ORIGINAL RESEARCH **Open access**

A Conditional Generative Adversarial Framework for Monocular Depth Estimation in Apple Orchards

Elena Fedorova[1*], Alexey Volkov[1], Irina Morozova[2], Dmitry Kuznetsov[3]

Abstract

Vision-based robotic techniques for managing apple orchards have received considerable attention in recent years. For the vision system, obtaining accurate depth data of tree canopies is critical for gaining clearer insight into the spatial relationships among objects within the orchard setting, which directly supports safe and effective robotic performance. Yet, sensors capable of providing depth are typically costlier and less widely available than ordinary RGB cameras, which restricts easy access to depth-related information. This research demonstrates that a data-driven strategy employing a conditional generative adversarial network (cGAN)—specifically the Pix2Pix model—can successfully predict depth from RGB orchard images captured with a single monocular camera. The Pix2Pix architecture was modified to produce a depth channel from a standard RGB input image. Training and evaluation of the network were performed using orchard images collected from two distinct apple-growing systems and different camera types. Findings revealed that the generated depth maps closely matched the ground-truth depth channel, with a root-mean-squared error (RMSE) of 1.83 cm (equivalent to a relative error of 3.5%). Furthermore, the predicted depth images exhibited strong structural similarity (index> 0.55) and textural characteristics that matched those of the real depth images. Overall, these outcomes confirm that the Pix2Pix model represents a feasible option for producing accurate depth maps of fruit orchards from monocular RGB cameras, serving as a cost-effective substitute for pricier RGB-D sensors.

Keywords Machine vision, Generative adversarial networks, RGB-D, Pix2Pix, Agricultural robots

*Correspondence:
Elena Fedorova
elena.fedorova@nsu.ru
[1] Department of AI for Spring Wheat and Rye Lodging Risk Assessment, Faculty of Agriculture, Novosibirsk State University, Novosibirsk, Russia
[2] Department of Smart Machinery Route Optimization, Faculty of Agriculture, Ural Federal University, Yekaterinburg, Russia
[3] Department of Machine Learning for Frost Damage Prediction, Faculty of Agriculture, Kazan Federal University, Kazan, Russia

Introduction

The labor requirements for maintaining specialty crop orchards, including those growing apples, berries, and citrus, remain high, making such operations less viable amid rising labor costs and a declining pool of qualified personnel. As a result, both academic researchers and industry players have invested heavily in recent years to create automated and robotic systems for orchard management [1]. In apple production, robotic technologies have been actively developed for demanding manual tasks such as pruning, thinning of flowers and fruitlets, and harvesting [2-4]. Robotic orchard platforms typically comprise three core components: a machine vision module, a manipulation mechanism, and an end effector. The vision module is responsible for detecting and positioning targets so the manipulator and end-effector can interact with them appropriately (for instance, cutting branches or grasping fruit). Thus, the vision system's reliability largely determines the success of the entire robotic operation.

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

Advancements in 3D sensing technology and artificial intelligence have substantially boosted the capabilities of vision systems, leading to better overall robot performance in orchards [5, 6]. A wide range of machine-vision solutions has been proposed for robotic applications in pruning, harvesting, and thinning operations [2, 7-9]. Before depth cameras, 3D localization of orchard objects mainly relied on triangulation methods applied to several images from a single color camera [10]. In some harvesting setups, fruits were first identified in 2D images, after which the robotic arm moved until the gripper contacted the fruit [11]. Other approaches estimated fruit location using its apparent size and position in the image combined with known camera parameters [12]. However, relying solely on color images for distance measurement has often led to inaccuracies, especially in complex, unstructured orchard conditions when using traditional image processing. Stereo vision setups, involving multiple cameras spaced apart, have also been widely used for canopy localization, but their deployment is hindered by their high complexity and processing demands [13].

Conversely, RGB-D cameras deliver both color images suitable for recognition tasks (including detection and segmentation) and a dedicated depth channel for precise 3D positioning. Active RGB-D sensors, which use an infrared light projector and receiver, have emerged as the preferred choice in robotics. These devices capture color images alongside 3D data more rapidly and with greater precision compared to alternatives. For instance, Tu *et al.* [14] used RGB and depth data for passion fruit localization and achieved superior results compared to stereo-vision systems. Several other works have incorporated RGB-D cameras in robotic apple harvesting, reporting detection rates exceeding 0.85 [15, 16]. Gené-Mola *et al.* [17] leveraged color, depth, and intensity channels from an RGB-D sensor to identify Fuji apples, attaining an F1-score of 0.90 and average precision of 94.8%. Additional research demonstrated localization accuracy of less than 10 mm in Cartesian coordinates [18]. Depth data has further enabled detailed point cloud generation of apple tree structures for 3D modeling in pruning applications [19, 20]. Despite these benefits, acquiring in-depth information typically requires more expensive hardware. Therefore, deriving depth estimates directly from affordable RGB cameras holds strong potential to accelerate the practical use and commercialization of orchard robotics.

Monocular depth estimation from individual images has emerged as a vibrant research field in recent years [21]. The core difficulty arises from fundamental ambiguity, where a single RGB image can correspond to multiple possible 3D configurations (a one-to-many relationship). Effective depth prediction thus requires robust techniques to translate visual appearance into depth values. Early methods depended on manually designed features, such as semantic object classification paired with machine learning that incorporated class-specific knowledge [22]. Alternative strategies used known object dimensions for scale estimation, spatial layout cues, and relative depth relationships [23]. More recent efforts have focused on refining these techniques through multi-task frameworks, improved network designs, and optimized loss functions within deep learning [24]. The pioneering work by Eigen *et al.* [25] introduced a neural network dedicated to depth prediction.

Nevertheless, such early models suffered from slow inference speeds unsuitable for real-time use, excessive parameters, heavy data requirements, and limited output resolution. These shortcomings were later mitigated by Laina *et al.* [26] using a ResNet-50-based convolutional neural network (CNN) with upsampling layers to produce higher-resolution outputs and reduce parameter count. Numerous CNN architectures for single-image depth estimation followed [27, 28]. Lately, conditional generative adversarial networks (cGANs) built on CNNs have shown strong promise for depth map synthesis [29, 30]. In cGANs, input images condition both the generator and discriminator in a supervised setup, where the model learns to regress depth from paired RGB-depth training data [31]. Although many depth estimation studies exist, none have yet explored 3D scene reconstruction tailored to agricultural environments like fruit orchards.

