

Article

Vision-Based Counting of Horticultural Seeds Using ViT-UNet Density Regression for Post-Harvest Quality Assessment

Mengxue Dong ^{1,2}, Chunxiang Zhang ¹, Jiegang Mou ¹, Ziheng Tang ¹, Yiming Zhang ³ and Maosen Xu ^{1,2,*} 

¹ College of Metrology Measurement and Instrument, China Jiliang University, Hangzhou 310018, China; mengxue_dong@cjl.u.edu.cn (M.D.); mjj@cjl.u.edu.cn (J.M.)

² Department of Mechanical Engineering, National University of Singapore, Singapore 117576, Singapore

³ School of Mechanical Engineering, Zhejiang University, Hangzhou 310058, China

* Correspondence: msxu@cjl.u.edu.cn

Highlights

What are the main findings?

- A ViT-UNet density regression framework was developed for point-supervised multi-class counting of cucumber, tomato, and pepper seeds.
- A horticultural seed image dataset containing 231 images and 7789 point-annotated seed instances was constructed and converted into category-specific density maps.
- The proposed model achieved a total-count MAE of 4.0678 and an R^2 of 0.8272 on the held-out test set, while bootstrap confidence intervals and paired statistical tests provided additional support for its lower image-level total-count absolute errors relative to the evaluated deep-learning baselines and the ablation model.

What are the implications of the main findings?

- The proposed method provides a non-destructive visual counting approach for post-harvest horticultural seed quality assessment.
- The method has potential as an efficient visual counting component for high-throughput seed assessment systems, although full-pipeline optimization and validation using independent seed batches are still required.

Abstract

Reliable counting of horticultural seeds is important for seed grading, batch consistency assessment, packaging, sowing-rate control, and post-harvest quality evaluation. Manual counting is labor-intensive, whereas conventional image-processing methods can be sensitive to seed adhesion, visual similarity, background variation, and illumination changes. This study developed a lightweight ViT-UNet density regression framework for point-supervised multiclass counting of cucumber, tomato, and pepper seeds. Vision Transformer (ViT) features were integrated with a U-Net-style decoder to predict category-specific density maps from RGB images. A dataset containing 231 images and 7789 annotated seed instances was constructed, and a held-out test set of 36 images was used for final evaluation. For total-count estimation, the proposed model achieved a mean absolute error (MAE) of 4.0678, root mean squared error (RMSE) of 5.1896, mean absolute percentage error (MAPE) of 14.7104%, and R^2 of 0.8272. Bootstrap resampling yielded a 95% confidence interval of 3.0517–5.1832 for total-count MAE, and paired Wilcoxon signed-rank tests with Holm–Bonferroni correction provided statistical support for lower image-level total-count absolute errors of the proposed model relative to the evaluated deep-learning baselines and the ablation model. The network forward pass reached 94.10 FPS, whereas the full image-processing pipeline reached 2.55 FPS, indicating real-time inference potential at the



Academic Editor: Noelia Castillejo

Received: 24 June 2026

Revised: 20 July 2026

Accepted: 29 July 2026

Published: 3 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

model level but not yet full-pipeline industrial deployment. These results suggest that ViT-UNet density regression is a useful basis for non-destructive multiclass seed counting, while category-level discrimination, external validation, and pipeline optimization remain priorities for future work.

Keywords: computer vision; density regression; horticultural seeds; point annotation; post-harvest quality assessment; seed counting; ViT-UNet

1. Introduction

Seeds are the starting point of horticultural crop production and play a fundamental role in vegetable propagation, breeding, nursery management, and commercial seed supply. For horticultural crops such as cucumber, tomato, and pepper, accurate seed counting is not only a basic measurement task but also an important prerequisite for seed grading, batch consistency evaluation, packaging, sowing-rate control, and post-harvest quality assessment. In practical seed production and processing, large numbers of seeds are often handled in batches, and the number and composition of seeds directly affect downstream management decisions. Manual counting remains common in small-scale operations and laboratory evaluation, but it is time-consuming, labor-intensive, and prone to subjective errors, especially when seeds are small, densely distributed, or visually similar. Recent studies on seed quality assessment and seed phenotyping have shown that computer vision, X-ray imaging, and deep learning provide effective tools for non-destructive seed evaluation and high-throughput seed trait analysis [1–4]. More broadly, deep learning-based image analysis and automated feature extraction have been applied across agricultural automation, food quality assessment, food processing, and biomedical image analysis [5–8]. Therefore, developing a rapid, non-destructive, and accurate seed counting method is important for improving the efficiency and standardization of horticultural seed production and quality evaluation.

Traditional image-processing approaches for seed and grain counting usually rely on threshold segmentation, morphological operations, watershed algorithms, and hand-crafted features. Although these methods are interpretable and easy to implement, their performance often depends strongly on controlled imaging conditions. Uneven illumination, background variation, and seed adhesion can cause segmentation errors that lead to overcounting or undercounting. Deep learning-based object detectors, including Faster R-CNN and YOLO-series models, have improved object localization and recognition performance [9–11]. These methods have also been applied to agricultural and horticultural object detection tasks, including rice grain counting, wheat grain detection, wheat head counting, pest detection, fruit or vegetable quality inspection, and lettuce growth-stage detection [12–15]. Compared with single-category seed-counting studies, relatively few investigations have addressed the simultaneous identification and quantification of multiple seed categories within the same image. Kumari et al. applied YOLOv5 and YOLOv7 to detect and count five cover-crop seed species in mixed-seed images, demonstrating the feasibility of bounding-box-based multi-seed quantification [16]. Ahmed et al. further used YOLOv8 to identify and quantify mixtures of multiple weed-seed species, while also reporting difficulties associated with seeds having similar morphological characteristics [17]. In addition to instance-detection approaches, Mehreen et al. used convolutional neural networks and Vision Transformers to estimate the composition of seed mixtures containing 11 seed varieties [18], although their task focused on image-level composition estimation rather than individual seed counting. These studies demonstrate the potential of deep

learning for multi-category seed-mixture analysis; however, existing approaches mainly rely on bounding-box detection or image-level composition regression. Point-supervised, category-specific density regression for simultaneous counting of visually similar horticultural seeds remains comparatively underexplored. Nevertheless, bounding-box-based detectors may still suffer from missed detections, repeated detections, and inaccurate localization when targets are small, densely distributed, or visually similar. These problems are particularly relevant to horticultural seed counting, where tomato and pepper seeds may show similar color and texture, whereas cucumber seeds differ substantially in size and shape.

Density map regression offers an alternative route for dense object counting. Instead of detecting each instance with a bounding box, the model predicts a continuous density map, and the final count is obtained by integrating the predicted density values. This strategy has been widely used in crowd counting and has increasingly been adapted for plant and agricultural counting tasks [19–23]. U-Net-based and hybrid U-Net architectures have also been explored for agricultural object localization and counting, demonstrating the value of encoder–decoder structures for recovering spatial details in plant images [24]. More recent studies have extended density estimation and deep learning methods to seed kernel counting and in situ seed counting, indicating that density-based learning is suitable for high-throughput seed counting and phenotyping applications [25]. Compared with detection-based methods, density regression is more suitable for densely distributed small objects and can be trained using point-level annotations, thereby reducing annotation cost compared with bounding-box or pixel-level mask annotations. Point-supervised object counting studies further demonstrate that center-point annotations can provide effective supervision for counting and localization [26,27]. For horticultural seed counting, density regression is especially attractive because each seed can be annotated by only one center point, and dense or partially touching seed distributions can be handled without explicitly separating every object boundary. In density-regression-based counting, ground-truth density maps are commonly generated by placing Gaussian responses at annotated object centers. For multiclass horticultural seeds, different categories may differ in apparent size, shape, and local visual scale. Because cucumber seeds were generally larger and more elongated than tomato and pepper seeds, category-specific Gaussian standard deviations were used as part of the density-map generation protocol.

Another challenge lies in feature representation. Convolutional neural networks and U-Net-like architectures are effective for local feature extraction and spatial reconstruction, but convolutional operations are inherently local and may have limited ability to model long-range contextual relationships. In multiclass seed images, local texture information alone may be insufficient to distinguish visually similar categories, especially under dense arrangements or variable imaging conditions. Vision Transformers have shown strong capability in global context modeling through self-attention mechanisms. Transformer-based attention has been increasingly explored for global contextual representation in computer-vision tasks [28]. However, directly applying pure Transformer models to relatively small horticultural image datasets may be challenging because Transformers often require large-scale training data and may lose fine spatial details during patch embedding. A hybrid architecture that combines a lightweight ViT encoder with a U-Net-style decoder offers a balanced design: the ViT encoder captures global relationships among seed instances, while the decoder and skip connections recover spatial detail for density-map prediction. The effectiveness of skip connections in training deep neural networks and preserving multilevel feature information has also been widely discussed [29]. Transformer-based attention and hybrid convolution–Transformer designs have also been explored for multiscale feature representation and image restoration [28,30].

The novelty of this study does not lie in proposing a completely new Transformer block. Instead, it lies in adapting and systematically validating a lightweight ViT-UNet density regression framework for point-supervised multiclass horticultural seed counting. The contribution includes category-specific density-map generation, comparison with conventional and deep-learning baselines, bootstrap confidence intervals, paired statistical testing, descriptive stratified analysis under different imaging conditions, and category-level false-positive analysis.

