AgriEngineering

Article

Heatmap Regression-Based Context-Aware Learning for Tillage Boundary Detection

Gyu-Sung Ham [1] and Kanghan Oh [1,2,*]

[1] AI Convergence Research Institute, Wonkwang University, Iksan 54538, Republic of Korea; ham1231@wku.ac.kr
[2] Department of Computer and Software Engineering, Wonkwang University, Iksan 54538, Republic of Korea
* Correspondence: khoh888@wku.ac.kr

Abstract: In agricultural automation, autonomous tractors play a crucial role in enhancing farming efficiency, particularly through the navigation and identification of tillage boundaries. Traditional approaches rely on machine vision techniques to identify paths by distinguishing between tilled and untilled soil areas. Although recent advancements in convolutional neural networks (CNNs) have shown promising results in agricultural automation, they still face challenges in fully capturing the global context of tillage boundaries. This limitation stems mainly from CNNs' small receptive fields, which often limit the network's capacity to capture broader contextual information in agricultural landscapes, potentially causing inaccuracies in boundary detection. These methods rely significantly on local feature analysis and necessitate complex, computationally intensive heuristic post-processing to enhance detected tillage lines, thus limiting their real-time application efficacy. We propose a line-context-aware learning method that combines a heatmap regression with a transformer to more effectively learn and extract global contextual features. The proposed end-to-end method streamlines detection, enhancing real-time agricultural applications and improving the accuracy and reliability of autonomous tractor navigation and operation. The proposed method was evaluated on a custom dataset, demonstrating competitive performance in accurately detecting tillage boundaries and proving its capability to handle the intricate details and variations present in agricultural landscapes.

Keywords: autonomous tractor; tillage boundary detection; CNN; deep learning; precision agriculture

Academic Editor: Simone Pascuzzi

Received: 9 December 2024
Revised: 14 January 2025
Accepted: 28 January 2025
Published: 30 January 2025

Citation: Ham, G.-S.; Oh, K. Heatmap Regression-Based Context-Aware Learning for Tillage Boundary Detection. *AgriEngineering* **2025**, *7*, 32. https://doi.org/10.3390/agriengineering7020032

1. Introduction

Precision agriculture is a key paradigm in industrial farming that contributes greatly to enhanced productivity. This field has increasingly focused on automation technologies that minimize labor requirements by using sophisticated computer vision techniques for various tasks with significant efficiency [1–10]. Autonomous tractors, which are pivotal in modern cultivation practices, serve as both vehicles and instrumental tools. The advancement of self-driving tractors is a major aim of agricultural automation [1,10]. In particular, the capability to automate path-finding is crucial because it significantly reduces manual intervention and ensures precise and efficient operations across various farming tasks. For self-driving tractors, detecting the tillage line during operation is essential for optimizing working paths, reducing redundant operations, and preventing missed areas. This

is particularly beneficial for family farms, as it significantly reduces manual labor and increases efficiency in various agricultural operations, including plowing, planting, and harvesting. Autonomous tractors, compared to traditional tractors, offer notable advantages in terms of production savings. By minimizing the need for human intervention, autonomous tractors reduce labor costs, enhance operational precision, and allow for continuous operation, even under challenging conditions, such as low visibility or extended hours. This leads to improved productivity and cost-effectiveness for family farms, ensuring sustainable agricultural practices.

Traditional autonomous tractor technologies often use global positioning systems (GPS) [11–13] that rely on absolute coordinates for navigation, thereby achieving significant success. In addition, light detection and ranging (LiDAR) technology has been employed to capture detailed spatial data and enhance the ability of tractors to navigate by identifying critical structural features, e.g., crops and trees [14,15]. Although these methods provide robust solutions for navigation and environmental perception, they are often limited by high costs and signal variability in diverse farming environments.

In recent years, advancements in machine vision have become increasingly prominent in agriculture, providing innovative solutions for various farming challenges. Machine vision has garnered significant interest in the realm of visual perception not only for its ability to mimic human-like visual recognition but also for its cost-effectiveness. In particular, convolutional neural networks (CNNs) have achieved outstanding performance in the field of computer vision, excelling in various tasks. Their sophisticated architecture allows them to achieve exceptional results, often surpassing human-level performance in image classification [16]. Moreover, CNNs have been proven to be effective in other vision tasks, e.g., object detection [17] and segmentation [18,19], where their ability to interpret complex visual information is crucial. This versatility makes CNNs invaluable for applications requiring detailed visual understanding and has led to their widespread adoption in both academic and industrial fields.

