3D Reconstruction-Based Seed Counting of Sorghum Panicles for Agricultural Inspection

Harry Freeman[1*], Eric Schneider[1*], Chung Hee Kim[1], Moonyoung Lee[1], George Kantor[1]

***Abstract*— In this paper, we present a method for creating high-quality 3D models of sorghum panicles for phenotyping in breeding experiments. This is achieved with a novel reconstruction approach that uses seeds as semantic landmarks in both 2D and 3D. To evaluate the performance, we develop a new metric for assessing the quality of reconstructed point clouds without having a ground-truth point cloud. Finally, a counting method is presented where the density of seed centers in the 3D model allows 2D counts from multiple views to be effectively combined into a whole-panicle count. We demonstrate that using this method to estimate seed count and weight for sorghum outperforms count extrapolation from 2D images, an approach used in most state of the art methods for seeds and grains of comparable size.**

I. Introduction

With recent advancements in data-driven computer vision, agriculture is widely adopting image-based techniques to efficiently inspect vast quantities of crops. Automated crop inspections, which were not easily done before, enable farmers and breeders to make real-time decisions to manage pests, disease, and drought, and to automate laborious tasks such as phenotyping and yield prediction. In this paper we propose a computer vision-based method for non-destructive counting of sorghum seeds for early forecasting of yield. Accurate forecasting is valuable for sorghum breeding programs, as it would allow faster decision-making on variant suitability, which could expedite the current five-year breeding process [1]. Seed count would be a valuable phenotypic trait, but it is currently not possible to sample in a non-destructive way.

Common deep learning techniques for visual agricultural inspection include disease classification [2], object detection [3], and fruit counting [4]. In contrast to the large and separated fruits typically inspected, we investigate seed modelling on a sorghum panicle, which is more challenging from a computer vision perspective. The seeds are much smaller than typically studied crops (average diameter 3.3mm), making them difficult to detect and track. In addition, there is significantly more occlusion due to dense packing and clutter from husks. Although there has been work on 2D image based instance counts for other crops [5], [6], [7], it is still difficult to obtain a high accuracy count with sorghum. We create a high-quality 3D model from stereo using a semantic landmark-based reconstruction approach, which we use to count seeds. Using our proposed method, we acquire a more realistic count than 2D image-based approximations.

[1]Carnegie Mellon University Robotics Institute, PA, USA {hfreeman, franzs, chunghek, moonyoul, kantor}@cs.cmu.edu

*These authors contributed equally to this work.

The specific contributions of this paper are:

- A novel 3D reconstruction method that utilizes seeds as semantic landmarks in 2D and 3D to produce a high quality model of a sorghum panicle.
- A new metric for assessing point cloud reconstruction quality in the absence of ground truth.
- A novel method for extracting seed counts from point clouds by extending 2D image processing techniques into 3D, along with an algorithm to identify local maxima in a point cloud.
- A dataset of sorghum stereo images with camera poses, a subset labeled with instance segmentation of seeds. [1].

II. Related Work

There has been a significant amount of recent work dedicated towards reconstruction and counting in agricultural settings. Mapping and estimating the yield of mangoes in occluded environments using a FasterRCNN segmenter is presented in [8] and [9]. Mapping and counting grapes in 3D by fitting spheres to point clouds is presented in [10]. While these methods work in their respective domains, they do not extend well to sorghum where the seeds are smaller and the level of density and occlusions are higher, making them harder to consistently segment and fit shapes to.

Phenotyping during the breeding process is laborious if done manually and important for expedited decision making. As a result, several works address automated phenotyping using robots. In [11], UAV images taken early in the season are used to predict end of season above-ground biomass. [12] and [13] show that images collected from mobile robots can be used to assess plant height and stalk size more easily than manual collection. Component traits such as these are used in genetic research to improve biomass yield. Our work explores seed counting, which was not possible at the resolution of these systems.

There has also been relevant work in estimating seed counts for smaller crops from single 2D images. Counting rice and soybeans with density maps using convolutional neural networks is addressed in [5] and [6] respectively. However, the rice and beans have been stripped from the plant and laid out such that there are few occlusions. Density maps have also been used to count corn kernels on the cob, where the final count is proportional to the density map count as a result of corn's symmetric shape [7]. Similarly, [14] uses a KD-Forest approach to detect grapes in clusters using keypoint-based features, and estimates yield using a scale