To address these challenges, this study proposes a vision-based multiclass counting method for horticultural seeds using ViT-UNet density regression. Cucumber, tomato, and pepper seeds were selected as representative vegetable seed categories because they differ in size, shape, and visual appearance. A point-annotation-based image dataset was constructed to support density-map supervision without requiring bounding-box or pixel-level mask annotations. The proposed framework integrates a lightweight ViT-Tiny encoder with a U-Net-style decoder and skip connections to predict three-channel density maps corresponding to cucumber, tomato, and pepper seeds. The final category-wise and total seed counts are obtained by integrating the predicted density maps.

2. Materials and Methods

2.1. Experimental Materials and Image Acquisition Setup

Three representative horticultural seed categories were used in this study: cucumber seeds, tomato seeds, and pepper seeds. These vegetable seed categories were selected because they are commonly used in horticultural production and differ in size, shape, and visual appearance. Cucumber seeds are relatively large and elongated, whereas tomato and pepper seeds are smaller and more similar in local color and texture, making multiclass visual counting challenging. Approximately 200 cucumber seeds, 400 tomato seeds, and 400 pepper seeds were prepared for image acquisition. Damaged, incomplete, or severely deformed seeds were removed before imaging. All seeds were kept dry under room conditions and were repeatedly rearranged during image acquisition to generate images with different densities and spatial distributions. Consequently, the same physical seed lots appeared in multiple images, increasing spatial diversity but not providing independent seed-batch replication.

A standardized image acquisition setup was established to simulate the visual sensing stage of horticultural seed quality assessment and post-harvest handling. Seed images were captured using an iPhone 15 smartphone camera. The rear camera incorporates a 48-megapixel image sensor, and the images used in this study were recorded at a resolution of approximately 5712×4284 pixels with a 4:3 aspect ratio. All images were saved in JPEG format. During image acquisition, the smartphone was mounted vertically above the seed plane using a fixed support to minimize perspective distortion. The distance between the smartphone camera and the seed plane was approximately 15 cm. Within each predefined imaging condition, the imaging distance and illumination setup were kept consistent. Controlled variations in illumination were introduced to include normal, weak-light, and shadow conditions. LED illumination was used as the main light source. Image acquisition was conducted at China Jiliang University, Hangzhou, China.

To improve the diversity of the image dataset, images were collected under different backgrounds, illumination, and nominal seed-density conditions. Three background conditions were considered: white, black, and blue. Illumination conditions included normal, weak-light, and shadow conditions. During image acquisition, each image was assigned a nominal low-, medium-, or high-density label according to the acquisition setting recorded for that image. These labels represented experimental acquisition conditions and were not generated retrospectively by applying fixed global thresholds to the annotated total

seed counts. Seeds were randomly distributed within the imaging area, and partial adhesion between seeds was allowed, but severe overlap that made category identification impossible was avoided. All images were named sequentially using a uniform naming rule. Before annotation and density map generation, all acquired images were checked for EXIF orientation information and standardized to ensure consistency among image display, image size, and annotation coordinates. Representative seed images acquired under different imaging conditions are shown in Figure 1.

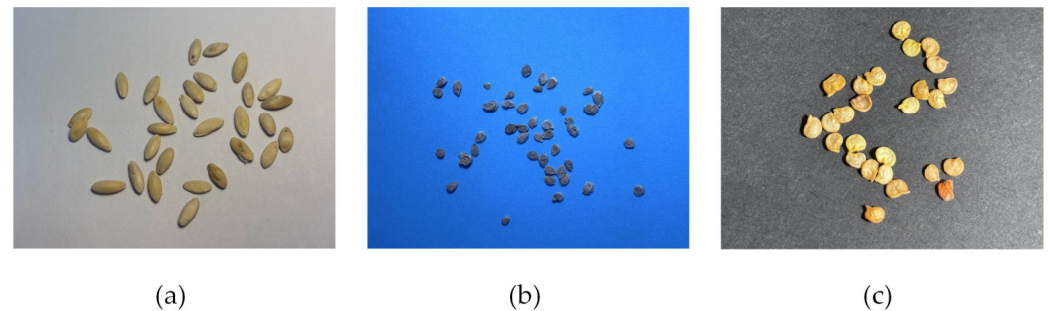


Figure 1. Representative horticultural seed images. (a) cucumber seeds; (b) tomato seeds; (c) pepper seeds.

2.2. Dataset Construction, Point Annotation, and Density Map Generation

The dataset was constructed using point-level annotations. For each visible seed in an image, one center point and one category label were manually annotated. The three category labels were defined as cucumber, tomato, and pepper. Compared with bounding-box annotation or pixel-level mask annotation, point-level annotation is more efficient for dense small-object counting tasks because each seed instance only requires one center coordinate. Therefore, bounding boxes and pixel-level masks were not used in this study. After annotation, the number of annotated points in each category was checked against manual counts to reduce missing annotations, duplicated annotations, and incorrect category labels. Before density map generation, invalid category labels, duplicate points, and out-of-bounds coordinates were checked. Completely duplicated annotation records were removed, and suspicious annotations were manually inspected. Images with ambiguous category labels, severe occlusion, or indistinguishable seed boundaries were excluded from the dataset.

In addition, a metadata file was prepared for each image to record the background color, illumination condition, and nominal seed-density acquisition condition. Specifically, three background conditions were recorded: white, black, and blue; illumination conditions included normal, weak-light, and shadow conditions; and the nominal seed-density condition was recorded as low, medium, or high according to the acquisition setting used for each image. These labels were retained for descriptive stratified analysis and were not retrospectively generated from fixed total-count thresholds.

The point annotations were converted into class-specific density maps for supervised learning. Each output channel corresponded to one seed category, namely cucumber, tomato, and pepper. For each annotated center point, a two-dimensional Gaussian response was generated at the corresponding location. Considering the scale differences among the three seed categories, different Gaussian standard deviations were assigned to different categories. Specifically, a relatively larger Gaussian standard deviation was used for cucumber seeds, while smaller Gaussian standard deviations were used for tomato and pepper seeds. The category-specific density maps were initially generated at the native image resolution using the original point-annotation coordinates. At this resolution, the Gaussian standard deviations for cucumber, tomato, and pepper were set to 3.5, 2.5, and

2.5 pixels, respectively. These class-specific values were empirically selected according to the approximate visual scales of the three seed categories and were fixed for all samples. At the native image resolution, numerical verification confirmed that the sum of each category-specific density map was consistent with the corresponding ground-truth count, up to negligible numerical error.

For category c , the Gaussian response generated from the k -th annotated point was defined as Equation (1):

$$G_{c,k}(x,y) = \frac{1}{Z_{c,k}} \exp \left[-\frac{(x-x_{c,k})^2 + (y-y_{c,k})^2}{2\sigma_c^2} \right] \quad (1)$$

where $(x_{c,k}, y_{c,k})$ is the center coordinate of the k -th seed belonging to category c , σ_c is the category-specific Gaussian standard deviation, and $Z_{c,k}$ is the normalization factor. The normalization factor was calculated as in Equation (2), so that the discrete integral of each single-seed Gaussian response was equal to 1:

$$Z_{c,k} = \sum_{x=1}^W \sum_{y=1}^H \exp \left[-\frac{(x-x_{c,k})^2 + (y-y_{c,k})^2}{2\sigma_c^2} \right] \quad (2)$$

The category-specific ground-truth density map was generated by summing all normalized Gaussian responses belonging to the same category, as shown in Equation (3):

$$D_c(x,y) = \sum_{k=1}^{N_c} G_{c,k}(x,y), \quad c \in \{\text{cucumber, tomato, pepper}\} \quad (3)$$

At the native image resolution, the discrete normalization was designed to make the sum of each category-specific density map consistent with the corresponding ground-truth count. The area-based correction used to approximately preserve the density-map integral during subsequent resizing is described in Section 2.3. This relationship is expressed as Equation (4):

$$N_c = \sum_{x=1}^W \sum_{y=1}^H D_c(x,y) \quad (4)$$

At the native image resolution, this unit-integral normalization established consistency between the point annotations and the discrete density-map sums and compensated for truncation of Gaussian kernels near image boundaries. During inference, the predicted number of seeds in each category was obtained by integrating the corresponding predicted density map. Thus, the proposed density regression framework transformed the seed counting task into a multichannel density map prediction problem. The complete workflow of point annotation and class-specific density map generation is illustrated in Figure 2.