CNNs have significantly advanced the field of autonomous agricultural machinery by enhancing the ability to navigate and perform tasks with greater precision. For example, CNNs have been applied to autonomous tractors for accurate field mapping [20], weed detection [21], and crop monitoring [22,23]. By recognizing different terrain types and plant health indicators, CNNs help optimize the path and operation of these tractors, ensuring efficient coverage and minimal damage to crops [24]. Recently developed CNN-based techniques have introduced innovative methods for autonomously controlling agricultural tractors by using only RGB images [1,2,5,6]. Existing methods [5,6] have proposed image-patch-wise classification to classify areas into tilled and non-tilled regions. Although effective, this process is computationally demanding because it requires independent feedforward operations for each image patch during testing, posing challenges for real-time processing, which is essential in autonomous tractor operations.

In contrast, the fully convolutional network (FCN)-based segmentation method [20] offers a more streamlined approach by directly segmenting the tillage boundary using a single feedforward pass. This reduces the computational overhead and aligns better with the requirements of real-time agricultural operations, ensuring that machinery can operate efficiently without the delay involved in processing multiple frames. Recently, a heatmap regression-based ensemble network [25] was introduced as a heatmap regression model for boundary detection, and demonstrated significant success. This study employed image-to-image mapping to produce a likelihood heatmap in which the tillage boundary points were recognized as the global maximum in the heatmap.

Although CNNs and FCNs have proven to be highly effective in numerous applications, they have inherent limitations that can affect their performance, especially in complex environments, e.g., agriculture. The convolutional kernels are typically small, which

inherently limits their ability to capture and learn large amounts of contextual information from input images. This can be particularly problematic in tasks, e.g., tillage boundary detection, where understanding the broader context and spatial relationships is crucial for obtaining accurate results. Because CNNs and FCNs have a limited ability to grasp larger contextual information, local noise or incomplete predictions often arise, requiring extensive heuristic post-processing to refine their outputs. Such post-processing is computationally expensive and time-consuming, making it less ideal for real-time applications where speed and efficiency are critical. Furthermore, diverse and natural field environments pose significant challenges for CNNs, especially because of the extreme variability in the local patterns of soil texture, which are heavily influenced by factors like weather [5,6]. To construct a robust model that operates under varied conditions, it is crucial to understand the global context of the farming environment, which helps mitigate the effects of local noise.

To address these issues, we propose a heatmap regression-based context-aware network to learn the contextual information within an agricultural environment. To this end, we defined three tillage boundary landmark points, which were subsequently used to calculate the tillage boundaries and determine the driving direction of the tractor. Figure 1 illustrates these three landmark points. The line connecting the red and green points indicates the direction of travel, whereas the green and blue points represent the positions of the tractor's U-turn. Our model incorporates advanced techniques, e.g., heatmap regression and transformers, to enhance its ability to accurately detect and refine tillage boundary points. Transformers, introduced by Vaswani et al. (2017) [26], have revolutionized natural language processing, computer vision, and other fields, owing to their powerful self-attention mechanisms. This method processes all the input features simultaneously, allowing them to capture long-range dependencies more effectively. The self-attention mechanism enables each position in the input sequence to attend to all other positions, thereby improving the ability to model contextual relationships.

Figure 1. Visualization of three landmark points for the tillage boundaries: The line connecting the red and green points indicates the direction of travel, while the green and blue points mark the positions for the tractor's U-turn.

In this study, we utilized a U-Net-based heatmap regression network to identify and learn the key landmarks of the tillage boundaries. This step ensured that the model could pinpoint critical areas with high precision. Subsequently, transformers were employed to further process these landmarks by learning the surrounding context, thereby enhancing the accuracy of the detected points. The focus of this study was solely on the detection of tillage boundaries. Therefore, controlling tractor operations is beyond the scope of this study.