[1]https://labs.ri.cmu.edu/aiira/resources/

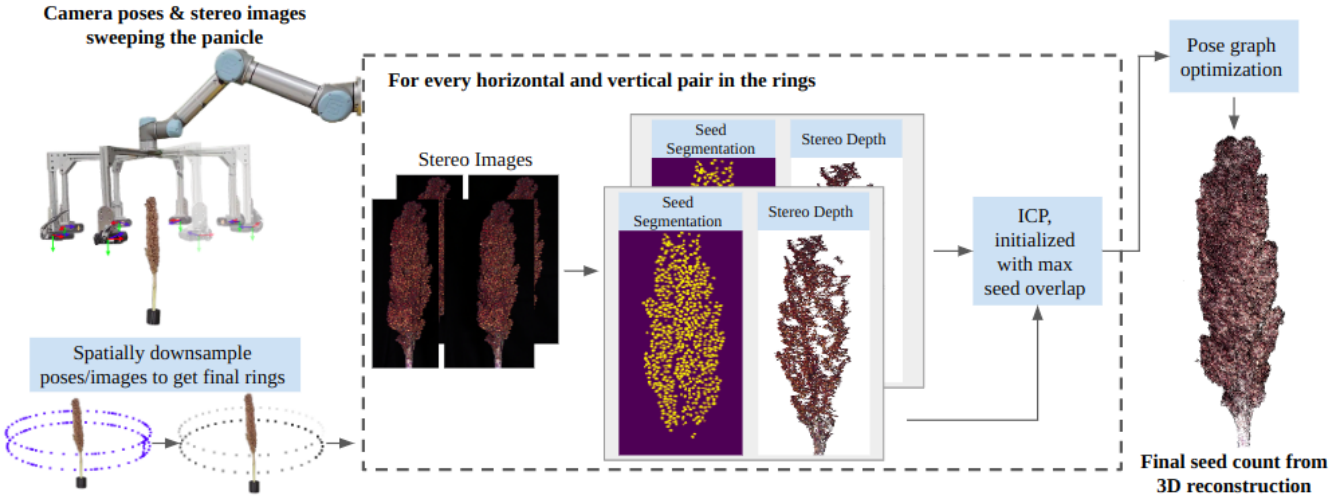


Fig. 1: 3D Reconstruction pipeline for the sorghum stalk

factor. These methods do not adapt well to sorghum seeds because of the asymmetric nature of sorghum panicles.

With regards to reconstruction in agriculture, most works are focused on larger maps and fields rather than single plants. For example, large field maps of different crops are reconstructed in [15] and [16]. Although localized views of flowers and vines are captured in [17] and [18], they do not get a complete 360° scan. Apple orchard rows are reconstructed in [19] by merging views from opposing sides using cylinders fit to trunks. This does not adapt well to sorghum as the stems are too small to effectively fit.

III. Methodology

A. Overview

In order to generate a high-quality 3D model of sorghum panicles, we set up an automatic data collection process by attaching a flash stereo camera [20] to the wrist of a UR5 robot arm. The robot follows a circular trajectory around the panicle as shown in Fig. 1. Although these images were captured in the lab, in-field image capture from an arm mounted on a mobile base would also be possible. From all images taken, we spatially downsample to only consider images $\mathbf{I}_i \in \mathbb{I}$ and poses $\mathbf{T}_i \in \mathbb{T}$ in the shape of a double ring, spaced 5cm apart, as seen in Fig. 1, leaving roughly 85 images per panicle. We then use RAFT-Stereo [21] to construct point clouds for each frame. Using Iterative Closest Point (ICP) on segmented seeds alone, we construct a pose graph that best aligns all point clouds to create the final high quality point cloud $\mathbf{C}$. Lastly, given $\mathbf{C}$, seed masks are combined between all images $\mathbf{I}_i$ to obtain a final seed count.

B. Instance Segmentation

Given a stereo image pair, we acquire a 3D point cloud semantically labeled with individual sorghum seeds. This is achieved through instance segmentation on 2D images, which is projected onto the 3D points. Our seed segmentation is based on a CenterMask [22] instance segmentation network. Seeds were hand segmented from 10 different 1440×1080 sorghum images across different species. After training, seed masks are projected onto the 3D point cloud.

C. Global Registration

We jointly register point clouds of a sorghum panicle imaged from different viewpoints via pose graph optimization, as presented in [23]. One challenge is that the point clouds are dense, and ICP optimization on the full cloud performed poorly due to bad correspondences, an example of ICP falling into local minima. Instead we choose a limited set of high-quality points in the cloud and run ICP only on those points, somewhat analogous to doing optical flow on higher quality landmarks like SIFT features. Semantically segmented seeds with high confidence are identified based on their inference scores. The set of good seeds from image $\mathbf{I}_i$ are then used as node $\mathbf{P}_i$ in the pose graph, and pose graph optimization is performed using the Levenberg-Marquardt algorithm [24]. An example of a reconstructed panicle is shown in Fig. 2(b).

We observe that using camera poses from arm kinematics to initialize ICP yields poor results on the scale of seeds. This is due to error in extrinsic camera parameters, despite using a standard hand-eye calibration process. Hence, we refine the camera pose priors by maximizing seed mask overlap. The seed masks of two neighboring nodes $\mathbf{P}_i$ and $\mathbf{P}_j$ are projected into a common image frame, at the average pose between $\mathbf{T}_i$ and $\mathbf{T}_j$. We search for the pixel shifts that yield maximum intersection over union (IOU) of seed masks as shown in Fig. 3. The *No Shift Maximize* ablation test in Fig. 9 shows that this IOU maximization improves reconstruction.