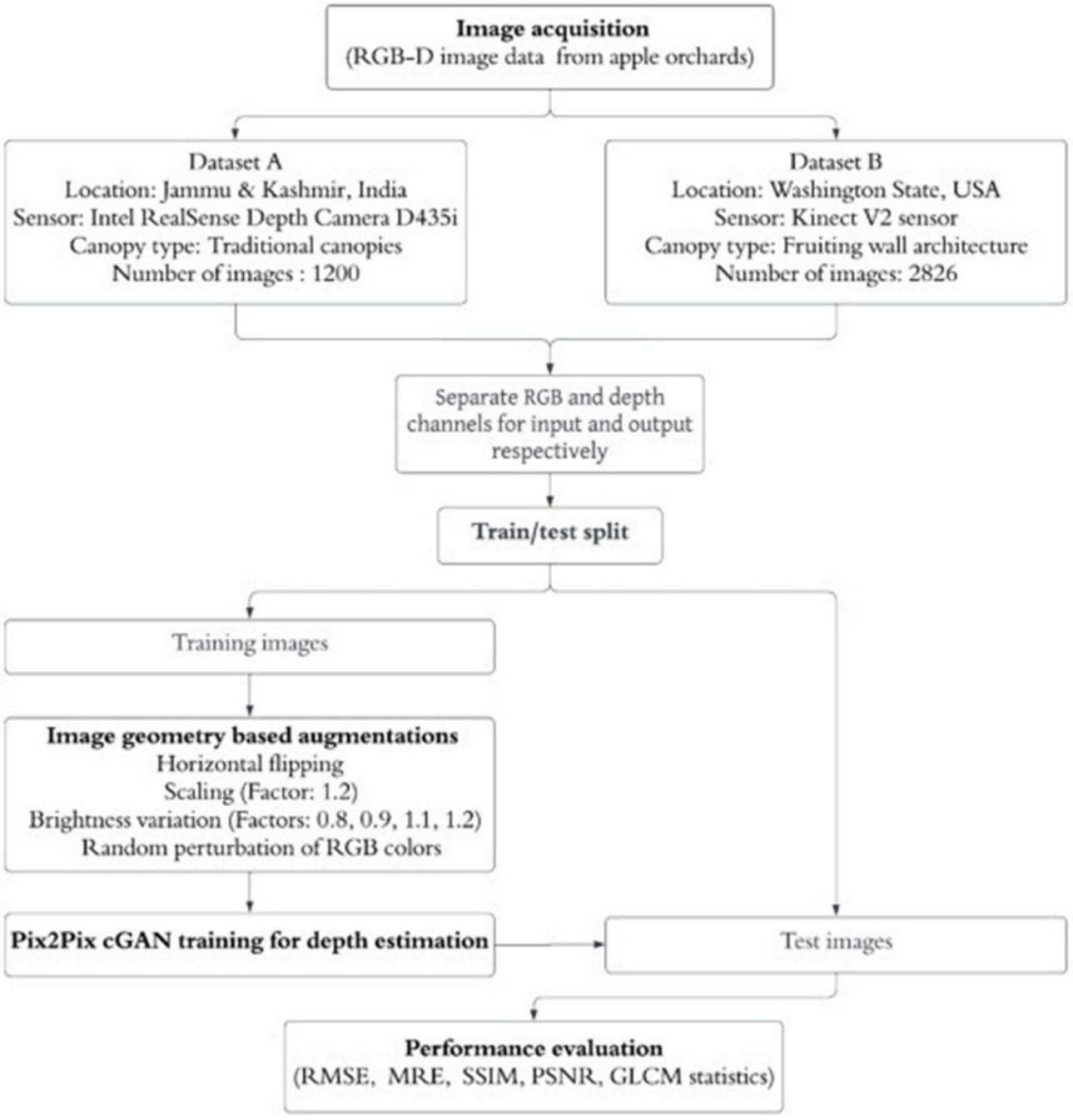
This work demonstrates the viability of GAN-based models for extracting depth information from conventional RGB images captured in apple orchards. The Pix2Pix framework, originally introduced by Isola *et al.* [32], is a popular GAN variant widely used in agriculture [33, 34] and remote sensing [35, 36]. Within farming contexts, prior research successfully applied Pix2Pix to derive normalized difference vegetation index (NDVI) maps from standard RGB UAV imagery, bypassing the need for NIR sensors and cutting system costs by as much as $5000 per unit [37, 38]. Building on this, the current study adopts the image-to-image translation approach of the Pix2Pix GAN [32]. Its central aim is to

© 2024 The Author(s). | Published by Agro Intelligence Research Press

establish a deep learning solution using Pix2Pix to obtain reliable depth data without requiring specialized depth-sensing equipment for robotic orchard tasks.

Materials and Methods

This investigation aimed to evaluate how effectively a cGAN could synthesize depth maps from RGB images. The Pix2Pix network [32] served as the foundational model due to its proven effectiveness across various image translation applications [36-38]. The following section outlines the dataset preparation, the detailed architecture and training procedure of the Pix2Pix model for depth prediction, and the performance metrics selected for assessment. A summary of the complete workflow appears in **Figure 1**.



© 2024 The Author(s). | Published by Agro Intelligence Research Press

Figure 1. Workflow for estimating a depth image from an RGB image of a fruit orchard using the Pix2Pix cGAN.

Dataset

Images were obtained from apple orchards employing RGB-D sensors. Collections originated from two separate production systems: one in Jammu & Kashmir (J&K), India, characterized by trees forming dense traditional bushy canopies, and another in Washington State, USA, where trees followed contemporary fruiting wall layouts **(Figure 2)**. J&K orchard data came via an Intel® RealSense® Depth Camera D435i (Intel Corporation, San Francisco, California, US) at sites operated by the Sher-e-Kashmir University of Agricultural Sciences and Technology (SKUAST) and the Indian Council of Agricultural Research – Central Institute of Temperate Horticulture (ICAR-CITH) in J&K, India [39]. Operators manually held the camera while standing 0.5-2.0 m from the trees. Further details on the collection are provided in Rathore *et al.* [39]. Twenty RGB videos, each 4–8 s at 60 fps, were recorded alongside per-frame depth information. From these, 1200 RGB images were extracted with their paired depth images **(Figure 2a)**.