The contributions of this study are summarized as follows. First, the proposed framework effectively detects tillage boundary lines by leveraging contextual line features, providing robust end-to-end learning (i.e., resistant to local noise), and eliminates the

need for any post-processing phases, unlike traditional methods. Second, the proposed method demonstrated strong performance across various environments and limited datasets.

The remainder of this paper is organized as follows: Section 2 provides a detailed description of the materials and the proposed method, Section 3 presents the experimental results and discussion, and finally, the conclusions are presented in Section 4.

2. Materials and Methods

2.1. Dataset Description

To acquire the dataset, we chose the LS-Mtron XU6168 (Figure 2, left) as the model for our tractor and equipped it with a GoPro Hero7 action camera strategically positioned at the center of the front-top edge of the tractor for an optimal field of view (FOV) (Figure 2, middle). The camera, capable of capturing 2-dimensional (2D) RGB images at a rate of 60 fps, ensured that each image maintained a resolution of 1920 × 1080 pixels and provided detailed visual data for analysis. The camera was mounted on the top of the windshield in the tractor's cabin, approximately 220 cm above-ground. When setting the camera's field of view, we ensured that the tractor's front did not obstruct the ground, and the view included not only the farmland but also the sky. During tillage, tractors typically operate on soil that has not been tilled or is adjacent to areas that have already been cultivated. Therefore, the boundary line between tilled and non-tilled soils serves as a crucial feature for guiding tractor movements. This boundary line allows for precise adjustment of the tractor's alignment, ensuring that it runs parallel to the edge of the previously tilled area, thereby optimizing the tillage process [1]. To collect data during tillage, the tractor was operated in a zigzag pattern across the field to ensure comprehensive coverage.

As Figure 2 (right) shows, the tractor starts from a preexisting tillage boundary and travels until it reaches the end of the field, where it performs a U-turn. After completing the turn, it was aligned with the previous tillage boundary for another pass, systematically collecting data along the path. This pattern was repeated throughout the field. When no previous tillage boundaries were present at the beginning of the tillage process, a preliminary pass was performed by using a tractor to establish the initial boundary. This initial line served as the reference for subsequent passes, allowing for structured and consistent data collection from the start of the operation.

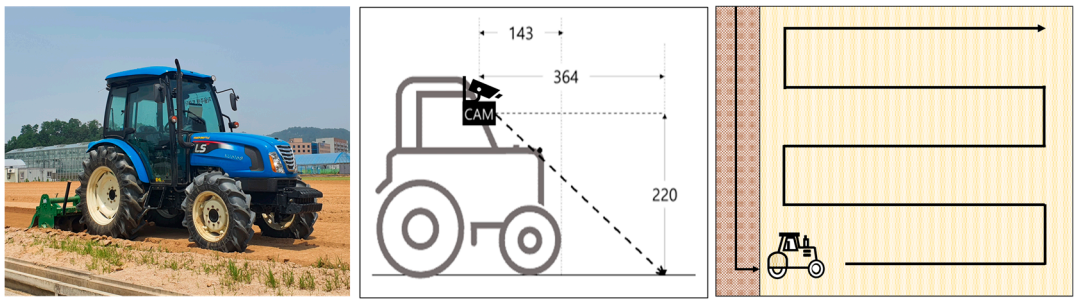

Figure 2. Data collection environment. From left to right: the tractor used in our experiment, the action camera position and its FOV (cm), and the tillage route.

To capture a comprehensive range of tillage scenarios, data collection was planned to include a variety of environmental conditions. We recorded images under different lighting conditions (morning and afternoon) and weather patterns (clear and overcast days) and across diverse agricultural fields to ensure variability in the dataset. This approach allowed us to simulate real-world conditions that a tractor might encounter in typical farming operations. Ultimately, our dataset was organized into three categories: 2590 images for training, 1800 for validation, and 1756 for testing. The testing dataset was

divided into two main categories: 863 images on sunny days and 860 images on cloudy days. The sunny day dataset was further split into two subcategories: 463 images from the afternoon and 407 images from the morning.