D. Counting

In order to obtain a final seed count, we use the 3D model to ensure that a single true seed segmented in multiple images will be counted only once. The following new 3D counting method is proposed to perform this combination of 2D counts while handling the close proximity of neighboring seeds, spurious detections, and noise in the point cloud.

First, 3D seed centers are clustered using density-based spatial clustering (DBSCAN) [25], as shown in Fig. 4(d).

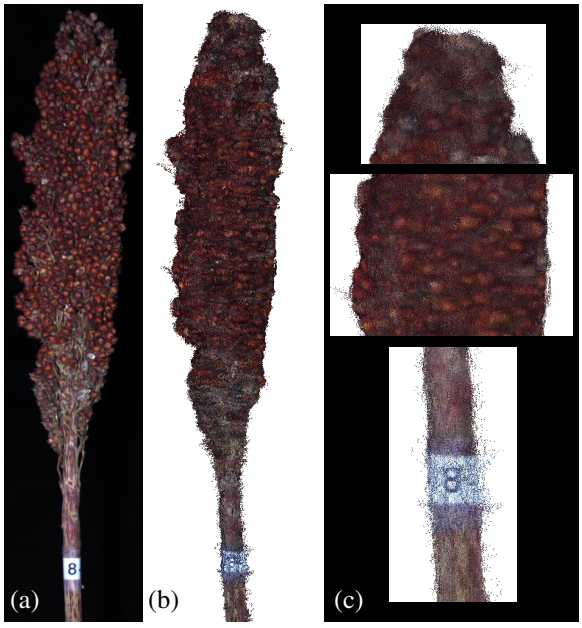


Fig. 2: Example reconstruction results. (a) one of the original RGB images, (b) the colorized point cloud, (c) zoomed view of the colorized point cloud at the stem, mid-body, and tip. Some points of interest include the "8" on the stem label, and the body outline which matches the RGB outline well.

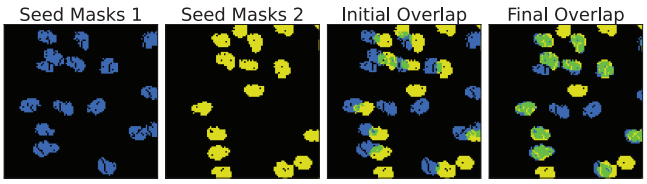


Fig. 3: Matching mask structure with maximum IOU. Seed masks 1, seed masks 2, and their intersection are colored blue, yellow, and green in respective order.

Seeds are then counted in each cluster, which breaks up the large problem of counting thousands of seeds across the whole panicle as there are often fewer than four true seeds in a given cluster. Next, we adapt the concept of 2D image smoothing and apply it to 3D point clouds. In image processing, a 2D Gaussian filter smooths an image by calculating a weighted average around each pixel's neighborhood. In our method, each seed center in the cluster is treated as a unit-impulse, and each impulse is smoothed around a volume of space using a 3D Gaussian filter. An example of this density map can be seen in Fig. 5(c).

Once the density values of the mask cloud points are created, the final step to calculate the number of seeds in each cluster is to find the local maxima within a defined radius. This is a type of non-maximal suppression (NMS) on the density values. Each local maximum corresponds to a unique seed and is the location of the seed's center. Algorithm 1 has a detailed description of the proposed algorithm.

Once all local maxima are found for each cluster, the total number of maxima becomes the final seed count. An example of this process on a single cluster can be seen in Fig. 5.

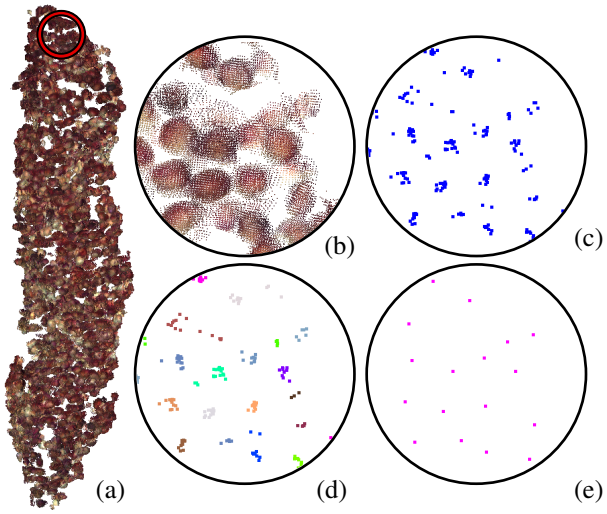


Fig. 4: (a) An example of a final point cloud seed mask, (b) zoomed seeds, (c) seed centers, (d) seed centers clustered with DBSCAN, and (e) final seed sites.