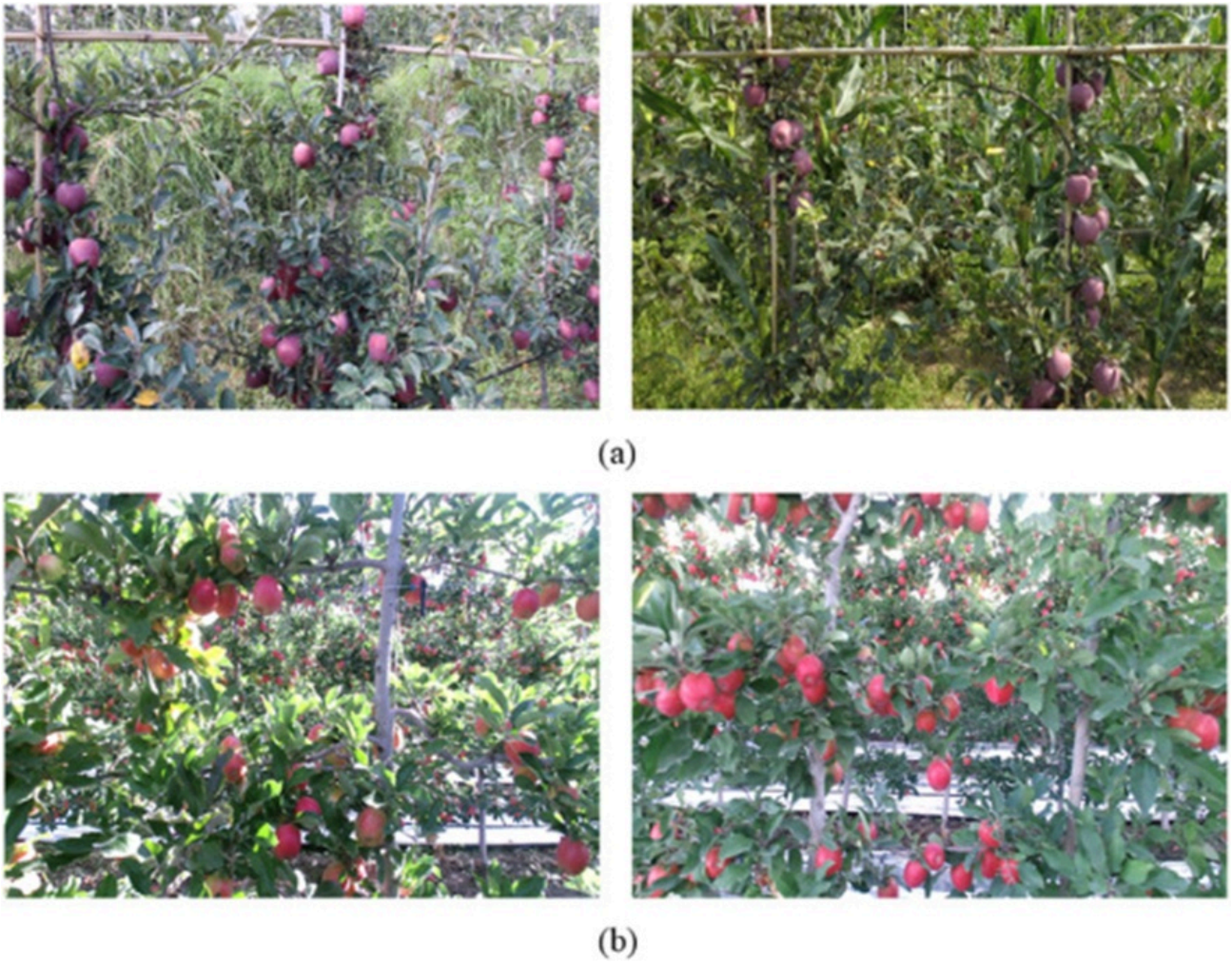


Figure 2. Sample RGB images acquired from apple orchards in (a) J&K (India) and (b) Washington State (USA).

Washington State materials consisted of 2826 RGB-D images of apple tree canopies **(Figure 2b)**. Capture was conducted at a commercial site in Prosser, WA, USA, using a Kinect V2 sensor (Microsoft Inc., Redmond, WA) [40, 41]. Complete acquisition methods and orchard background are covered in Zhang *et al.* [42]. In total, this study used 4026 images: 3626 for training and 400 for testing (200 per orchard type). Testing samples originated from different tree rows

© 2024 The Author(s). | Published by Agro Intelligence Research Press

than the training ones. Each image paired its RGB channels with the depth channel **(Figure 3)** by joining them along the initial dimension, consistent with Pix2Pix training guidelines [32].

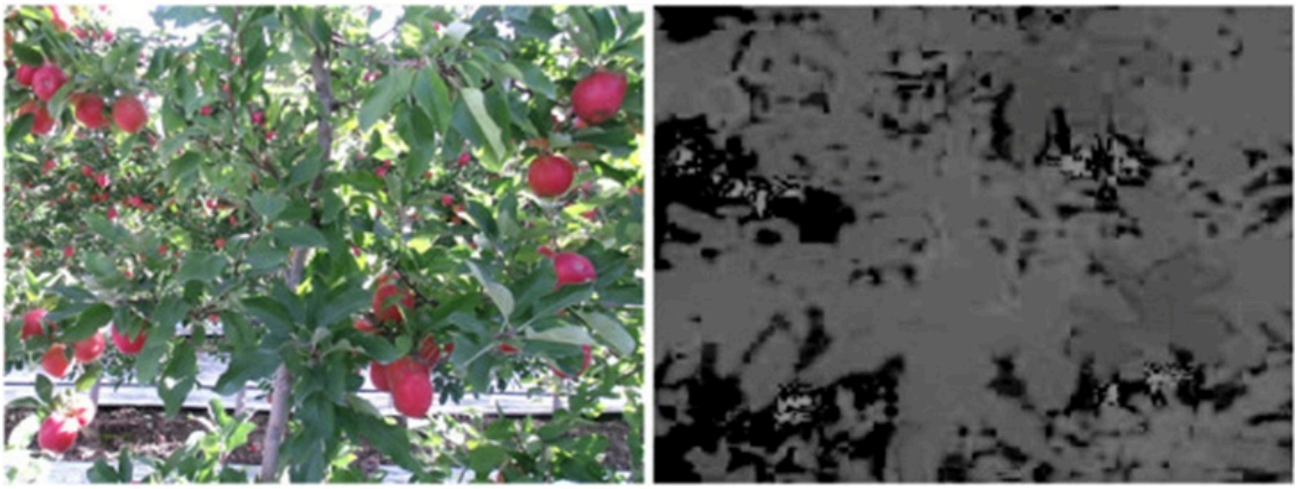


Figure 3. Example of an RGB image of an apple orchard fruiting wall and corresponding depth channel, acquired using the Kinect V2 RGB-D sensor.

Depth estimation using Pix2Pix GAN

Training the Pix2Pix network [32] involved matched pairs — each RGB image paired with its depth counterpart — enabling the system to learn mappings between their characteristics. The GAN framework featured a generator that generated the depth channel from RGB input, alongside a discriminator tasked with detecting inaccuracies in the synthesized depth relative to the genuine one. Adversarial training drives the generator to produce realistic fakes that fool the discriminator [31]. For conditional GANs (cGANs) of this kind, the loss function incorporates conditioning on specific input traits to yield properly featured outputs. Optimization follows a min-max process that learns a mapping from random noise z to the output y given the input x, as formulated in **Eq. 1**. In this setup, the RGB image serves as x, the true depth image as y, and the synthesized depth as the outcome.

$$\bar{y} = G(x).$$

$$\min_G \max_D L_{cGAN}(G, D) = E_{x.y}[\log D(x, y)] + E_{x.z}[\log(1 - D(x, G(x, y)))] \quad (1)$$

An encoder-decoder U-Net structure [43] powered the generator, while the discriminator adhered to PatchGAN [44] principles. Skip links connect the i-th layer to the (n−i)-th layer in the generator (where n denotes the total number of layers). PatchGAN, in turn, conducts localized patch evaluations on outputs, tagging them as real or synthetic, and reaches a verdict via majority patch consensus. The implementation specifics of these elements in Pix2Pix are thoroughly explained by Isola *et al.* [32]. Hyperparameters for training are listed in **Table 1**. Augmentations on the training data included geometric and intensity transformations: horizontal flips, scaling by 1.2, brightness shifts by 0.8, 0.9, 1.1, and 1.2, and random RGB color disturbances. Such enhancements enlarged the effective dataset and strengthened model resilience.

Table 1. Hyperparameters used to train the Pix2Pix network for depth estimation from RGB images.

Hyperparameter	Value

© 2024 The Author(s). | Published by Agro Intelligence Research Press

Optimizer	Adam
Learning rate	0.002
Gradient decay factor	0.5
Batch size	16
Epochs	1000

Evaluation of model performance

Accuracy and realism of synthesized depth images versus actual ones were gauged through these metrics:

Root mean squared error (RMSE)

Root mean squared error (RMSE) quantifies the discrepancy between model-estimated depths and measured depths. Calculation for an individual image follows:

$$RMSE = \sqrt{\frac{1}{m \times n} \sum_{0}^{m-1} \sum_{0}^{n-1} ||r(i,j) - g(i,j)||^2} \quad (2)$$

Where r and g stand for the actual and generated depth images having m rows and n columns.

Mean relative error (MRE)

Mean relative error (MRE) calculates depth deviation relative to true depth values. It quantifies the proportional departure of estimates from the source depth channel.

$$MRE = \frac{1}{m \times n} \sum_{0}^{m-1} \sum_{0}^{n-1} \frac{|r(i,j) - g(i,j)|}{r(i,j)} \quad (3)$$

Structural similarity index measure (SSIM)

The Structural Similarity Index Measure (SSIM) [45] assesses local image properties such as brightness, contrast, and organization using a single score. Structures involve intensity patterns among adjacent pixels, normalized for brightness and contrast. SSIM for matching windows x and y is expressed as:

$$SSIM\ (x,y) = \frac{((2 \times \mu_x \times \mu_x) + C_1) + ((2 \times \sigma_{xy}) + C_2)}{(\mu_x^2 + \mu_x^2 + C_1) \times (\sigma_x^2 + \sigma_x^2 + C_2)} \quad (4)$$

Consequently, SSIM values near zero indicate no resemblance, while values of 1 denote perfect identity.