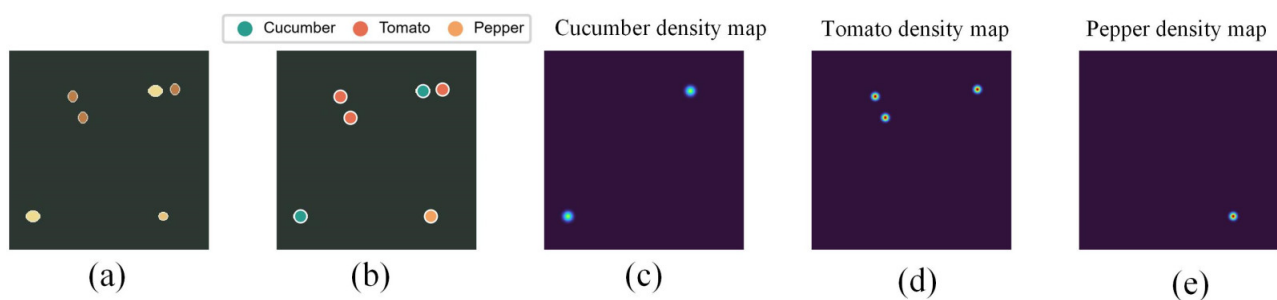


Figure 2. Point annotation and density map generation workflow. (a) Raw RGB seed image; (b) point-level annotations with category labels for cucumber, tomato, and pepper seeds; (c) cucumber ground-truth density map; (d) tomato ground-truth density map; (e) pepper ground-truth density map. The category-wise counts were obtained by integrating the corresponding density maps, and the total seed count was calculated as $N_{\text{total}} = N_{\text{cucumber}} + N_{\text{tomato}} + N_{\text{pepper}}$.

2.3. Dataset Preprocessing and Splitting

All raw images and their corresponding density maps were converted into standardized sample folders. Each sample folder contained the standardized image, three category-specific density maps, category counts, point annotations, and metadata. This organization ensured that image data, annotation data, density maps, and experimental condition labels were matched at the sample level.

The dataset was divided at the image level into training, validation, and held-out test sets using a ratio of 70%, 15%, and 15%, respectively. This procedure prevented the same image from appearing in more than one subset, and a fixed random seed of 42 was used for dataset splitting. The distributions of seed categories, background conditions, illumination conditions, and nominal seed-density acquisition conditions were kept as balanced as possible among the three subsets. The validation set was used for model selection and hyperparameter tuning, whereas the held-out test set was reserved for final performance evaluation. Because the same physical seed lots were repeatedly rearranged during image acquisition, this split represents image-level validation rather than independent seed-batch validation.

Before training, each RGB image and its corresponding category-specific density maps were resized synchronously from the native image resolution to 224×224 pixels. The RGB images were resized using bilinear interpolation with antialiasing enabled, whereas the density maps were resized using bilinear interpolation. After interpolation, each density map was multiplied by the area correction factor $H_0W_0/(224 \times 224)$, where H_0 and W_0 denote the height and width of the original image, respectively. This area-based correction was used to approximately preserve the discrete density-map integral during resizing and maintain consistency with the corresponding ground-truth seed count. Data augmentation was applied only to the training set and included horizontal flipping, rotation within $\pm 5^\circ$, and photometric adjustments to brightness, contrast, and saturation. Horizontal flipping and rotation were applied synchronously to the RGB images and corresponding density maps, whereas the photometric adjustments were applied only to the RGB images. No random cropping was used. The area correction described above was applied only during the initial resizing operation; no additional density-map renormalization was performed after rotation. The validation and held-out test sets were resized and normalized without random augmentation.

2.4. ViT-UNet Density Regression Model

A lightweight ViT-UNet density regression model was developed for multiclass horticultural seed counting. The model consisted of a ViT-Tiny encoder and a U-Net-style decoder. The input to the model was an RGB image with a size of 224×224 pixels, and the output was a three-channel density map corresponding to cucumber, tomato, and pepper seeds. The mapping from the input image to the predicted density map was formulated as Equation (5):

$$\hat{D} = f_{\theta}(I) \quad (5)$$

where I denotes the input RGB image, f_{θ} denotes the ViT-UNet model with trainable parameters θ , and $\hat{D}$ denotes the predicted three-channel density map.

The ViT-Tiny encoder used a patch size of 16×16 pixels to transform the input image into patch tokens. Intermediate features from multiple Transformer blocks were extracted and reshaped into two-dimensional feature maps. These multilevel features were introduced into the decoder through skip connections. The decoder progressively restored the spatial resolution by upsampling and feature fusion, thereby producing high-resolution density maps. A Softplus activation function was used at the output layer to ensure nonnegative density predictions. The overall architecture of the proposed ViT-UNet density regression model is illustrated in Figure 3.

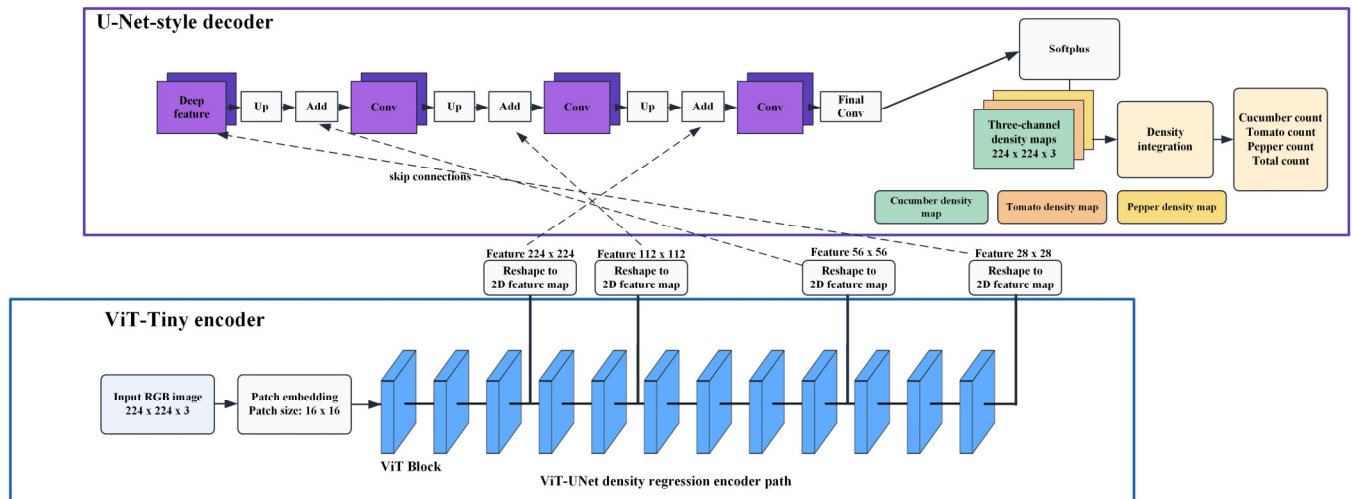


Figure 3. Architecture of the proposed ViT-UNet density regression model for multiclass horticultural seed counting. The RGB input image is transformed into patch tokens and encoded by a ViT-Tiny encoder. Multilevel feature representations are reshaped into two-dimensional feature maps and introduced into a U-Net-style decoder through skip connections. The decoder restores spatial resolution through upsampling, feature fusion, and convolution, followed by a Softplus layer that outputs three-channel density maps for cucumber, tomato, and pepper seeds. Category-wise counts are obtained by density integration, and the total count is calculated by summing the category-wise counts.

The model takes a 224×224 RGB image as input and predicts three-channel density maps for cucumber, tomato, and pepper seeds. Category-wise counts are obtained by integrating the corresponding predicted density maps, and the total seed count is calculated by summing the category-wise counts.

The predicted count of category c was obtained by integrating the predicted density map of that category, as shown in Equation (6):

$$\hat{N}_c = \sum_{x=1}^W \sum_{y=1}^H \hat{D}_c(x, y) \quad (6)$$

The total predicted seed count was calculated as the sum of the predicted category-wise counts, as shown in Equation (7):

$$\hat{N}_{\text{total}} = \sum_{c=1}^C \hat{N}_c, \quad C = 3 \quad (7)$$

The use of ViT-Tiny was intended to balance global feature representation and computational efficiency. The Transformer encoder captured long-range contextual relationships among seed instances, while the U-Net-style decoder recovered spatial details required for accurate density map prediction. This hybrid design is suitable for dense small-object counting tasks in horticultural seed images.

2.5. Training Strategy

The models were trained using supervised learning with the generated three-channel density maps as ground truth. The density map regression loss was defined as the mean squared error between the predicted and ground-truth density maps, as shown in Equation (8):

$$\mathcal{L}_{\text{den}} = \frac{1}{CHW} \sum_{c=1}^C \sum_{x=1}^W \sum_{y=1}^H (\hat{D}_c(x, y) - D_c(x, y))^2 \quad (8)$$

The final category-wise and total counts were obtained during evaluation by integrating the predicted density maps.

The optimizer was AdamW with an initial learning rate of 1×10^{-4} and a weight decay of 0.01. The batch size was set to 4, and the models were trained for 200 epochs. A cosine annealing learning rate scheduler was used to adjust the learning rate during training. The checkpoint with the lowest validation loss was selected for final evaluation on the held-out test set. The model backbone was configured as vit_tiny_patch16_224, and the ViT-Tiny encoder was initialized using the corresponding ImageNet-pretrained weights provided by timm. The archived execution environment contained Python 3.13.5, PyTorch 2.12.1, torchvision 0.27.1, timm 1.0.27, NumPy 2.4.6, Pillow 12.2.0, and OpenCV 4.13.0. The experiments were executed in a CUDA-enabled Windows environment on an NVIDIA GeForce RTX 4060 Ti GPU. A fixed random seed of 42 was used for dataset splitting and model training.

2.6. Evaluation Metrics

The counting performance was evaluated at both category-wise and total-count levels. For each seed category, the predicted count and ground-truth count were compared using mean absolute error, root mean squared error, mean absolute percentage error, and coefficient of determination. The same metrics were also calculated for the total seed count.