Algorithm 1 Find Cluster Local Maxima

Inputs: $M \in \mathbb{R}^{k\times 3}$, $r \in \mathbb{R}^1$, D
Output: $L \in \mathbb{R}^{l\times 3}$

1: $U \leftarrow M, L \leftarrow \emptyset$
2: **while** $U \neq \emptyset$ **do**
3: $\quad s \leftarrow \max_{p\in U} D(p)$
4: $\quad U \leftarrow U \setminus \{s\}$
5: $\quad R \leftarrow \{p \mid p \in M \text{ and } 0 < \|s - p\| < r\}$
6: $\quad m \leftarrow \max_{p\in R} D(p)$
7: $\quad$ **if** $D(s) > D(m)$ **then**
8: $\quad\quad L \leftarrow L \cup \{s\}$
9: $\quad U \leftarrow U \setminus \{p \mid p \in R\}$

Summary: All points p in the cluster's cloud M are initialized as unvisited (line 1). The unvisited point with the highest density value $D(p)$ is iteratively retrieved (line 4), and if that point has a higher density value than all neighbors in a defined radius r, it is a local maximum (lines 6-9). All points within the radius are marked as visited (line 10), and the process repeats until all points have been visited.

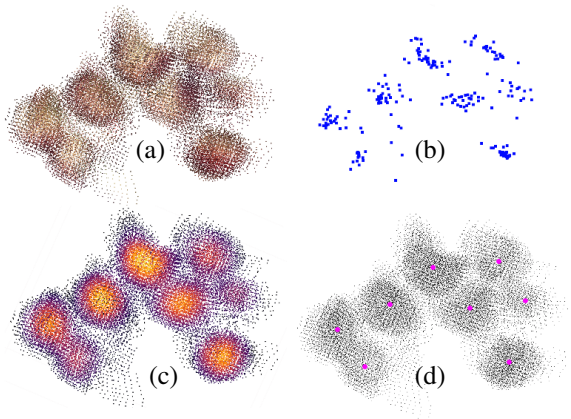


Fig. 5: (a) Seed point cloud that has been put in a single cluster by DBSCAN, (b) seed centers from individual images, (c) seed points weighted by seed-center density, and (d) local maxima (pink) that have been chosen as seeds.

E. 3D Reconstruction Metric

Several prior works [26], [27] discuss quantitative reconstruction evaluation in the absence of ground truth, but they require that the final output to evaluate against is a mesh. Our reconstruction method produces a dense point cloud, so we developed and validated a novel cloud-only rendering based method for assessing reconstruction quality in the absence of ground truth. We compare a small circle of pixels sampled from an RGB image $\mathbf{I}_i \in \mathbb{I}$, centered on a sampled seed, against a projected render of the same seed made using the full reconstructed cloud. A sampling function λ is defined so that K seeds are sampled per image along the center of the vertical axis where the projections are cleanest. This method experimentally indicates relative levels of noise in the reconstructed point clouds by comparing rendered sections to the original RGB images.

To validate this framework, noise was purposefully introduced in the camera poses $\mathbf{T}_i$ when creating the reconstructed cloud. A variety of comparisons were run on pairs of RGB image patches and the corresponding rendered patches. The strongest response to introduced noise came from normalized grayscale image patches. Both the mean-squared error (MSE) on image gradients, and the Structural Similarity [28] (SSIM) on image Laplacians responded well to the introduced noise, shown in Fig. 6. In order to settle on these two metrics we checked all combinations of RGB/grayscale, normalized/unnormalized, and intensity/gradient/Laplacian. Two examples of our image-to-render comparison with their corresponding MSE and SSIM scores are shown in Fig. 7.

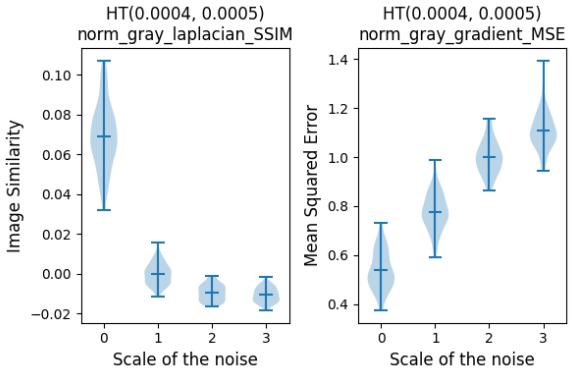


Fig. 6: Response of chosen metrics to introduced noise. Noise took the form of homogeneous transforms, with (x, y, z) translations drawn from a Gaussian $\mathcal{N}(0, \sigma = \text{scale} * 0.4\text{mm})$ and rotational noise (θ, ϕ, ψ) drawn from a Gaussian $\mathcal{N}(0, \sigma = \text{scale} * 0.0005\text{rad})$. After the random transforms were applied the cloud was recalculated and rendered.

Our reconstruction quality metrics "$\alpha\beta$-MSE" and "$\alpha\beta$-SSIM" are defined as follows. For each image, λ samples K seeds from $\mathbb{S}_i$, where $\mathbb{S}_i$ are the seeds in image $\mathbf{I}_i$. For a sampled seed $s_{ik} \in \mathbb{S}_i$, the image patch α_{ik} and rendering of the point cloud β_{ik} are generated, both of which are grayscaled and normalized. The MSE and SSIM of α_{ik} and β_{ik} are calculated, then averaged over all seeds and panicles.