Peak signal-to-noise ratio (PSNR)

The peak signal-to-noise ratio (PSNR) is calculated using **Eq. 5**. It measures the quality of the generated depth image in decibels (dB); higher values indicate better quality. Computation derives from MSE, with the true depth image as signal and prediction discrepancies as noise. MSE indicates the average deviation, and PSNR scales it relative to the image's peak intensity.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

$$PSNR = 20\ log_{10}\left(\frac{MAX_f}{\sqrt{MSE}}\right) \quad (5)$$

Where MAXf equals the peak signal level in the source image, and MSE is the mean squared error.

Gray-level co-occurrence matrix (GLCM) statistics

Examining textural properties of generated and real depth images is significant. GLCM is a widely used approach in pattern analysis for texture extraction [9, 46, 47]. Correlations in GLCM-derived statistics were thus computed across actual and synthetic depth images. Key measures from the GLCM (Gn×n) illuminate textural properties:

$$Contrast = \sum_{i,j} |i-j|^2\ G(i,j) \quad (6)$$

$$Correlation = \sum_{i.j} \frac{(i-\mu_i)(j-\mu_j)G(i,j)}{\sigma_i\sigma_j} \quad (7)$$

$$Homogeneity = \sum_{i,j} \frac{G(i,j)}{1+|i-j|} \quad (8)$$

Where G(i,j) represents the entry at row i and column j in the GLCM; μ and σ indicate matrix mean and standard deviation.

Results and Discussion

Qualitative evaluation of model performance

Assessments covered both visual and numerical angles. Visual checks relied on direct observation. Representative RGB inputs with associated translated depth outputs and ground-truth references appear in **Figure 4**. General distance variations among orchard elements — branches, fruits, trellis wires — remained evident in produced depth maps. Tree component renderings displayed close alignment between the real and generated versions. This supports the method's capacity to furnish practical depth approximations for robotic orchard uses.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

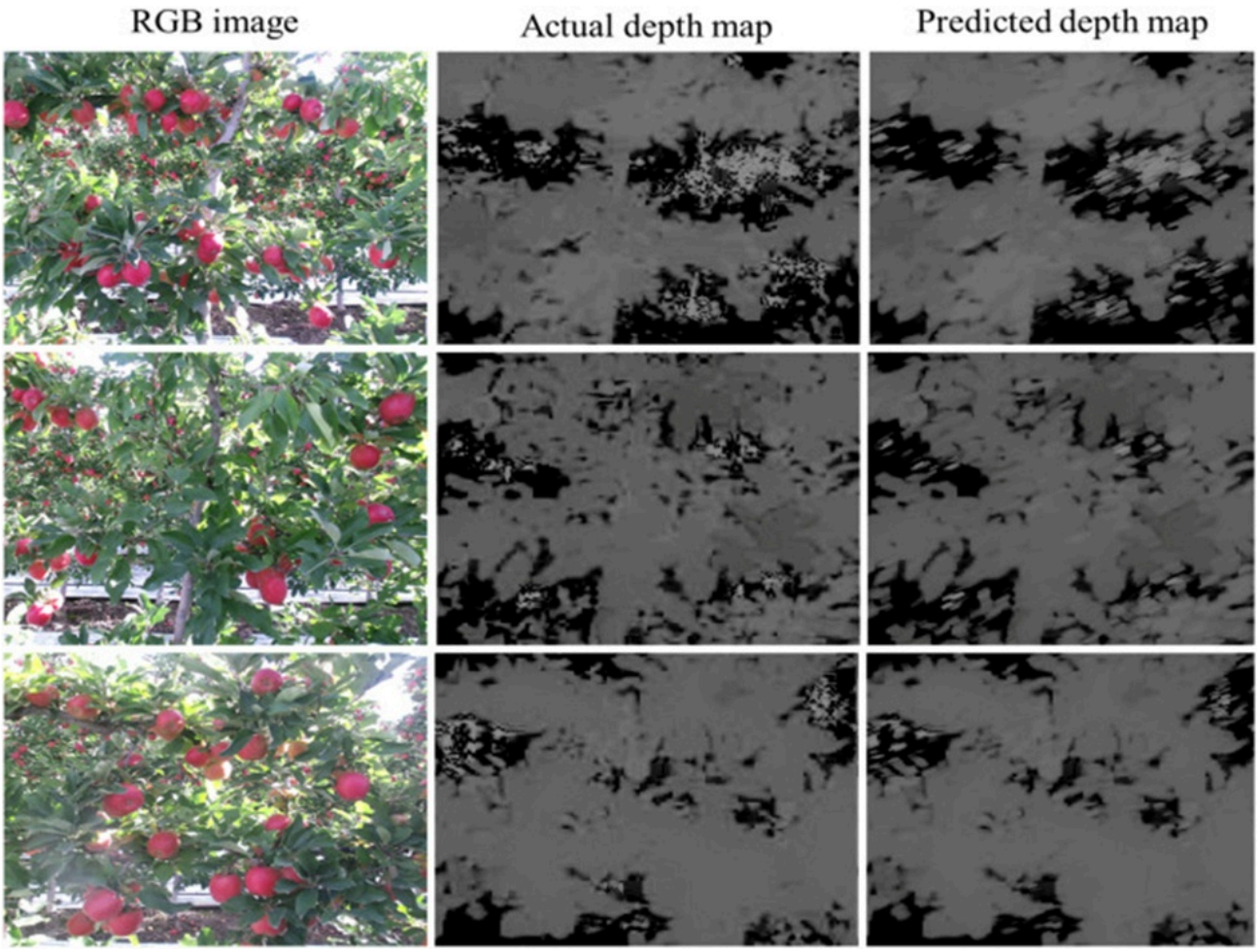


Figure 4. The RGB image and corresponding actual and model-generated/predicted depth maps are shown for comparison.

Performance was stronger overall on Washington State samples than on J&K samples. The structured fruiting wall design in Washington likely aided this, confining canopy elements to a tight depth span from the wall and presenting the model with a narrower, predictable range. Traditional J&K layouts spread canopy outward in varied directions from the trunk, yielding intricate 3D forms. Image canopy density exerted little influence on outcomes. Challenges, however, persisted in generating crisp boundaries at canopy edges **(Figure 4)**, consistent with recognized GAN constraints [36]. These small discrepancies probably have a negligible impact on the broad depth portrayal from RGB inputs.

Quantitative evaluation of model performance

Deeper insights came from a statistical review of the entire test collection, as well as isolated reviews per location (J&K and Washington State), applying metrics from Section 2.3. Divergences in growing styles (canopy designs) and sensor types justified separate checks to probe model stability. **Table 2** aggregates average RMSE, MRE, SSIM, and PSNR across test images. The global depth estimation RMSE was 1.833 cm, with an ~3.5% MRE. Boundary ambiguities observed visually explained much of the error. Over 70% of pixels registered absolute errors of less than 0.8 cm in J&K data and less than 0.65 cm in Washington State data. The average generation time per depth channel from RGB was 0.494 s, aligning with typical fruit detection processing durations [48]. Such levels are well-suited for deploying Pix2Pix depth outputs in robotics, notably for target positioning [49].