The mean absolute error was defined as Equation (9):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\hat{N}_i - N_i| \quad (9)$$

The root mean squared error was defined as Equation (10):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{N}_i - N_i)^2} \quad (10)$$

The mean absolute percentage error was calculated as Equation (11):

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^n \frac{|\hat{N}_i - N_i|}{N_i} \quad (11)$$

The coefficient of determination was calculated as Equation (12):

$$R^2 = 1 - \frac{\sum_{i=1}^n (N_i - \hat{N}_i)^2}{\sum_{i=1}^n (N_i - \bar{N})^2} \quad (12)$$

where n is the number of test images, $\hat{N}_i$ is the predicted count, N_i is the ground-truth count of the i -th image, and $\bar{N}$ is the mean ground-truth count. For total-count MAPE, all test images were included because the total ground-truth count was nonzero. For category-wise MAPE, images with zero ground-truth count for the corresponding category were excluded, and the denominator n was replaced by the number of valid nonzero samples.

To evaluate the effectiveness of the proposed architecture, several baseline and ablation models were implemented. A lightweight U-Net was used as a convolutional baseline. A pure Transformer-based density regression model was used to evaluate the effect of using a Transformer architecture without U-Net-style spatial reconstruction. A ViT-UNet model without intermediate skip connections was constructed to analyze the contribution of multilevel feature fusion. The final proposed model was the ViT-UNet with skip connections. All baseline and ablation models were trained using the same dataset split, input resolution, optimizer, learning rate, batch size, number of epochs, and evaluation metrics to ensure fair comparison.

To quantify uncertainty, support nonparametric statistical comparison, and provide conventional reference methods, additional analyses were conducted using the same held-out test set. Bootstrap confidence intervals were calculated by resampling the 36 test images with replacement for 10,000 iterations. For each bootstrap replicate, MAE, RMSE, MAPE,

and R^2 were recalculated. The 2.5th and 97.5th percentiles of the bootstrap distribution were used as the 95% confidence interval. Bootstrap replicates with zero variance in the ground-truth total count were excluded only for R^2 calculation.

Paired statistical comparisons were performed using image-level total absolute errors. The proposed ViT-UNet was compared with U-Net, Pure Transformer, and ViT-UNet without skip connections using the Wilcoxon signed-rank test. The paired difference was defined as the baseline absolute error minus the proposed model absolute error, so positive values indicate lower error for the proposed model. Holm–Bonferroni correction was applied to control the family-wise error rate across the three pairwise comparisons. These paired tests quantified image-level error differences for the selected model checkpoints but did not capture variability across repeated training runs with different random seeds or across alternative dataset splits.

Two conventional image-processing baselines were additionally implemented as non-learning reference methods for total seed counting: Otsu thresholding followed by morphological processing and connected-component counting, and adaptive thresholding followed by distance-transform-based watershed segmentation. These methods used fixed heuristic area filtering and were not supervised or tuned using the test labels. They were included as practical reference methods for evaluating density-regression-based counting under mixed background, illumination, and nominal seed-density acquisition conditions.

Descriptive stratified analyses were conducted for the proposed ViT-UNet according to background color, illumination condition, and nominal seed-density acquisition condition. Because several subgroups contained limited numbers of test images, the subgroup results were interpreted descriptively rather than as strong evidence of robustness. Category-level presence/absence and zero-ground-truth false-positive analyses were also performed as auxiliary analyses of category confusion and were reported in the Supplementary Materials.

3. Results

3.1. Dataset Composition and Annotation Characteristics

The constructed horticultural seed image dataset consisted of 231 images with 7789 manually annotated seed instances. Three representative vegetable seed categories were included: cucumber, tomato, and pepper. Among all annotated seed instances, cucumber seeds accounted for 2805 instances, tomato seeds for 2477 instances, and pepper seeds for 2507 instances, corresponding to 36.01%, 31.80%, and 32.19% of the total annotations, respectively. The relatively balanced distribution among the three categories provided a suitable basis for evaluating multiclass seed-counting performance.

The three seed categories showed clear differences in morphology and visual appearance. Cucumber seeds were generally larger, elongated, and visually more distinguishable, whereas tomato and pepper seeds were smaller and more similar in color and local texture. This difference in inter-class separability makes the dataset suitable for evaluating not only total-count estimation but also the model's ability to distinguish visually similar seed categories in mixed samples.

The dataset also included images collected under different imaging conditions. Specifically, 105 images were acquired using a white background, 63 using a blue background, and 63 using a black background. Regarding illumination conditions, 141 images were collected under normal illumination, 45 under weak illumination, and 45 under shadow illumination. According to the recorded acquisition-condition labels, the dataset contained 65 images under the nominal low-density condition, 83 under the nominal medium-density condition, and 83 under the nominal high-density condition. These acquisition conditions increased the diversity of the dataset and supported a descriptive examination of performance variation across imaging conditions.

The dataset was divided at the image level into training, validation, and held-out test subsets. The held-out test set contained 36 images and was reserved for final model evaluation. Because the same physical seed lots were rearranged across multiple images, this evaluation should be interpreted as image-level rather than independent seed-batch validation. The dataset composition and annotation statistics are summarized in Table 1.

Table 1. Dataset composition and annotation statistics.

Item	Value
Total images	231
Total annotated seed instances	7789
Cucumber instances	2805
Tomato instances	2477
Pepper instances	2507
White-background images	105
Blue-background images	63
Black-background images	63
Normal-illumination images	141
Weak-light images	45
Shadow-illumination images	45
Images under the nominal low-density condition	65
Images under the nominal medium-density condition	83
Images under the nominal high-density condition	83
Held-out test images	36

Overall, the dataset provided a practical multiclass benchmark for horticultural seed counting. The inclusion of three seed categories with different visual characteristics enabled the evaluation of both category-wise and total-count prediction. In addition to overall performance, model comparison, error distribution, and computational efficiency, descriptive stratified analyses were conducted to examine performance variation across background, illumination, and nominal seed-density acquisition conditions.

3.2. Overall and Category-Wise Counting Performance

The proposed ViT-UNet model was evaluated on the held-out test set using MAE, RMSE, MAPE, and R^2 . The model produced three category-specific density-map outputs corresponding to cucumber, tomato, and pepper seeds. Category-wise counts were obtained by integrating each predicted density map, and the total seed count was calculated by summing the three category-wise predictions.

The quantitative performance of the proposed model is shown in Table 2. For total seed counting, the proposed ViT-UNet achieved an MAE of 4.0678, RMSE of 5.1896, MAPE of 14.7104%, and R^2 of 0.8272. These results indicate that the model captured the main variation trend of total seed counts across the held-out test set and achieved a moderate absolute counting error. A total-count MAE of approximately four seeds per image suggests that the proposed density-regression framework can provide useful total-count estimates for batch-level seed assessment.

Table 2. Counting performance of the proposed ViT-UNet model on the held-out test set with bootstrap 95% confidence intervals.

Category	MAE (95% CI)	RMSE (95% CI)	MAPE (%) (95% CI)	R^2 (95% CI)
Cucumber	3.27 (2.61–4.01)	3.91 (3.05–4.82)	23.22 (13.44–37.45)	0.915 (0.863–0.943)
Tomato	2.90 (2.40–3.42)	3.29 (2.81–3.75)	30.36 (12.65–61.31)	0.865 (0.726–0.922)
Pepper	3.50 (2.86–4.20)	4.05 (3.34–4.78)	18.55 (14.03–23.10)	0.913 (0.866–0.938)
Total	4.07 (3.05–5.18)	5.19 (4.00–6.29)	14.71 (9.64–21.89)	0.827 (0.683–0.903)

Note: Bootstrap confidence intervals were calculated using 10,000 image-level resamples. Category-wise MAPE excluded images with zero ground-truth count for the corresponding category.

Bootstrap analysis quantified the uncertainty of the proposed model's performance estimates on the held-out test set. For total count, MAE was 4.07 seeds (95% CI: 3.05–5.18), RMSE was 5.19 (95% CI: 4.00–6.29), MAPE was 14.71% (95% CI: 9.64–21.89%), and R^2 was 0.827 (95% CI: 0.683–0.903).

At the category level, the model achieved an MAE of 3.27, RMSE of 3.91, MAPE of 23.22%, and R^2 of 0.915 for cucumber seeds. For tomato seeds, the MAE, RMSE, MAPE, and R^2 were 2.90, 3.29, 30.36%, and 0.865, respectively. For pepper seeds, the corresponding values were 3.50, 4.05, 18.55%, and 0.913. The R^2 values of all three categories were above 0.86, indicating that the model captured the general count variation of each seed category.

The category-wise results also reveal differences in counting difficulty among the three seed types. Although tomato had the lowest MAE, it showed the highest MAPE. This suggests that percentage error was more sensitive to samples with low tomato counts. In addition, tomato and pepper seeds were visually more similar than cucumber seeds, which may have increased category-level ambiguity in some mixed samples. By contrast, cucumber seeds were larger and more elongated, making their visual features easier to distinguish.

The total-count MAPE was lower than most category-wise MAPE values, indicating that category-level errors were partly compensated when the three predicted counts were summed. However, the total-count R^2 was slightly lower than the category-wise R^2 values, suggesting that total-count performance should be interpreted jointly using both percentage error and goodness-of-fit metrics.