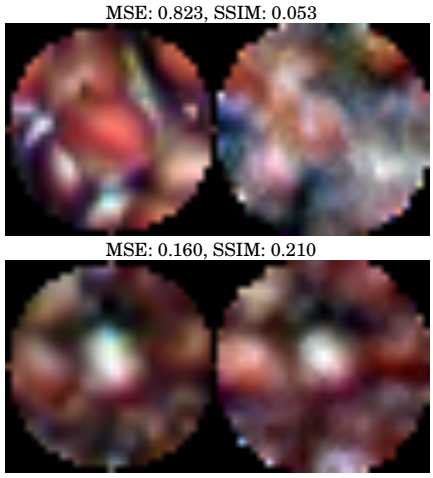


Fig. 7: Qualitative examples of the reconstruction metrics, drawn from low and high scoring samples. On the left are image patches from the original RGB images, on the right are image patches rendered from the reconstructed point cloud. All patches are normalized so each channel has min/max values of 0/255.

$$\text{MSE}_{ik} = \frac{1}{N} \sum_{pixels} \left[\nabla\alpha_{ik} - \nabla\beta_{ik}\right]^2$$

$$\alpha\beta\text{-MSE} = \frac{1}{P} \sum_{p} \frac{1}{IK} \sum_{i} \sum_{k \in \lambda(\mathbb{S}_i)} \text{MSE}_{ik}$$

$$\text{SSIM}_{ik} = \text{SSIM}\big(\mathcal{L}(\alpha_{ik}), \mathcal{L}(\beta_{ik})\big)$$

$$\alpha\beta\text{-SSIM} = \frac{1}{P} \sum_{p} \frac{1}{IK} \sum_{i} \sum_{k \in \lambda(\mathbb{S}_i)} \text{SSIM}_{ik}$$

Here ∇ is the image gradient, $\mathcal{L}$ is the image Laplacian, $\frac{1}{IK}\sum_i \sum_{k\in\lambda(\mathbb{S}_i)}$ indicates an average over sampled seeds in all images, and $\frac{1}{P}\sum_p$ indicates an average over all panicles.

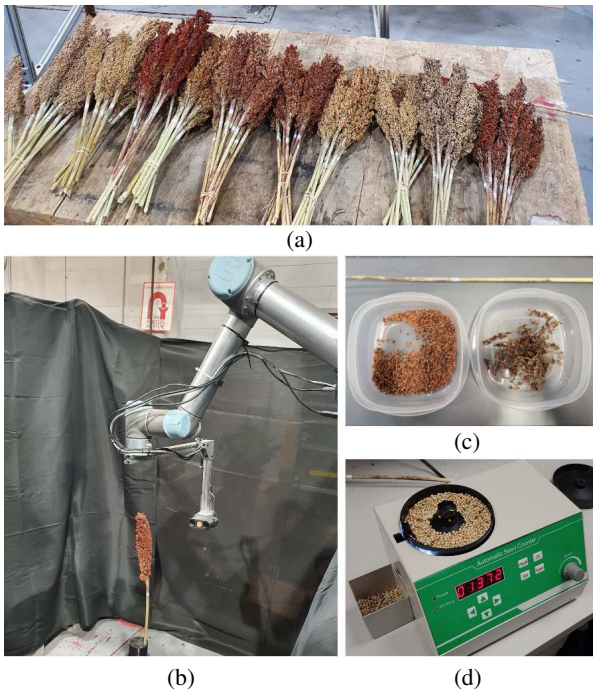
IV. Experimental Results

A. Dataset

Our dataset consists of stereo images of 100 sorghum panicles collected in two 360° rings using a custom stereo camera [20]. There were 10 panicles from 10 different species as seen in Fig. 8(a). To evaluate our proposed method, we manually stripped panicles (Fig. 8(c)) and counted all seeds using an automatic seed counting machine[2] (Fig. 8(d)), which serves as ground truth. The process of stripping seeds, removing husks, and counting took significant effort, on average 40 minutes per panicle, which reinforces the usefulness of an automated method for yield estimation. Random errors in the seed count include some lost seeds that fell off panicles between image collection and hand-counting. Affecting the count in the opposite direction, some unremoved husks were counted as seeds by the counting machine despite manual efforts to separate seeds from husks. We expect the effect on the ground truth to be small. The stereo images, camera poses, human-labeled seed segmentations, panicle weights, and human-counted seed counts can be found in our dataset[3].

[2]Wadoy Automatic Seeds Counter, Sly-C

[3]https://labs.ri.cmu.edu/aiira/resources/



(a) (b) (c) (d)

Fig. 8: (a) 100 sorghum panicles from 10 different sorghum species. (b) Our data collection system, a stereo camera attached to the UR5 robot arm. (c) Seeds were manually stripped and (d) counted using a seed counting machine.