Table 2. Statistical results produced by the Pix2Pix model for predicting a depth map from a color (RGB) image.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

Statistical metric	J&K orchard images	Washington State orchard images	Overall
RMSE (cm)	2.019	1.647	1.833
MRE (%)	3.774	3.164	3.469
SSIM	0.595	0.538	0.564
PSNR (dB)	18.538	15.943	17.241

SSIM exceeded 0.5 across test images from both sites, verifying the model's proficiency at crafting depth estimates comparable to sensor measurements. A high SSIM indicates strong spatial correspondence in intensity patterns across large image regions. **Figure 5** illustrates actual depth samples, model predictions, and local SSIM overlays. Background regions achieved top alignment due to straightforward modeling. Effective background distinction enables applications such as depth-guided filtering to improve target identification, as in fruit spotting [15] or branch outlining [50]. Sharp depth shifts at trunk and branch perimeters yielded the lowest similarity **(Figure 5)**. Even so, SSIM distributions confirm solid capture of tree forms and details. The mean PSNR across tests reached 17.2 dB, with J&K at PSNRJ = 18.5 dB and Washington State at PSNRW = 15.9 dB. The literature on image creation views PSNR over 15 dB as a reliable indicator of fidelity to reference (real depth) images [51, 52].

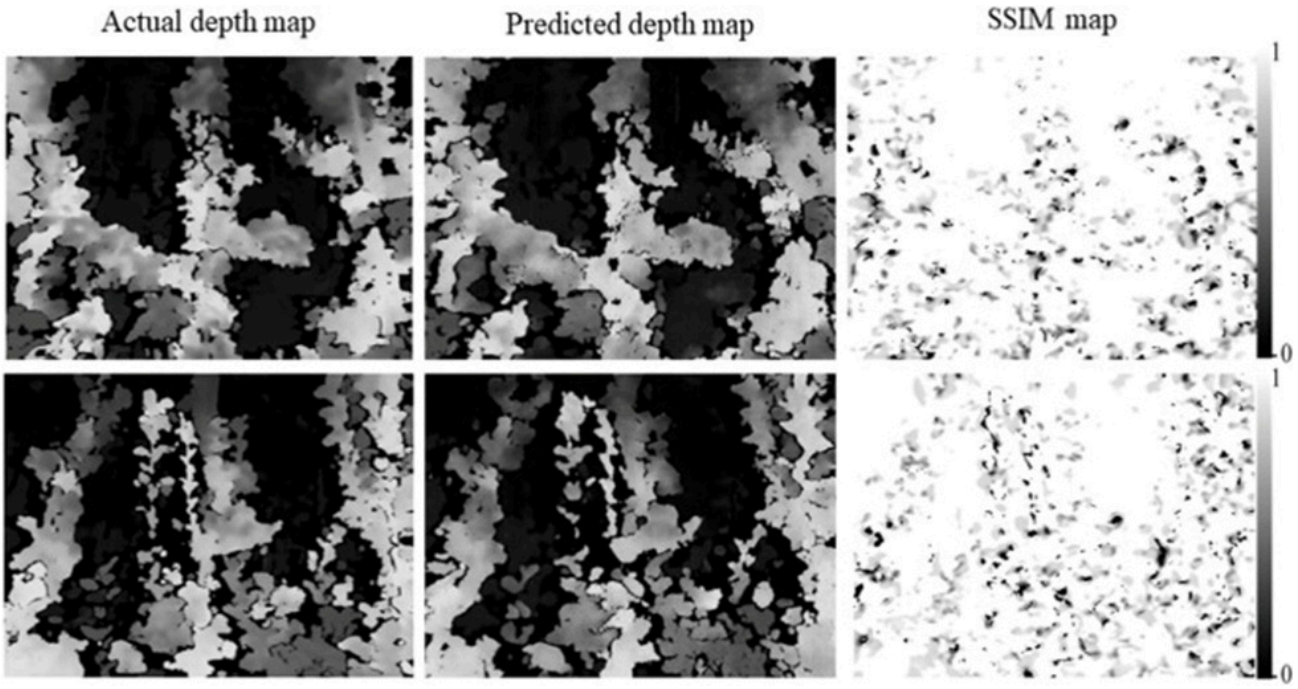


Figure 5. Column I: real/actual depth images, column II: Pix2Pix model predicted depth images, and column III: local SSIM map between real and predicted images. Higher values of local SSIM are represented by brighter pixels, and small SSIM values appear as darker pixels.

Additional checks correlated the real and generated depth images with the RGB source to assess the strength of the relationship [21]. SSIM and PSNR used RGB as a baseline. Regression produced R2 above 0.92 for SSIM **(Figure 6)**. PSNR R2 reached 0.822 on J&K data and 0.774 on Washington State data **(Figure 7)**.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

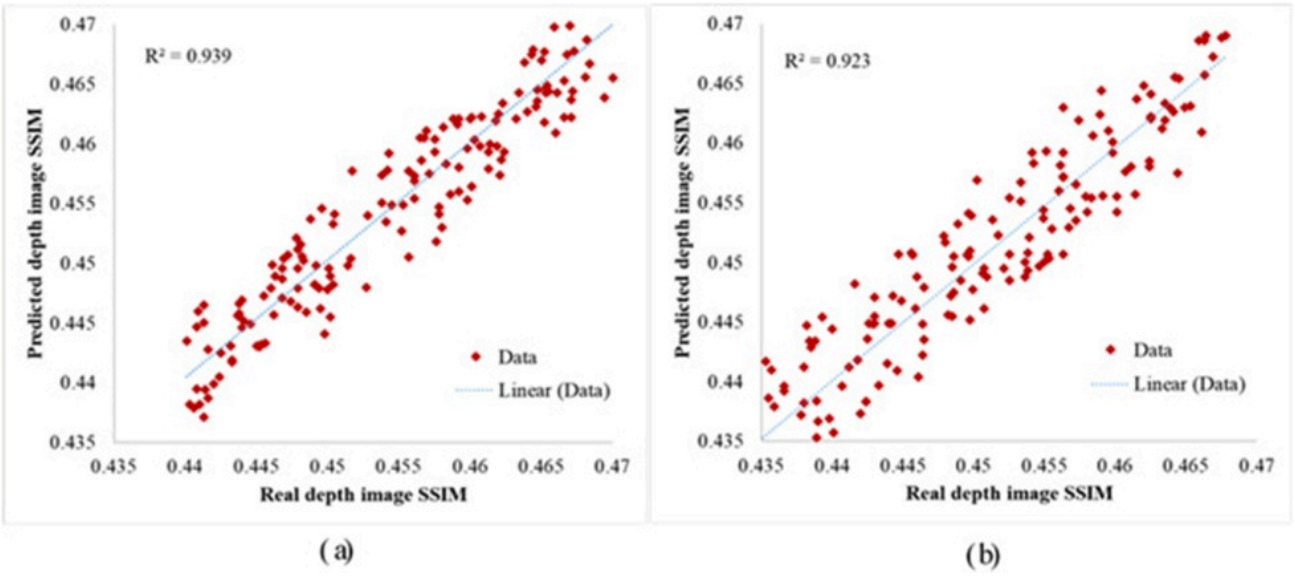


Figure 6. Linear regression for SSIM for the test images of (a) J&K orchard and (b) Washington state orchards dataset.