3.3. Prediction Agreement Between Ground-Truth and Predicted Counts

To further evaluate prediction consistency, the predicted counts were compared with the corresponding manually annotated ground-truth counts for each test image. Figure 4 shows the predicted-versus-ground-truth scatter plots for cucumber, tomato, pepper, and total seed counts. The dashed 1:1 reference line represents ideal agreement between prediction and ground truth.

Predicted counts showed clear positive relationships with ground-truth counts for all three categories and for the total count. The total-count prediction achieved an R^2 of 0.827, indicating that the model captured the overall variation in seed number across the held-out test set. Total-count prediction showed a lower percentage error than most category-wise predictions, suggesting that some category-level deviations were partly compensated after summing the three output channels.

However, Figure 4 also shows that prediction errors were not uniformly distributed. Several samples deviated from the 1:1 reference line, especially in category-wise predictions. Tomato and pepper showed nonzero predicted counts in some samples with zero ground-truth counts, suggesting the presence of inter-class false-positive responses. This observation is consistent with the visual similarity between tomato and pepper seeds. In addition, for some high-count samples, cucumber, pepper, and total predictions showed a tendency toward underestimation, suggesting that dense arrangements or crowded seed distributions may remain challenging.

These scatter plots provide evidence beyond aggregate metrics. While Table 2 summarizes average performance, Figure 4 reveals sample-level agreement and potential bias patterns. Therefore, the model achieved meaningful overall count prediction but still showed category-level ambiguity and high-count underestimation in some difficult cases.

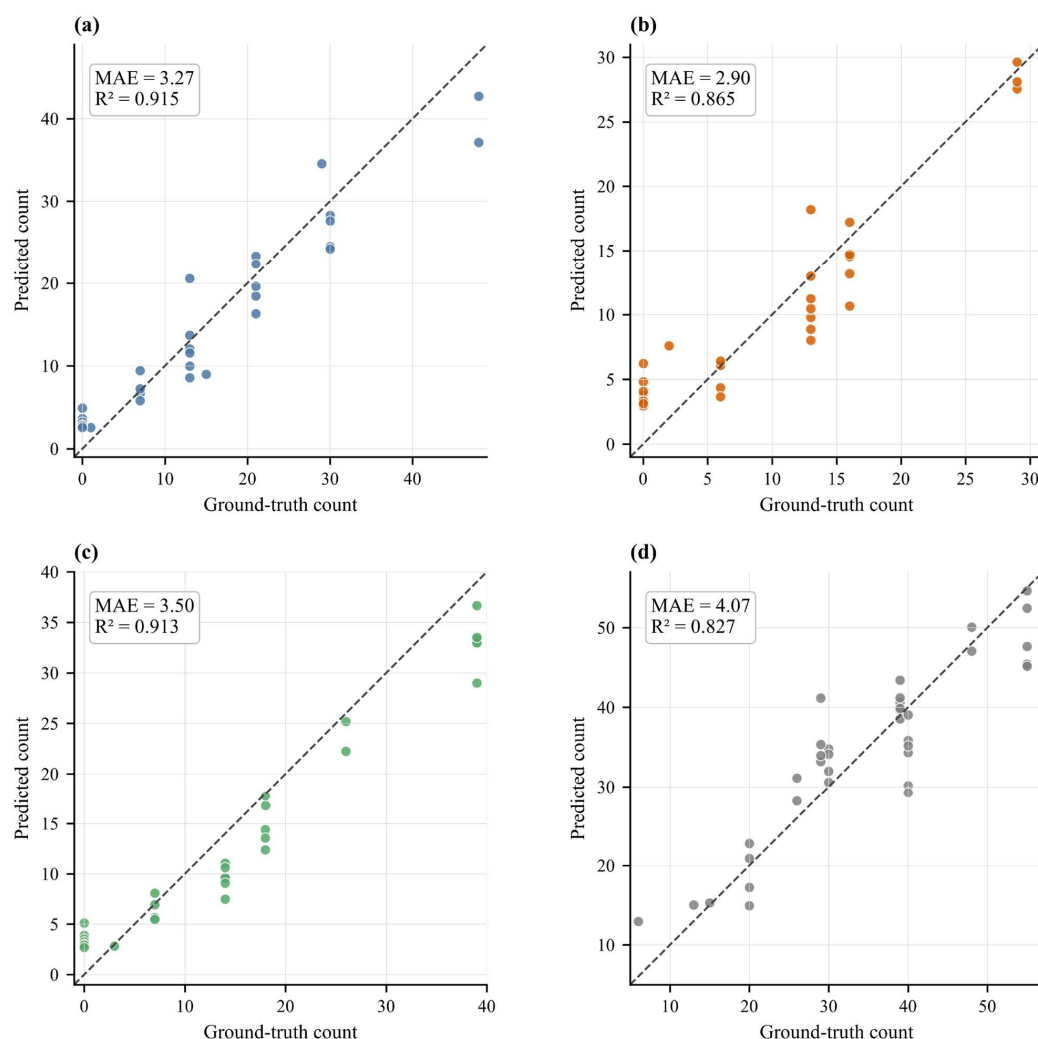


Figure 4. Predicted versus ground-truth seed counts on the held-out test set: (a) cucumber; (b) tomato; (c) pepper; (d) total. The dashed line represents the 1:1 reference line.

3.4. Comparison with Baseline and Ablation Models

To evaluate the contribution of the proposed architecture and to provide conventional non-learning references, the proposed ViT-UNet was compared with two conventional image-processing baselines, two deep-learning baselines, and one ablation model. The deep-learning baselines were U-Net and Pure Transformer, whereas the ablation model was ViT-UNet without skip connections. The conventional baselines included Otsu thresholding with connected-component counting and adaptive thresholding with watershed segmentation. The comparison results are summarized in Table 3.

Table 3. Comparison with conventional image-processing baselines, deep-learning baselines, and the ablation model on the held-out test set.

Model	Total MAE	Total RMSE	Total MAPE (%)	Total R^2
Otsu + connected components	21.89	24.47	61.46	−3.496
Adaptive thresholding + watershed	10.92	13.77	34.24	−0.424
U-Net	8.65	11.44	25.81	0.161
Pure Transformer	22.49	25.37	62.80	−3.130
ViT-UNet w/o skip connections	8.49	11.03	31.40	0.219
Proposed ViT-UNet	4.07	5.19	14.71	0.827

Note: Conventional baselines were evaluated only for total-count reference performance and used fixed heuristic segmentation settings.

The conventional image-processing methods were less accurate than the proposed density-regression model. Otsu thresholding with connected components produced a total-count MAE of 21.89 and a negative R^2 of -3.496 , while adaptive thresholding with watershed segmentation reduced the MAE to 10.92 but still remained substantially less accurate than the proposed ViT-UNet. These results indicate that fixed segmentation rules were sensitive to touching seeds, background color, and illumination variation.

Compared with U-Net, the proposed ViT-UNet reduced the total MAE by approximately 53.0%. Compared with Pure Transformer, the total MAE was reduced by approximately 81.9%. Compared with ViT-UNet without skip connections, the total MAE was reduced by approximately 52.1%. The same performance trend was also reflected by MAPE and R^2 , with the proposed ViT-UNet producing the lowest percentage error and the highest coefficient of determination. Under the present dataset split and evaluation conditions, these results indicate that the proposed architecture was associated with improved counting accuracy and prediction consistency.

The weak performance of the Pure Transformer model suggests that Transformer-based global representation alone was insufficient for this high-resolution density-regression task. Although self-attention can capture global relationships, a pure Transformer architecture may lose fine spatial information during patch embedding and may not effectively reconstruct local density responses. The lower performance of ViT-UNet without skip connections further supports the contribution of multilevel feature fusion under the present evaluation conditions. Without skip connections, the decoder mainly relies on deeper semantic features, while shallow spatial details are less effectively preserved. The superior performance of the proposed ViT-UNet therefore supports the benefit of combining global feature extraction, U-Net-style decoding, and skip-connection-based spatial feature fusion for the present seed-counting task.

3.5. Error Distribution Analysis

To further examine the error characteristics of the proposed model, the distribution of prediction errors was analyzed, as shown in Figure 5. The figure includes the absolute error distribution, total-count residual pattern, histogram of total-count absolute error, and signed error distribution for category-wise and total-count predictions. The absolute error distribution shows that most category-wise and total-count errors were concentrated within a moderate range, while a small number of samples produced larger deviations. This pattern indicates that the model performed reliably for many test samples but still encountered difficult cases with substantially larger errors. Such cases may be related to dense seed arrangements, category similarity, or local visual ambiguity.

The residual plot provides additional insight into the relationship between ground-truth total count and prediction error. Several high-count samples showed negative residuals, indicating a tendency toward underestimation when the number of seeds was large. This observation is consistent with the deviations from the 1:1 line observed in Figure 4. Dense samples may cause local seed appearances to merge, making individual contributions to the predicted density response less separable.

The signed error distribution further suggests that prediction bias differed among categories. Category-wise signed errors were more variable than total-count errors, reflecting the difficulty of simultaneously localizing and classifying visually similar seeds. In particular, tomato and pepper predictions were more likely to show category-level deviations, consistent with the false-positive patterns observed in Figure 4. These results indicate that total-count prediction was less affected by inter-class similarity than category-wise prediction.