2.2. Methodology

An overview of the proposed method for detecting tillage boundary points is shown in Figure 3. Our model incorporates advanced techniques, e.g., heatmap regression, ensemble methods, and transformers, to enhance its ability to accurately detect and refine tillage boundary points. In the final stage, we integrated the results of the heatmap and coordinate regression.

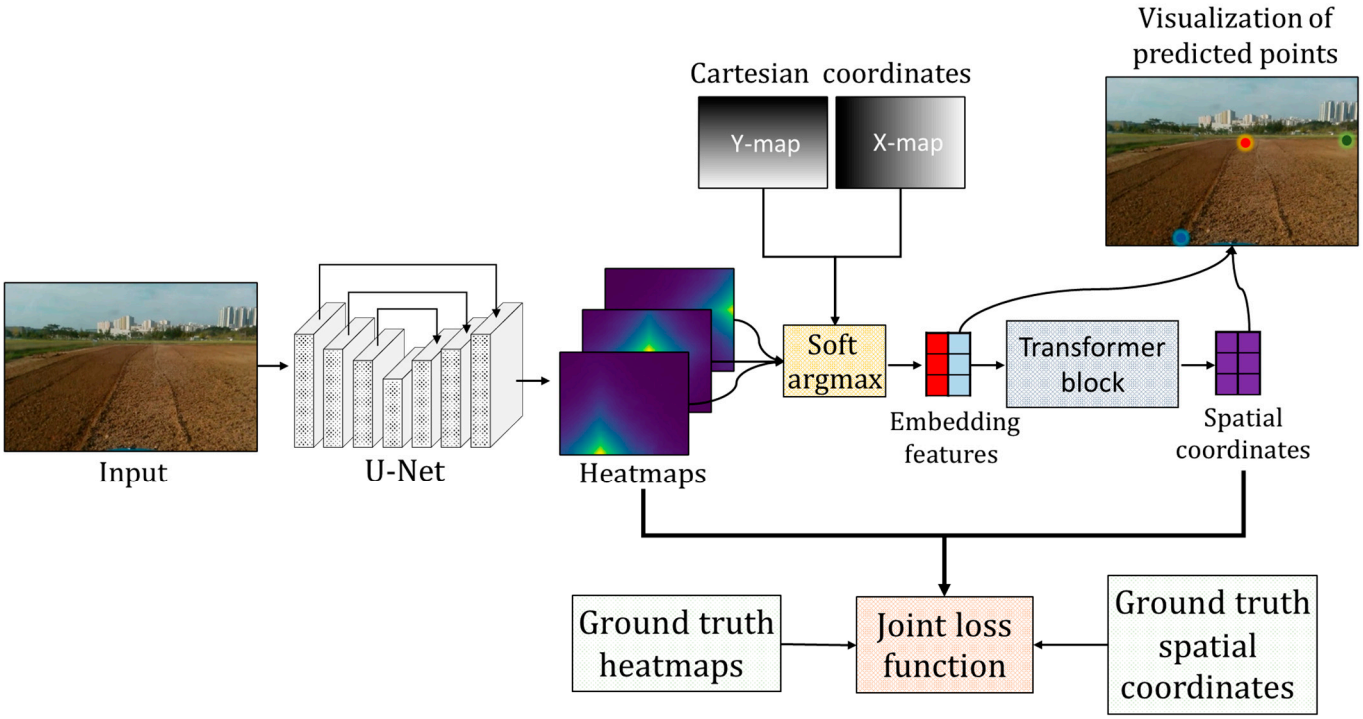

Figure 3. A heatmap regression-based context-aware network for boundary landmark point detection. The U-Net generates heatmap outputs, and spatial coordinates are extracted via soft-argmax. The transformer block captures contextual relationships between the tillage boundary points. The network is trained using a joint loss function combining heatmap regression and coordinate regression.

2.2.1. U-Net Based Heatmap Regression

Our methodology employed the U-Net architecture [18] to detect tillage boundary points based on heatmap regression, specifically utilizing three key points within the tillage field. This approach allows for the precise detection of tillage boundaries by creating detailed heatmaps around the tillage boundary points. These heatmaps serve as a normalized Gaussian distribution with a variance parameter of five, and in the final heatmap output, the highest intensity value within each channel indicates the location of these landmarks. By utilizing the maximum value within the heatmap to determine the boundary points, our model ensured an accurate representation of the tillage boundaries. The U-Net architecture (Figure 4) is structured around a series of repeating modules.