B. 3D Reconstruction Quality

We assess the effectiveness of our approach with ablation tests using the reconstruction metrics described in III-E. Below references to "$\alpha\beta$-MSE" and "$\alpha\beta$-SSIM" are referring to these specific operations on image and rendered patches. Note that growing $\alpha\beta$-MSE (error) and dropping $\alpha\beta$-SSIM (similarity) both indicate a worse match. Fig. 9 shows results of ablation and comparison tests on reconstruction quality.

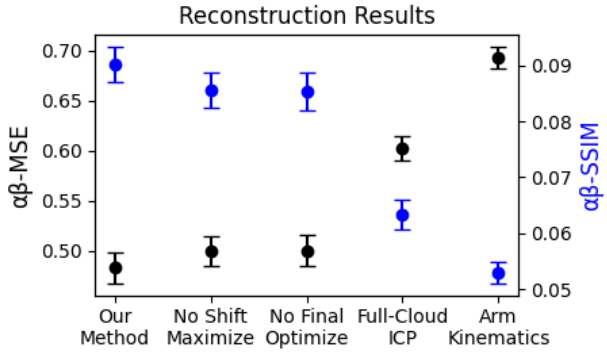


Fig. 9: Noise metric results showing growing error and dropping similarity for reconstruction experiments. The vertical bars are the 95% confidence intervals for the mean of the per-panicle scores.

1) *Our Method*: Our final method, as described in Section III. All experiments below are tweaks to this approach. Averaged across all panicles, this had the best $\alpha\beta$-MSE and $\alpha\beta$-SSIM scores.
2) *No Shift Maximize*: The mask overlap maximization discussed in Section III-C is not used. This resulted in a slight decrease in reconstruction quality.
3) *No Final Optimize*: The pair-wise ICP transformations discussed in Section III-C are still used to adjust cameras relative to the first frame, but the final optimization is not applied.
4) *Full-Cloud ICP*: Instead of running pair-wise ICP on masked seed points, ICP as described in Section III-C was run on the full point clouds. This test showed a significant drop in reconstruction quality.
5) *Arm Kinematics*: Views were combined using the arm kinematics, with no pose optimization. Although kinematically reconstructed panicles could be used for applications like collision avoidance, they had the worst reconstruction scores and could not be used for counting. Single seeds were clearly represented in multiple 3D locations, "smeared" cylindrically around the panicle.

The best reconstruction results came from pose adjustment using ICP on points determined to be high-quality seeds, and did notably better than ICP naively done using the full cloud from each image. Our hypothesis on why full-cloud ICP is worse is that sorghum is very organic and complex, and picking out meaningful, high-quality areas for ICP to operate on reduces the likelihood of ICP falling into a local minimum. As was discussed in *Arm Kinematics*, the required quality of reconstruction depends on your application. When using 3D structure to identify overlaps in 2D segmentation, decreasing reconstruction quality will lead to counting errors as identifications of the same seed drift apart in space.

C. Prediction of Sorghum Characteristics

As shown in Fig. 10, the count produced by our method has a strong linear fit to the ground truth count, with an R^2 of 0.875. The 10-fold RMSE using a 75/25 train/test split calculates an average prediction error of 295 seeds. There will always be some error in non-destructive counts, since sorghum panicles have internal, hidden seeds that cannot be seen from an outside view. The only way to expose all seeds is to strip them off the panicles, a time-consuming process.

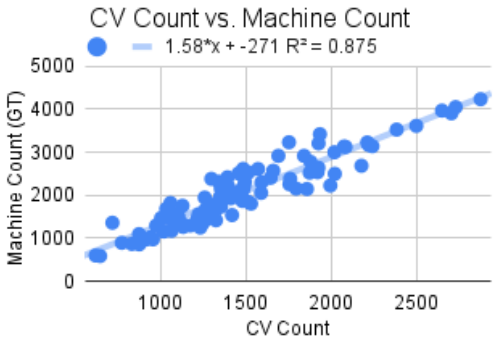


Fig. 10: Fit between our method's count (CV Count) and the ground truth count as described in Section IV-A.

Ultimately, the characteristic most worth measuring for sorghum is its yield weight, which represents a sellable

quantity of the crop. The fit between count and seed weight is still reasonably representative, with an R^2 linear fit of 0.819 in Fig. 11, but it fits slightly less well than Fig. 10. The 10-fold RMSE using a 75/25 train/test split calculates an average prediction error of 8.5 grams per panicle. This may be due to variations in seed weight.

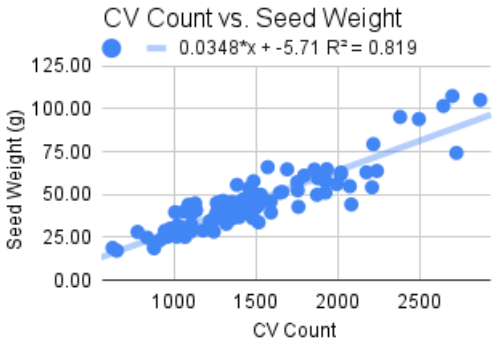


Fig. 11: Fit between counted seeds and seed weight, which is the weight of seeds after they have been stripped off a panicle and cleaned of husks.