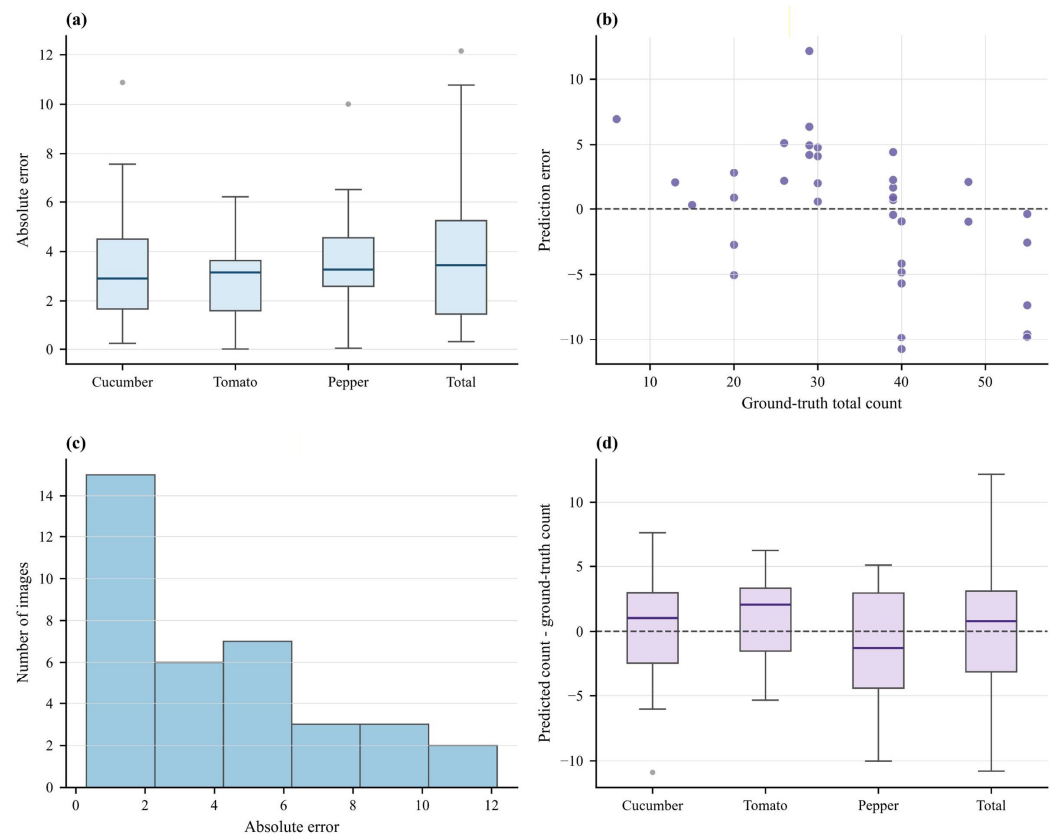


Figure 5. Error distribution analysis of the proposed ViT-UNet on the held-out test set: (a) absolute error; (b) total-count residuals; (c) total-count absolute error; (d) signed error.

Overall, the error distribution analysis indicates that the proposed ViT-UNet achieved useful total-count estimation but still produced larger errors for certain difficult samples. These findings provide an empirical basis for further improving category-level discrimination and counting performance for high-count or crowded samples.

A descriptive stratified analysis was further conducted according to background color, illumination condition, and nominal seed-density acquisition condition. Total-count MAE ranged from 3.43 to 5.05 across background conditions, from 4.05 to 6.56 across illumination conditions, and from 2.30 to 6.08 across the nominal seed-density acquisition conditions. Because some strata contained fewer than five samples, these subgroup results are reported as descriptive evidence in Supplementary Table S2 rather than as a main inferential performance table.

Auxiliary category-level presence/absence and zero-ground-truth analyses indicated systematic nonzero residual category-channel responses when a specific seed class was absent, revealing residual inter-class cross-talk in the current density-regression framework. These results are reported in Supplementary Table S3 and should be interpreted cautiously as evidence of category-level cross-talk rather than as a primary classification metric.

3.6. Statistical Comparison of Model Errors

Paired Wilcoxon signed-rank tests were performed using image-level total-count absolute errors to determine whether the observed differences between the proposed ViT-UNet and the evaluated deep-learning baselines or the ablation model were statistically supported. The statistical comparison results are summarized in Table 4. After Holm–Bonferroni correction, the proposed model showed lower total-count absolute error than U-Net, Pure Transformer, and ViT-UNet without skip connections. The mean paired error reductions were 4.58, 18.42, and 4.42 seeds per image, respectively.

Table 4. Paired Wilcoxon signed-rank comparisons of image-level total-count absolute errors between the proposed ViT-UNet and each deep-learning baseline or ablation model.

Comparison	Mean Paired Difference	Median Difference	Wilcoxon Statistic	Raw <i>P</i>	Holm-Adjusted <i>P</i>
Proposed ViT-UNet vs U-Net	4.58	2.68	124.00	<0.001	<0.001
Proposed ViT-UNet vs Pure Transformer	18.42	17.07	0.00	<0.001	<0.001
Proposed ViT-UNet vs ViT-UNet w/o skip connections	4.42	5.42	115.00	<0.001	<0.001

Note: All comparisons were based on 36 paired test images. Positive paired differences indicate lower image-level total-count absolute errors for the proposed ViT-UNet. *P* values below 0.001 are reported as <0.001. All comparisons remained statistically significant after Holm–Bonferroni correction.

3.7. Computational Efficiency and Forward-Inference Throughput

The computational efficiency of the proposed ViT-UNet model was evaluated on an NVIDIA GeForce RTX 4060 Ti GPU. As shown in Table 5, the model contained approximately 6.22 million parameters, with a model size of 23.75 MB. The input image size was 224 × 224 pixels. The average forward inference time was 10.63 ms per image, corresponding to an inference speed of 94.10 FPS. This result indicates that the neural network inference stage itself is computationally efficient and achieved high forward-inference throughput under the tested hardware configuration. When the complete image-processing pipeline was considered, including image loading, preprocessing, model inference, and post-processing, the average processing time was 392.30 ms per image, corresponding to 2.55 FPS.

Table 5. Computational efficiency of the proposed ViT-UNet model.

Metric	Value
Parameters	6.22 M
Model size	23.75 MB
Input size	224 × 224 pixels
Forward inference time	10.63 ms/image
Forward inference speed	94.10 FPS
Full pipeline time	392.30 ms/image
Full pipeline speed	2.55 FPS
Hardware	NVIDIA GeForce RTX 4060 Ti

The large difference between forward inference speed and full-pipeline speed indicates that the current throughput bottleneck lies mainly in image I/O, preprocessing, and post-processing rather than in the neural network architecture. Therefore, although the proposed model has real-time inference potential at the network level, deployment of the method in high-throughput seed assessment or automated visual inspection applications would require further optimization of the complete processing pipeline.

Taken together, the quantitative comparison, bootstrap confidence intervals, paired statistical tests, prediction-agreement analysis, error distribution analysis, descriptive stratified analysis, and computational-efficiency evaluation indicate that the proposed ViT-UNet achieved the lowest overall counting errors among the evaluated methods while maintaining high forward-inference throughput at the model level. Representative image-level prediction cases are provided in Supplementary Figure S2.

4. Discussion

4.1. Effectiveness of ViT-UNet Density Regression for Multiclass Seed Counting

The present study demonstrates the feasibility of using a ViT-UNet density regression framework for multiclass horticultural seed counting. Compared with direct image-level count regression, density regression provides a spatially structured learning target by converting point annotations into category-specific density maps. This design is particularly suitable for seed-counting tasks because horticultural seeds are usually small, numerous, and may appear in dense or partially touching arrangements. In such cases, precise instance boundary separation is difficult, whereas density regression can estimate seed numbers through continuous spatial responses without requiring bounding-box or pixel-level mask annotations.

The proposed framework achieved useful total-count prediction on the held-out test set. The total-count results suggest that the model learned the overall relationship between image appearance and seed number. This is important for post-harvest seed assessment, where total seed number is often the primary output for packaging, sampling, sowing-rate control, and batch-level quality inspection. Compared with category-wise prediction, total-count estimation showed a lower percentage error, suggesting that some category-level deviations could partly compensate for each other after summing the three output channels.

The category-wise results further indicate that the model captured the main count variations in cucumber, tomato, and pepper seeds. However, the difficulty was not uniform across categories. Cucumber seeds were relatively large and elongated, making them more visually distinguishable. In contrast, tomato and pepper seeds were smaller and more similar in color, shape, and local texture. This visual similarity increased the possibility of inter-class confusion, especially in mixed-category images. Therefore, the results indicate that the proposed framework is effective for total seed counting, while category-level discrimination remains a more challenging component of the task.

4.2. Architectural Contribution of the ViT-UNet Model

The comparison with U-Net, Pure Transformer, and ViT-UNet without skip connections provides insight into the architectural contribution of the proposed model. The full ViT-UNet achieved the best overall performance, indicating that the combination of Transformer-based global feature extraction and U-Net-style spatial decoding is beneficial for horticultural seed counting.

