[21] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017. 8
[22] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025. 2
[23] A. Ullrich and M. Pfennigbauer. Noisy lidar point clouds: impact on information extraction in high-precision lidar surveying. In *Laser Radar Technology and Applications XXIII*, page 106360M. International Society for Optics and Photonics, SPIE, 2018. 6
[24] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. 2
[25] Noel Varela, Jesus Silva, Omar Bonerge Pineda, and Danelys Cabrera. Prediction of the corn grains yield through artificial intelligence. *Procedia Computer Science*, 170:1017–1022, 2020. The 11th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 3rd International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops. 1
[26] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph.*, 38(5), 2019. 4
[27] Cedar Warman, Christopher M. Sullivan, Justin Preece, Michaela E. Buchanan, Zuzana Vejlupkova, Pankaj Jaiswal, and John E. Fowler. A cost-effective maize ear phenotyping platform enables rapid categorization and quantification of kernels. *The Plant Journal*, 106(2):566–579, 2021. 8
[28] Christopher Weber, Stefanie Hahmann, and Hans Hagen. Methods for Feature Detection in Point Clouds. In *Visualization of Large and Unstructured Data Sets - Applications in Geospatial Planning, Modeling and Engineering (IRTG*

1131 Workshop), pages 90–99, Dagstuhl, Germany, 2011. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. 6

[29] Shuo Zhou, Xiujuan Chai, Zixuan Yang, Hongwu Wang, Chenxue Yang, and Tan Sun. Maize-IAS: A maize image analysis software using deep learning for high-throughput plant phenotyping. *Plant Methods*, 17(1):48, 2021. 6