D. Benefits of 3D Data over 2D

In [7] it was shown that it is sufficient to take a 2D count of one side of an ear of corn and scale that to a full kernel count. To test this, ears were rotated around their long axis by 90° increments and imaged, and it was found that single-image kernel counts had low variation because kernels were generally evenly distributed. In contrast, sorghum is more complex in shape, and therefore has more variation when a full count is extrapolated from a single image. In Fig. 12 and Fig. 13 we compare the predictiveness of 2D and 3D counts. It may seem unfair to compare 2D and 3D extrapolation because 3D methods have more data available (dozens of images vs. a single image), but it is important to evaluate for hardware considerations. Getting images surrounding a plant for 3D reconstruction is more costly in terms of system complexity, requiring the camera to be actuated rather than fixed to a mobile base such as a tractor, so it is important to assess what relative benefit the 3D method brings.

In order to test the extrapolation principle, we obtained 2D segment counts from images spaced 90° apart. This was complicated by the fact that some panicles were too tall to be captured in a single frame. To avoid trying to combine segmentation counts from multiple images, we only use counts where the full panicle is visible in four 2D views. 36 out of the 100 panicles met this criteria, enough to get a reasonable representation.

As seen in Fig. 12, 3D counts have a significantly better linear fit to the ground truth counts, with an R^2 of 0.885 compared to 0.623 for 2D counts (sampled randomly from the 90° separated views), demonstrating that 3D count is a better predictor of the desired feature. The variation in 2D count within each panicle can be seen in Fig. 13. There are significant variations in extrapolated counts within each panicle, often stretching to 20-40% of the ground truth value.

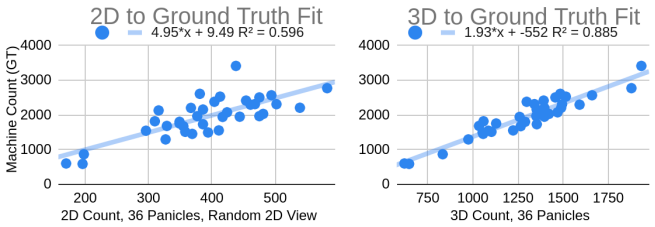


Fig. 12: Comparison of 2D and 3D counts fit to ground truth. 2D count comes from a single image per available panicle and has a lower R^2 score, indicating worse predictive performance for linear regression. The 10-fold RMSE for these 2D and 3D counts are 353 and 204 respectively.

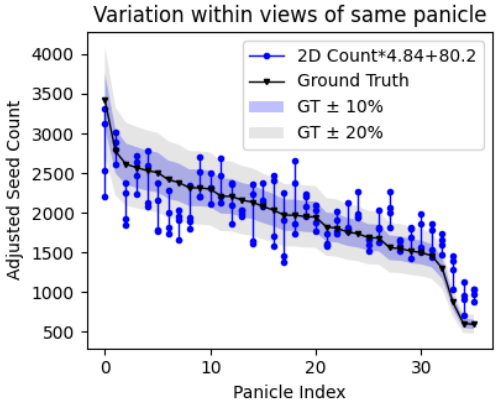


Fig. 13: Variation across viewpoints among the 36 panicles, using a linear fit to extrapolate from 2D count to an estimated full count. Linear fit parameters have been recalculated to use four 90° separated images per panicle instead of a random one as in Fig. 12. R^2 on the increased views was 0.634.

V. Conclusion

One of the benefits to this approach is the integration of segmentation counts across multiple 2D views, using the 3D model to determine which detections are unique. Future detection and segmentation improvements could be folded in to improve estimates while still getting the benefit of view combination. However, the use of multiple views is an intensive process, and uses many images of each panicle. It would be worthwhile to find the minimal image set that could reliably create a high-quality model, reducing runtime and resource requirements. Dense panicle models could also be put to other uses - in addition to extracting counts, other phenotyping or health characteristics could be evaluated, perhaps based on crop volume, color, or texture. The model could also be used to plan physical interactions between robots and the modelled crop. This fits into our lab's larger goal of modelling plants for analysis and manipulation.

Acknowledgments

We would like to thank Stephen Kresovich, Rick Boyles, and the Clemson team for the panicles. This work was partially supported by: ARPA-E TERRA DE-AR0001134, USDA NIFA 20216702135974, NSF Robust Int. 1956163.

REFERENCES

[1] L. R. House, "A guide to sorghum breeding," 1985.

[2] A. A. Author, B. B. Author, and C. Author, "Title of article," *Title of Journal*, vol. 10, no. 2, pp. 49–53, 2005.

[3] Y.-Y. Zheng, J.-L. Kong, X.-B. Jin, X.-Y. Wang, T.-L. Su, and M. Zuo, "Cropdeep: The crop vision dataset for deep-learning-based classification and detection in precision agriculture," *Sensors*, vol. 19, no. 5, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/5/1058

[4] L. Mosley, H. Pham, Y. Bansal, and E. Hare, "Image-based sorghum head counting when you only look once," 2020.