The weak performance of the evaluated Pure Transformer baseline suggests that global self-attention without an effective spatial reconstruction decoder was insufficient under the present dataset and training conditions. Although Transformer architectures are effective at modeling long-range dependencies, patch-based tokenization may reduce fine spatial details that are important for local density reconstruction. For seed counting, especially when seeds are small and closely distributed, the model must preserve both local appearance and spatial arrangement. Without an effective decoder for spatial recovery, a pure Transformer model may struggle to reconstruct high-resolution density responses, which may partly explain its large error and negative R^2 .

The comparison between ViT-UNet without skip connections and the full ViT-UNet further supports the contribution of multilevel feature fusion. Skip connections allow shallow and intermediate encoder features to be introduced into the decoder, preserving spatial details that may be weakened in deeper feature representations. The performance degradation observed after removing skip connections suggests that shallow and intermediate spatial features contributed substantially to seed density prediction under the present evaluation conditions.

Therefore, the proposed ViT-UNet architecture is effective because it addresses two complementary requirements of seed counting: global context modeling and local spatial reconstruction. Related hybrid architectures have also combined local spatial representation with global contextual modeling in agricultural image analysis and dense-prediction tasks [24,28,31]. The ViT encoder contributes global representation capability, while the U-Net-style decoder and skip connections help recover spatially detailed density information. Under the present dataset split and experimental conditions, this combination performed better than the evaluated convolutional U-Net and Pure Transformer baselines for small-object density regression.

4.3. Error Sources and Category-Wise Prediction Behavior

The prediction-agreement and error-distribution analyses reveal several important characteristics of model behavior. First, the predicted and ground-truth counts showed clear positive relationships, indicating that the model captured the overall counting trend. However, deviations from the 1:1 reference line were still observed, especially in category-wise predictions. These deviations suggest that the remaining errors were not random numerical fluctuations alone but were related to the visual and structural difficulty of specific samples.

One important error source was inter-class false-positive prediction. Tomato and pepper showed nonzero predictions in some samples where the corresponding ground-truth count was zero. This indicates that the model sometimes confused visually similar small seeds. Such category-level ambiguity is reasonable because tomato and pepper seeds share similar visual features, particularly in color and local texture. In mixed-category or dense scenes, the model may detect seed-like regions correctly but assign part of the response to an incorrect category channel.

Another error source was underestimation in high-count samples. The scatter plots and residual analysis indicate that some samples with larger ground-truth counts tended to be underestimated. This may be associated with dense arrangements, partial adhesion, and local overlap among seeds. When multiple seeds are close to one another, their local visual features may merge, and the density response may become less separable. As a result, the integrated predicted count may be lower than the actual count.

4.4. Practical Implications for Post-Harvest Seed Assessment

The proposed method has practical relevance for horticultural seed assessment and automated visual inspection. Traditional manual counting is labor-intensive and inefficient, especially when large numbers of small seeds need to be processed. Conventional image-processing methods, such as threshold segmentation and connected-component analysis, may be sensitive to background color, illumination variation, adhesion, and seed similarity. The proposed deep density regression framework offers a more flexible alternative by learning the mapping from seed images to category-wise and total-count estimates.

The total-count performance is especially meaningful for practical applications. In many post-harvest scenarios, the total number of seeds in a sample or batch is the most important measurement. The proposed model achieved useful total-count prediction on the held-out test set, suggesting that it may provide a basis for a future visual counting module for seed and batch-level quality assessment. The category-wise outputs provide additional information when multiple seed types need to be counted separately, although further improvement is still needed for visually similar categories.

The computational results also show practical potential. The model achieved a forward inference speed of 94.10 FPS on an NVIDIA GeForce RTX 4060 Ti GPU, indicating that the neural network itself is sufficiently lightweight for real-time inference. However, the full processing pipeline reached only 2.55 FPS, showing that the current system-level

bottleneck lies mainly in image loading, preprocessing, and post-processing rather than model inference. This distinction is important for deployment: the model has real-time potential, but the complete system must be optimized before use in high-throughput seed sorting or industrial inspection lines.

For practical deployment, several engineering improvements are needed. Batch inference, parallel preprocessing, hardware-triggered image acquisition, optimized data loading, and edge-device implementation could substantially increase the full-pipeline throughput. In addition, stable illumination, controlled imaging distance, and standardized background conditions would further improve prediction reliability in practical environments.

4.5. Limitations and Future Work

Several limitations should be acknowledged. First, the dataset size was limited. Although the dataset contained 231 images and 7789 annotated seed instances, the held-out test set contained only 36 images and may not fully represent all seed varieties, imaging conditions, nominal seed-density acquisition conditions, or practical deployment scenarios. In addition, because the same physical seed lots were repeatedly rearranged, images in different subsets may share seed-specific appearance characteristics. The held-out test set therefore represents image-level validation rather than independent seed-batch validation. Moreover, one fixed dataset split and one selected checkpoint for each model were used, and variability across repeated training runs or alternative dataset splits was not quantified. Future work should include additional seed species, independently collected seed batches, more imaging devices, repeated training runs, and broader environmental variations.

Second, although the dataset included images acquired under different background, illumination, and nominal seed-density conditions, subgroup sample sizes were limited. The analysis therefore reports descriptive performance stratified by background color, illumination condition, and nominal seed-density acquisition condition in Supplementary Table S2. These results provide useful information about performance consistency across imaging conditions, but they should not be over-interpreted as strong statistical evidence of robustness because several subgroups contained only a small number of test images.

Third, the comparison includes two conventional image-processing baselines, two deep-learning baselines, and one ablation model. The conventional thresholding and watershed baselines provide a practical non-learning reference and show the difficulty of applying fixed segmentation rules to seeds under variable background, illumination, and nominal seed-density acquisition conditions. Nevertheless, these baselines were implemented as fixed reference methods rather than extensively tuned classical pipelines, and future work could evaluate additional color-space, shape-based, or hybrid segmentation approaches.

Fourth, category-wise prediction remains more challenging than total-count estimation. Supplementary category-level presence/absence and zero-ground-truth false-positive analyses indicate that nonzero residual category-channel responses can occur for absent categories, particularly among visually similar seeds. These auxiliary results should be interpreted cautiously because the model is optimized for density regression rather than instance-level classification. Future work may improve category-level discrimination by using larger mixed-category datasets, class-balanced sampling, higher-resolution local features, attention mechanisms for category-specific channels, or auxiliary classification supervision.

Finally, the current study evaluated the model under an offline image-processing setting. Although the forward inference speed was high, deployment in real post-harvest sorting systems requires validation under continuous image acquisition, moving samples, varying sample placement, and hardware constraints. Future work should therefore integrate the model with practical imaging platforms and evaluate its performance in real-time sorting or inspection workflows.

Overall, the results demonstrate the feasibility of the proposed ViT-UNet density regression framework for multiclass horticultural seed counting under the tested acquisition conditions. However, category ambiguity, underestimation in some high-count or crowded samples, limited subgroup sizes, and the absence of independent seed-batch validation constrain the generalizability of the present findings. With larger datasets, independently collected seed batches, more comprehensive robustness validation, and optimized deployment pipelines, the method may support practical post-harvest seed quality assessment and high-throughput visual inspection.

5. Conclusions

This study developed a ViT-UNet-based density regression framework for vision-based counting of horticultural seeds. Cucumber, tomato, and pepper seeds were selected as representative vegetable seed categories with different morphological and visual characteristics. A point-annotation-based dataset was constructed, and category-specific density maps were generated to support multiclass seed counting without requiring bounding-box or pixel-level mask annotations. The proposed model integrated a ViT-Tiny encoder, a U-Net-style decoder, and multilevel skip connections to combine global contextual representation with spatially resolved density-map reconstruction.

On the held-out test set, the proposed ViT-UNet achieved a total-count MAE of 4.0678, RMSE of 5.1896, MAPE of 14.7104%, and R^2 of 0.8272. At the category level, the model obtained R^2 values of 0.915, 0.865, and 0.913 for cucumber, tomato, and pepper seeds, respectively. These results indicate that the proposed framework captured the major count variations in the three seed categories and provided useful total-count estimates on the held-out test set. The prediction-agreement and error-distribution analyses further showed that the model achieved meaningful overall counting performance, although category-wise false-positive responses and high-count underestimation were still observed in some difficult samples.

The statistical analysis showed that the total-count MAE had a bootstrap 95% CI of 3.05–5.18 seeds, and paired Wilcoxon tests with Holm–Bonferroni correction provided statistical support for lower image-level total-count absolute errors of the proposed model relative to the evaluated deep-learning baselines and the ablation model. The two fixed conventional image-processing reference pipelines also produced higher total-count errors than the proposed model under the tested acquisition conditions. However, these results should not be interpreted as a general comparison with all optimized conventional image-processing methods.

Compared with U-Net, Pure Transformer, and ViT-UNet without skip connections, the proposed ViT-UNet achieved the lowest total MAE and MAPE and the highest total R^2 . The comparison indicated that the combination of Transformer-based global feature extraction, U-Net-style spatial decoding, and skip-connection-based feature fusion was beneficial under the present evaluation conditions. In particular, the weak performance of the Pure Transformer model and the performance degradation after removing skip connections suggested that spatial reconstruction and multilevel feature fusion contributed to accurate small-object counting.