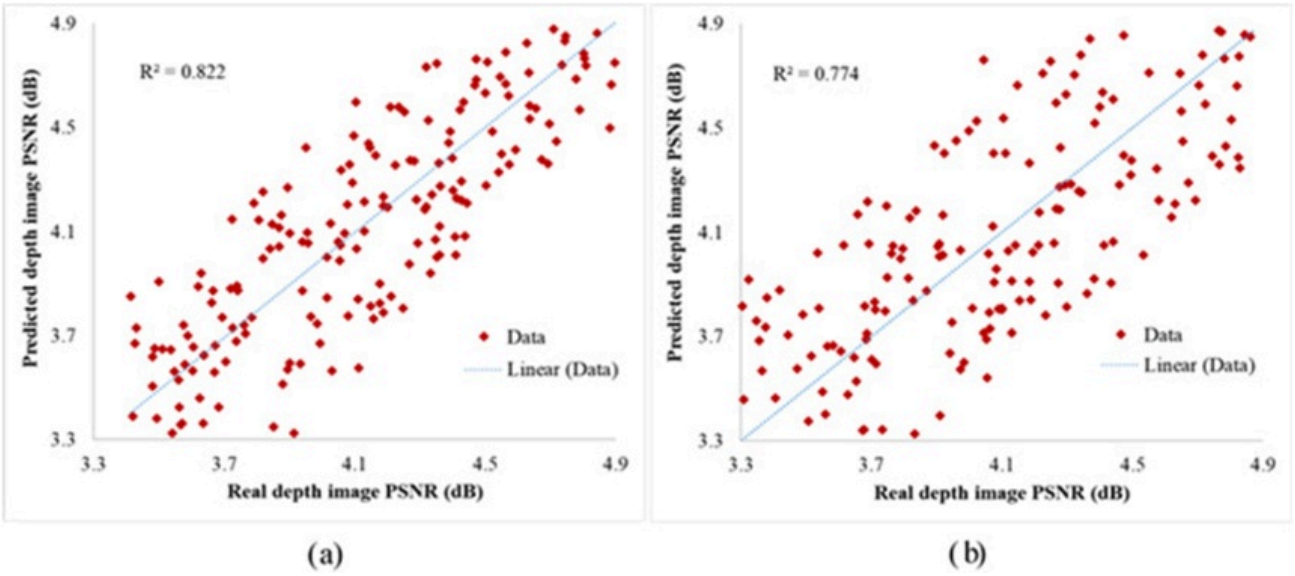


Figure 7. Linear regression for PSNR for the test images of (a) J&K orchards and (b) Washington state orchards.

Furthermore, texture characteristics within depth images can reveal spatial differences in depth values associated with objects in the corresponding RGB images. Accordingly, regression analysis evaluated the agreement among GLCM statistics from the actual and synthesized depth images **(Figure 8)**. The three GLCM measures — correlation, contrast, and homogeneity — captured the textural properties of these depth images. Across all three measures, R2 values between predicted and real depth images exceeded 0.89 **(Figure 8)**. Contrast is a key indicator of local intensity changes within images. Notably, very high R2 values (> 0.95) appeared for ‘contrast’ in both datasets when comparing real and generated images **(Figure 8b)**. Likewise, correlation, reflecting the likelihood of particular depth value pairs occurring together, reached R2 levels of 0.90 and above. Depth images from J&K orchards generally showed greater homogeneity (0.06 to 0.27) compared to those from Washington State orchards (0.04 to 0.22). Strong alignment between ground-truth and predicted depth data across these metrics confirmed that the proposed Pix2Pix model effectively extracts depth details from RGB color channels.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

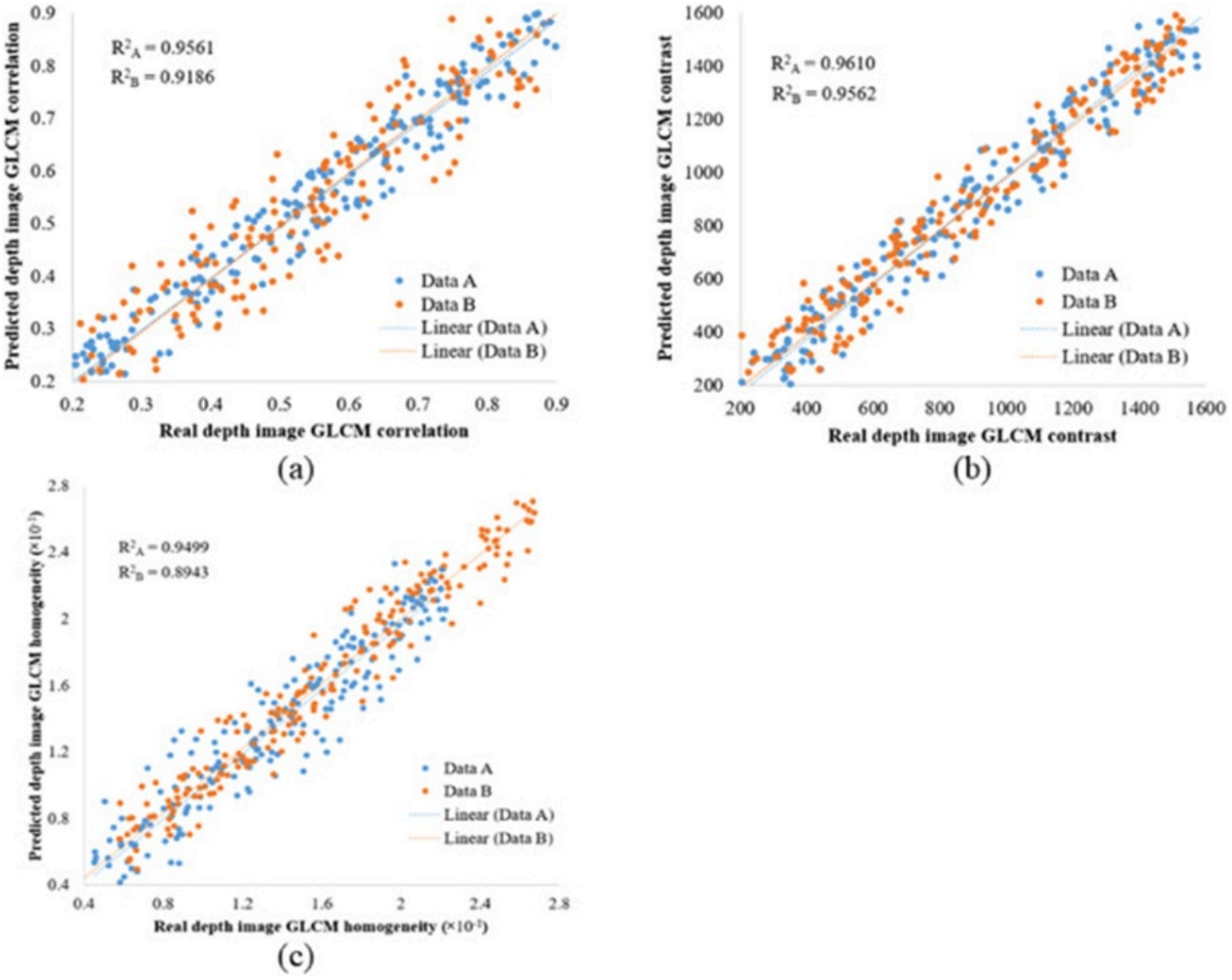


Figure 8. Linear regression for GLCM statistical metrics derived from the actual and estimated depth images: (a) correlation, (b) contrast, and (c) homogeneity. Data points A and B represent test images from orchards of Washington State (US) and J&K (India), respectively.

The outcomes of this research indicate that the deep-learning Pix2Pix model can produce accurate depth approximations of apple orchards from monocular RGB images while keeping errors low. Depth estimation has remained a vibrant research domain, with various works showing that GANs offer promising and superior alternatives to traditional hand-designed features, including texton features, SIFT, GIST, and Markov Random Fields (MRF) [53]. Yet, the majority of existing literature centers on models for generating depth maps in indoor environments, yielding outcomes suitable for integration with conventional color cameras [54, 55]. Outdoor environments often involve wide depth ranges, occasionally exceeding sensor capabilities, as observed in fruit orchards. Thus, a robust technique is necessary to handle the uncertainty in converting monocular images to depth maps. This investigation demonstrated that the developed method supports depth prediction to assist multiple autonomous, vision-driven robotic tasks in specialty crop orchard handling that need 3D target positioning. Both visual and numerical assessments revealed strong model suitability for pairing with standard color cameras to acquire depth maps. Training and testing occurred across images from two varied cameras and canopy structures to gauge broader applicability. Comparable results were obtained across both orchard types, suggesting that incorporating data from diverse sensors could further strengthen model stability. Still, thorough robustness verification would require cross-validation on datasets from additional sensors with depth ranges differing from those in the training collection.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

Additionally, integration into machine-vision workflows must address the increased computational requirements of GANs and the additional frame processing time. When merging with other deep learning components, parallel processing pipelines are advisable over sequential ones to avoid excessive inference delays. Upcoming research should emphasize tailoring GANs for orchard-specific applications to reduce reconstruction loss in Pix2Pix, thereby reducing overly smooth edges in predicted depth outputs. Enhancements to the study's findings are needed to support tasks that demand precise depth resolution, such as robotic pollination. Expansion into additional specialty crops, such as cherries and wine grapes, could also be pursued later.