[5] A. Feng, H. Li, Z. Liu, Y. Luo, H. Pu, B. Lin, and T. Liu, "Research on a Rice Counting Algorithm Based on an Improved MCNN and a Density Map," *Entropy (Basel)*, vol. 23, no. 6, Jun 2021.

[6] Y. Li, J. Jia, L. Zhang, A. M. Khattak, S. Sun, W. Gao, and M. Wang, "Soybean seed counting based on pod image using two-column convolution neural network," *IEEE Access*, vol. 7, pp. 64 177–64 185, 2019.

[7] S. Khaki, H. Pham, Y. Han, A. Kuhl, W. Kent, and L. Wang, "Deepcorn: A semi-supervised deep learning method for high-throughput image-based corn kernel counting and yield estimation," *Knowledge-Based Systems*, vol. 218, p. 106874, 2021.

[8] M. Stein, S. Bargoti, and J. Underwood, "Image based mango fruit detection, localisation and yield estimation using multiple view geometry," *Sensors*, vol. 16, no. 11, p. 1915, 2016.

[9] X. Liu, S. W. Chen, C. Liu, S. S. Shivakumar, J. Das, C. J. Taylor, J. Underwood, and V. Kumar, "Monocular camera based fruit counting and mapping with semantic data association," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2296–2303, 2019.

[10] A. K. Nellithimaru and G. A. Kantor, "Rols: Robust object-level slam for grape counting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[11] S. Varela, T. Pederson, C. J. Bernacchi, and A. D. B. Leakey, "Understanding growth dynamics and yield prediction of sorghum using high temporal resolution uav imagery time series and machine learning," *Remote Sensing*, vol. 13, no. 9, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/9/1763

[12] Y. Bao, L. Tang, M. W. Breitzman, M. G. Salas Fernandez, and P. S. Schnable, "Field-based robotic phenotyping of sorghum plant architecture using stereo vision," *Journal of Field Robotics*, vol. 36, no. 2, pp. 397–415, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rob.21830

[13] S. N. Young, E. Kayacan, and J. M. Peschel, "Design and field evaluation of a ground robot for high-throughput phenotyping of energy sorghum," *Precision Agriculture*, vol. 20, no. 4, pp. 697–722, 2019. [Online]. Available: https://link.springer.com/article/10.1007/s11119-018-9601-6

[14] S. T. Nuske, K. Wilshusen, S. Achar, L. Yoder, S. G. Narasimhan, and S. Singh, "Automated visual yield estimation in vineyards," *J. Field Robotics*, vol. 31, pp. 837–860, 2014.

[15] C. Potena, R. Khanna, J. Nieto, R. Siegwart, D. Nardi, and A. Pretto, "Agricolmap: Aerial-ground collaborative 3d mapping for precision farming," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1085–1092, 2019.

[16] N. Chebrolu, P. Lottes, A. Schaefer, W. Winterhalter, W. Burgard, and C. Stachniss, "Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields," *The International Journal of Robotics Research*, vol. 36, p. 027836491772051, 07 2017.

[17] N. Ohi, K. Lassak, R. Watson, J. Strader, Y. Du, C. Yang, G. Hedrick, J. Nguyen, S. Harper, D. Reynolds, C. Kilic, J. Hikes, S. Mills, C. Castle, B. Buzzo, N. Waterland, J. Gross, Y.-L. Park, X. Li, and Y. Gu, "Design of an autonomous precision pollination robot," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 7711–7718.

[18] A. Silwal, F. Yandún, A. K. Nellithimaru, T. Bates, and G. Kantor, "Bumblebee: A path towards fully autonomous robotic vine pruning," *CoRR*, vol. abs/2112.00291, 2021. [Online]. Available: https://arxiv.org/abs/2112.00291

[19] P. Roy, W. Dong, and V. Isler, "Registering reconstructions of the two sides of fruit tree rows," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–9.

[20] A. Silwal, T. Parhar, F. Yandun, H. Baweja, and G. Kantor, "A robust illumination-invariant camera system for agricultural applications," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 3292–3298.

[21] L. Lipson, Z. Teed, and J. Deng, "RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching," *arXiv preprint arXiv:2109.07547*, 2021.

[22] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," in *CVPR*, 2020.

[23] S. Choi, Q.-Y. Zhou, and V. Koltun, "Robust reconstruction of indoor scenes," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5556–5565.

[24] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quarterly of applied mathematics*, vol. 2, no. 2, pp. 164–168, 1944.

[25] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. of 2nd International Conference on Knowledge Discovery and*, 1996, pp. 226–231.

[26] X. Zhao, R. Wu, Z. Zhou, and W. Wu, "A new metric for measuring image-based 3d reconstruction," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012, pp. 1030–1033.

[27] G. Zhang and Y. Chen, "A metric for evaluating 3d reconstruction and mapping performance with no ground truthing," in *ICIP*, 2021.

[28] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.