The proposed model also showed promising computational efficiency. With approximately 6.22 million parameters and a model size of 23.75 MB, the network achieved a forward inference speed of 94.10 FPS on an NVIDIA GeForce RTX 4060 Ti GPU. This indicates that the model itself has real-time inference potential. However, the full processing pipeline reached 2.55 FPS, suggesting that practical deployment still depends on further optimization of image loading, preprocessing, and post-processing procedures.

Overall, this study demonstrates the feasibility of using ViT-UNet density regression for multiclass horticultural seed counting under the tested conditions. Future work will focus on expanding the dataset using independently collected seed batches, improving category-level discrimination for visually similar seeds, evaluating variability across repeated training runs, validating the method under practical imaging conditions, and optimizing the complete processing pipeline for high-throughput deployment.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/horticulturae12080961/s1>. Table S1, Bootstrap 95% confidence intervals for category-wise and total-count performance of the proposed ViT-UNet model; Table S2, Descriptive stratified total-count performance of the proposed ViT-UNet model under different background, illumination, and nominal seed-density acquisition conditions; Table S3, Auxiliary category-level presence/absence, threshold-sensitivity, and zero-ground-truth false-positive analyses; Figure S1, Training and validation loss curves of the proposed ViT-UNet model; Figure S2, Representative conventional image-processing results and image-level prediction cases. The representative prediction cases are presented at the image-count level, including category-wise groundtruth and predicted counts.

Author Contributions: Conceptualization, M.X.; methodology, Z.T.; software, M.D.; validation, C.Z.; formal analysis, M.D.; investigation, Z.T.; resources, J.M.; data curation, M.X.; writing—original draft preparation, M.D.; writing—review and editing, M.X.; visualization, M.D.; supervision, J.M.; project administration, Y.Z.; funding acquisition, M.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Fundamental Research Funds for the Provincial Universities of Zhejiang, grant number 2025YW106.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors acknowledge the financial support from the China Scholarship Council (CSC) for this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Musaev, F.; Priyatin, N.; Potrakhov, N.; Beletskiy, S.; Chesnokov, Y. Assessment of Brassicaceae Seeds Quality by X-ray Analysis. *Horticulturae* **2021**, *8*, 29. [\[CrossRef\]](#)
2. Altizani-Júnior, J.C.; Cicero, S.M.; de Lima, C.B.; Alves, R.M.; Gomes-Junior, F.G. Optimizing Basil Seed Vigor Evaluations: An Automatic Approach Using Computer Vision-Based Technique. *Horticulturae* **2024**, *10*, 1220. [\[CrossRef\]](#)
3. Margapuri, V.; Thapaliya, P.; Neilsen, M. Seed Kernel Counting using Deep Learning for High-Throughput Phenotyping. In *2024 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, Bali, Indonesia, 28–30 November 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–7. [\[CrossRef\]](#)
4. Jin, C.; Zhou, L.; Pu, Y.; Zhang, C.; Qi, H.; Zhao, Y. Application of Deep Learning for High-Throughput Phenotyping of Seed: A Review. *Artif. Intell. Rev.* **2025**, *58*, 76. [\[CrossRef\]](#)
5. Tian, H.; Wang, T.; Liu, Y.; Qiao, X.; Li, Y. Computer vision technology in agricultural automation—A review. *Inf. Process. Agric.* **2020**, *7*, 1–19. [\[CrossRef\]](#)
6. Meenu, M.; Kurade, C.; Neelapu, B.C.; Kalra, S.; Ramaswamy, H.S.; Yu, Y. A concise review on food quality assessment using digital image processing. *Trends Food Sci. Technol.* **2021**, *118*, 106–124. [\[CrossRef\]](#)
7. Zhu, L.; Spachos, P.; Pensini, E.; Plataniotis, K. Deep learning and machine vision for food processing: A survey. *Curr. Res. Food Sci.* **2021**, *4*, 233–249. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Sheng, H.; Zhang, Y.; Cong, R.; Wang, S.; Yang, D.; Cui, Z.; Huang, X.; Chen, R.; Liu, J.; Ke, W.; et al. Deep-learning based colorectal cancer pathological analysis with hyperspectral light field microscopy. *iScience* **2025**, *28*, 112987. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016*; IEEE: Piscataway, NJ, USA, 2016; pp. 779–788.

10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
11. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A Review of Yolo Algorithm Developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [[CrossRef](#)]
12. Deng, R.; Tao, M.; Huang, X.; Bangura, K.; Jiang, Q.; Jiang, Y.; Qi, L. Automated Counting Grains on the Rice Panicle Based on Deep Learning Method. *Sensors* **2021**, *21*, 281. [[CrossRef](#)] [[PubMed](#)]
13. Tian, Y.; Wang, S.; Li, E.; Yang, G.; Liang, Z.; Tan, M. MD-YOLO: Multi-scale Dense YOLO for small target pest detection. *Comput. Electron. Agric.* **2023**, *213*, 108233. [[CrossRef](#)]
14. Wang, X.; Li, C.; Zhao, C.; Jiao, Y.; Xiang, H.; Wu, X.; Chai, H. GrainNet: Efficient detection and counting of wheat grains based on an improved YOLOv7 modeling. *Plant Methods* **2025**, *21*, 44. [[CrossRef](#)] [[PubMed](#)]
15. Gao, Q.; Ji, Y.; Shi, C.; Wang, M. Design of a Machine Vision Detection System for Lettuce Growth Stages Based on the CCASF-YOLOv10 Model. *Horticulturae* **2026**, *12*, 379. [[CrossRef](#)]
16. Kumari, K.; Won, K.; Nafchi, A.M. Computer Vision for Cover Crop Seed-Mix Detection and Quantification. *Seeds* **2025**, *4*, 59. [[CrossRef](#)]
17. Ahmed, S.; Revolinski, S.R.; Maughan, P.W.; Savic, M.; Kalin, J.; Burke, I.C. Deep learning-based detection and quantification of weed seed mixtures. *Weed Sci.* **2024**, *72*, 655–663. [[CrossRef](#)]
18. Mehreen, S.; Goëau, H.; Bonnet, P.; Chau, S.; Champ, J.; Joly, A. Estimating compositions and nutritional values of seed mixes based on Vision Transformers. *Plant Phenomics* **2023**, *5*, 0112. [[CrossRef](#)] [[PubMed](#)]
19. Amirholipour, S.; He, X.; Jia, W.; Wang, D.; Zeibots, M. A-CCNN: Adaptive CCNN for density estimation and crowd counting. In *Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018*; IEEE: Piscataway, NJ, USA, 2018; pp. 948–952.
20. Xiong, H.; Cao, Z.; Lu, H.; Madec, S.; Shen, C. TasselNetv2: In-field counting of wheat spikes with context-augmented local regression networks. *Plant Methods* **2019**, *15*, 150. [[CrossRef](#)] [[PubMed](#)]
21. Ranjan, V.; Sharma, U.; Nguyen, T.; Hoai, M. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 3394–3403.
22. Zhang, Y.; Zhao, H.; Duan, Z.; Huang, L.; Deng, J.; Zhang, Q. Congested crowd counting via adaptive multi-scale context learning. *Sensors* **2021**, *21*, 3777. [[CrossRef](#)] [[PubMed](#)]
23. Gong, S.; Zhang, S.; Yang, J.; Dai, D.; Schiele, B. Class-agnostic object counting robust to intraclass diversity. In *Computer Vision—ECCV 2022*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 388–403.
24. Zaji, A.; Liu, Z.; Xiao, G.; Bhowmik, P.; Sangha, J.S.; Ruan, Y. Wheat spike localization and counting via hybrid UNet architectures. *Comput. Electron. Agric.* **2022**, *203*, 107439. [[CrossRef](#)]
25. Sung, B.-G.; Lee, C.-G.; Kang, Y.-H.; Yu, S.-H.; Lee, D.-H. Semi-Supervised Density Estimation with Background-Augmented Data for In Situ Seed Counting. *Agriculture* **2025**, *15*, 1682. [[CrossRef](#)]
26. Zheng, H.; Fan, X.; Bo, W.; Yang, X.; Tjahjadi, T.; Jin, S. A Multiscale Point-Supervised Network for Counting Maize Tassels in the Wild. *Plant Phenomics* **2023**, *5*, 0100. [[CrossRef](#)] [[PubMed](#)]
27. Tabernik, D.; Muhovič, J.; Skočaj, D. Dense center-direction regression for object counting and localization with point supervision. *Pattern Recognit.* **2024**, *153*, 110540. [[CrossRef](#)]
28. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM Comput. Surv.* **2022**, *54*, 1–41. [[CrossRef](#)]
29. Oyedotun, O.K.; Al Ismaeil, K.; Aouada, D. Why is everyone training very deep neural network with skip connections? *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 5961–5975. [[CrossRef](#)] [[PubMed](#)]
30. Zhu, M.; Xu, Z.; Zhang, Q.; Liu, Y.; Gu, D.; Xu, S.S.-D. GCSTormer: Gated Swin Transformer with channel weights for image denoising. *Expert Syst. Appl.* **2025**, *284*, 127924. [[CrossRef](#)]
31. Xie, J.; Wu, H.; Wu, W.; Hong, L.; He, L.; Yu, Q.; Liu, L.; Lin, A.; Som-ard, J. A CNN–Transformer hybrid network with boundary guidance for mapping cropland field parcels from high-resolution remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2026**, *64*, 4401122. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.