Conclusion

This study examined the capabilities of a deep-learning cGAN, Pix2Pix, to predict depth maps using solely a monocular RGB image as input. Evaluation was conducted using both quantitative and qualitative methods across two separate apple orchard datasets, confirming that the method delivers effective depth-feature predictions in apple orchard canopy scenes. The RMSE between predicted and ground-truth depth was 1.83 cm, with an MRE of 3.5 %. SSIM and PSNR reached 0.56 and 17.2 dB, respectively. Additional numerical indicators, such as GLCM-based textural features including contrast, correlation, and homogeneity, likewise showed strong agreement between the estimated and ground-truth depth maps.

Furthermore, R2 values for linking actual and estimated depth images ranged from 0.77 to 0.96 across all metrics. Depth estimate generation required 0.494 s per image. The ability to derive depth from a single image could reduce reliance on costlier RGB-D sensors and intricate stereo setups for depth data in robotic orchard management, thereby boosting the commercial feasibility of these solutions on resource-limited farms globally. Subsequent developments will focus on creating a comprehensive vision framework for 3D target localization from RGB images.

Acknowledgements

Research funding received from the Ministry of Electronics and Information Technology (Government of India) through the National Program on Electronics and ICT Applications in Agriculture and Environment is gratefully acknowledged.

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 14 Sep 2023 Revised: 19 Nov 2023 Accepted: 05 Dec 2023
Published online: 25 May 2025

© 2024 The Author(s). | Published by Agro Intelligence Research Press

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Zhang Q. Opportunity of robotics in specialty crop production. IFAC Proc. 2013;46:38-9.

2. Hussain M, He L, Schupp J, Lyons D, Heinemann P. Green fruit segmentation and orientation estimation for robotic green fruit thinning of apples. Comput Electron Agric. 2023;207:107734. https://doi.org/10.1016/j.compag.2023.107734.

3. Bhattarai U, Zhang Q, Karkee M. Design, integration, and field evaluation of a robotic blossom thinning system for tree fruit crops. 2023https://arxiv.org/abs/2304.04919v1 (accessed August 30, 2023).

4. Wang X, Kang H, Zhou H, Au W, Wang MY, Chen C. Development and evaluation of a robust soft robotic gripper for apple harvesting. Comput Electron Agric. 2023;204:107552. https://doi.org/10.1016/j.compag.2022.107552.

5. Gongal A, Amatya S, Karkee M, Zhang Q, Lewis K. Sensors and systems for fruit detection and localization: A review. Comput Electron Agric. 2015;116:8-19. https://doi.org/10.1016/j.compag.2015.05.021.

6. Wang T, Chen B, Zhang Z, Li H, Zhang M. Applications of machine vision in agricultural robot navigation: A review. Comput Electron Agric. 2022;198:107085. https://doi.org/10.1016/j.compag.2022.107085.

7. Bhattarai U, Karkee M. A weakly-supervised approach for flower/fruit counting in apple orchards. Comput Ind. 2022;138:103635. https://doi.org/10.1016/j.compind.2022.103635.

8. Montoya-Cavero LE, Díaz de León Torres R, Gómez-Espinosa A, Escobedo Cabello JA. Vision systems for harvesting robots: Produce detection and localization. Comput Electron Agric. 2022;192:106562. https://doi.org/10.1016/j.compag.2021.106562.

9. Divyanth LG, Soni P, Pareek CM, Machavaram R, Nadimi M, Paliwal J. Detection of coconut clusters based on occlusion condition using attention-guided faster R-CNN for robotic harvesting. Foods. 2022;11:3903. https://doi.org/10.3390/foods11233903.

10. Wang Q, Nuske S, Bergerman M, Singh S. Automated crop yield estimation for apple orchards. STAR. 2013;88:745-58. https://doi.org/10.1007/978-3-319-00065-7_50.

11. De-An Z, Jidong L, Wei J, Ying Z, Yu C. Design and control of an apple harvesting robot. Biosyst Eng. 2011;110:112-22. https://doi.org/10.1016/j.biosystemseng.2011.07.005.

12. Mehta SS, Burks TF. Vision-based control of robotic manipulator for citrus harvesting. Comput Electron Agric. 2014;102:146-58. https://doi.org/10.1016/j.compag.2014.01.003.

13. Plebe A, Grasso G. Localization of spherical fruits for robotic harvesting. Mach Vis Appl. 2001;13:70-9. https://doi.org/10.1007/PL00013271.

14. Tu S, Xue Y, Zheng C, Qi Y, Wan H, Mao L. Detection of passion fruits and maturity classification using Red-Green-Blue Depth images. Biosyst Eng. 2018;175:156-67. https://doi.org/10.1016/j.biosystemseng.2018.09.004.

15. Fu L, Majeed Y, Zhang X, Karkee M, Zhang Q. Faster R–CNN–based apple detection in dense-foliage fruiting-wall trees using RGB and depth features for robotic harvesting. Biosyst Eng. 2020;197:245-56. https://doi.org/10.1016/j.biosystemseng.2020.07.007.

16. Yoshida T, Kawahara T, Fukao T. Fruit recognition method for a harvesting robot with RGB-D cameras. Robomech J. 2022;9:1-10. https://doi.org/10.1186/S40648-022-00230-Y.

17. Gené-Mola J, Vilaplana V, Rosell-Polo JR, Morros JR, Ruiz-Hidalgo J, Gregorio E. Multi-modal deep learning for Fuji apple detection using RGB-D cameras and their radiometric capabilities. Comput Electron Agric. 2019;162:689-98. https://doi.org/10.1016/j.compag.2019.05.016.

18. Nguyen TT, Vandevoorde K, Wouters N, Kayacan E, De Baerdemaeker JG, Saeys W. Detection of red and bicoloured apples on tree with an RGB-D camera. Biosyst Eng. 2016;146:33-44. https://doi.org/10.1016/j.biosystemseng.2016.01.007.

19. Karkee M, Adhikari B, Amatya S, Zhang Q. Identification of pruning branches in tall spindle apple trees for automated pruning. Comput Electron Agric. 2014;103:127-35. https://doi.org/10.1016/j.compag.2014.02.013.

20. Zahid A, Mahmud MS, He L, Heinemann P, Choi D, Schupp J. Technological advancements towards developing a robotic pruner for apple trees: A review. Comput Electron Agric. 2021;189:106383. https://doi.org/10.1016/j.compag.2021.106383.

21. Mertan A, Duff DJ, Unal G. Single image depth estimation: An overview. Digit Signal Process. 2021;123:103441. https://doi.org/10.1016/j.dsp.2022.103441.

22. Liu B, Gould S, Koller D. Single image depth estimation from predicted semantic labels. Proc IEEE Conf Comput Vis Pattern Recognit. 2010:1253-60. https://doi.org/10.1109/CVPR.2010.5539823.

23. Karsch K, Liu C, Kang SB. Depthtransfer: Depth extraction from video using non-parametric sampling. IEEE Trans Pattern Anal Mach Intell. 2019;36:2144-58. https://doi.org/10.1109/TPAMI.2014.2316835.

24. Qi X, Liao R, Liu Z, Urtasun R, Jia J. GeoNet: Geometric neural network for joint depth and surface normal estimation. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2018:283-91. https://doi.org/10.1109/CVPR.2018.00037.

25. Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. Adv Neural Inf Process Syst. 2014;3:2366-74. https://arxiv.org/abs/1406.2283v1, accessed August 31, 2023.

26. Laina I, Rupprecht C, Belagiannis V, Tombari F, Navab N. Deeper depth prediction with fully convolutional residual networks. Proc 4th Int Conf 3D Vis (3DV). 2016:239-48. https://doi.org/10.1109/3DV.2016.32.

27. Garg R, Vijay Kumar BG, Carneiro G, Reid I. Unsupervised CNN for single view depth estimation: Geometry to the rescue. Lect Notes Comput Sci. 2016;9912:740-56. https://doi.org/10.1007/978-3-319-46484-8_45.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

28. Hambarde P, Dudhane A, Patil PW, Murala S, Dhall A. Depth estimation from single image and semantic prior. Proc Int Conf Image Process ICIP;2020. p. 1441-5. https://doi.org/10.1109/ICIP40778.2020.9190985.

29. Hambarde P, Murala S, Dhall A. UW-GAN: Single image depth estimation and image enhancement for underwater images. IEEE Trans Instrum Meas. 2021. https://doi.org/10.1109/TIM.2021.3120130.

30. Kumar ACS, Bhandarkar SM, Prasad M. Monocular depth prediction using generative adversarial networks. IEEE/CVF Conf Comput Vis Pattern Recognit (CVPR) Work ;2018. p. 413-21. https://doi.org/10.1109/CVPRW.2018.00068.

31. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM. 2020;63:139-44. https://doi.org/10.1145/3422622.

32. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. Proc 30th IEEE Conf Comput Vis Pattern Recognit CVPR;2017. p. 5967-76. https://doi.org/10.1109/CVPR.2017.632.

33. Divyanth LG, Marzougui A, González-Bernal MJ, McGee RJ, Rubiales D, Sankaran S. Evaluation of effective class-balancing techniques for CNN-based assessment of Aphanomyces root rot resistance in pea (Pisum sativum L.). Sensors. 2022;22:7237. https://doi.org/10.3390/s22197237.

34. Divyanth LG, Guru DS, Soni P, Machavaram R, Nadimi M, Paliwal J. Image-to-image translation-based data augmentation for improving crop/weed classification models for precision agriculture applications. Algorithms. 2022;15:401. https://doi.org/10.3390/a15110401.

35. Creswell A, White T, Dumoulin V, Arulkumaran K, Sengupta B, Bharath AA. Generative adversarial networks: An overview. IEEE Signal Process Mag. 2018;35:53-65. https://doi.org/10.1109/MSP.2017.2765202.

36. Lu Y, Chen D, Olaniyi E, Huang Y. Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review. Comput Electron Agric. 2022;200:107208. https://doi.org/10.1016/j.compag.2022.107208.

37. Davidson C, Jaganathan V, Sivakumar AN, Czarnecki JMP, Chowdhary G. NDVI/NDRE prediction from standard RGB aerial imagery using deep learning. Comput Electron Agric. 2022;203:107396. https://doi.org/10.1016/j.compag.2022.107396.

38. de Lima DC, Saqui D, Mpinda SAT, Saito JH. Pix2Pix network to estimate agricultural near infrared images from RGB data. Can J Remote Sens. 2022;48:299-315. https://doi.org/10.1080/07038992.2021.2016056.

39. Rathore D, Divyanth LG, Reddy KLS, Chawla Y, Buragohain M, Soni P, et al. A two-stage deep-learning model for detection and occlusion-based classification of Kashmiri orchard apples for robotic harvesting. J Biosyst Eng. 2023:1-15. https://doi.org/10.1007/s42853-023-00190-0.

40. Zhang X, Karkee M, Zhang Q. Full-foliage Apple canopies in modern fruiting wall architecture (RGB-D Images). 2023. https://doi.org/10.7273/000004762.

41. Fu L, Karkee M, Zhang Q. Scifresh apple orignial and DepthFilter RGB images. n.d. https://doi.org/10.7273/000001845.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

42. Zhang X, Karkee M, Zhang Q, Whiting MD. Computer vision-based tree trunk and branch identification and shaking points detection in dense-foliage canopy for automated harvesting of apples. J Field Robot. 2021;38:476-93. https://doi.org/10.1002/rob.21998.

43. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. Lect Notes Comput Sci. 2015;9351:234-41. https://doi.org/10.1007/978-3-319-24574-4_28.

44. Li C, Wand M. Precomputed real-time texture synthesis with markovian generative adversarial networks. Lect Notes Comput Sci. 2016;9907:702-16. https://doi.org/10.1007/978-3-319-46487-9_43.

45. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process. 2004;13:600-12. https://doi.org/10.1109/TIP.2003.819861.

46. Saifullah S, Drezewski R. Non-destructive egg fertility detection in incubation using SVM classifier based on GLCM parameters. Procedia Comput Sci. 2022;207:3254-63. https://doi.org/10.1016/j.procs.2022.09.383.

47. Mathew A, Antony A, Mahadeshwar Y, Khan T, Kulkarni A. Plant disease detection using GLCM feature extractor and voting classification approach. Mater Today Proc. 2022;58:407-15. https://doi.org/10.1016/j.matpr.2022.02.350.

48. Koirala A, Walsh KB, Wang Z, McCarthy C. Deep learning – Method overview and review of use for fruit detection and yield estimation. Comput Electron Agric. 2019;162:219-34. https://doi.org/10.1016/j.compag.2019.04.017.

49. Liu T, Kang H, Chen C. ORB-Livox: A real-time dynamic system for fruit detection and localization. Comput Electron Agric. 2023;209:107834. https://doi.org/10.1016/j.compag.2023.107834.

50. Zhang J, He L, Karkee M, Zhang Q, Zhang X, Gao Z. Branch detection for apple trees trained in fruiting wall architecture using depth features and Regions-Convolutional Neural Network (R-CNN). Comput Electron Agric. 2018;155:386-93. https://doi.org/10.1016/j.compag.2018.10.029.

51. Chow LS, Paramesran R. Review of medical image quality assessment. Biomed Signal Process Control. 2016;27:145-54. https://doi.org/10.1016/j.bspc.2016.02.006.

52. Ghrare SE, Ali MAM, Ismail M, Jumari K. Diagnostic quality of compressed medical images: Objective and subjective evaluation. Proc 2nd Asia Int Conf Model Simul AMS. 2008:923-7. https://doi.org/10.1109/AMS.2008.10.

53. Islam NU, Park J. Depth estimation from a single RGB image using fine-tuned generative adversarial network. IEEE Access. 2021;9:32781-94. https://doi.org/10.1109/ACCESS.2021.3060435.

54. Gwn K, Reddy K, Giering M, Bernal EA. Generative adversarial networks for depth map estimation from RGB video. Proc IEEE Comput Vis Pattern Recognit Work ; 2018. p. 1258-66. https://doi.org/10.1109/CVPRW.2018.00163.

55. Chen L, Lin SY, Xie Y, Lin YY, Fan W, Xie X. DGGAN: Depth-image guided generative adversarial networks for disentangling RGB and depth images in 3D hand pose estimation. Proc Winter Conf Appl Comput Vis WACV. 2020:400-8. https://doi.org/10.1109/WACV45572.2020.9093380.

© 2024 The Author(s). | Published by Agro Intelligence Research Press