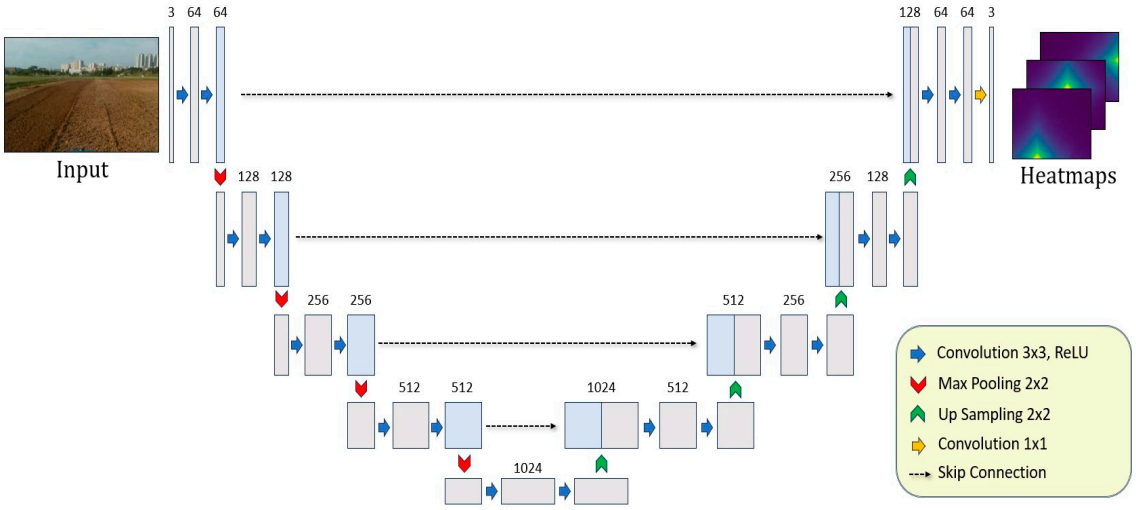

Figure 4. U-Net based heatmap regression for the boundary landmark point detection.

We adopted U-Net architecture [18] as the backbone of our network, leveraging its well-established design for effective feature extraction. The feature map sizes, derived from the U-Net structure, balance performance and computational efficiency. Increasing channel dimensions enhance feature representation at higher levels, which is critical for accurate boundary detection, while decreasing spatial dimensions minimize computational cost. Each module consisted of two 3 × 3 convolutional layers with padding to preserve the spatial dimensions. Following these convolutions, each layer was subjected to batch normalization and rectified linear unit activations. Each convolutional step is then succeeded by a layer-specific operation: max-pooling with a 2 × 2 window is used in the encoding layers to reduce the spatial dimensions while enhancing the feature depth and upsampling in the decoding layers to reconstruct the spatial dimensions. During the encoding phase, the network progressively increased the number of feature maps from 64 to 512, intensifying its ability to capture and process more complex features from the input images. Conversely, in the decoding phase, the feature maps were gradually reduced from 512 to 64, focusing on the precise localization and reconstruction of the high-resolution output. Importantly, skip connections were utilized to merge the feature maps from the corresponding encoding and decoding layers, thereby preserving spatial information that might otherwise be lost during downsampling.

$$O_h = Conv_1(Decoder(Encorder(x)_{skip}) \quad (1)$$

where $Encorder(\cdot)_{skip}$ is the function representing the encoder path with skip connections, and $Decoder(\cdot)$ is the function representing the decoder path that processes both the encoder output and the skip connections. $Conv_1$ denotes the final 1×1 convolutional layer of the output.

2.2.2. Spatial Feature Extraction Using Soft-Argmax

Given the heatmaps from the U-Net, the tillage boundary points were calculated using the soft-argmax function [27]. The soft-argmax function is a differentiable approximation of the argmax function, allowing the extraction of spatial coordinates within heatmaps without sacrificing the network's ability to learn through gradient descent. This operation converts the heatmap values, which represent the likelihood of the presence of a landmark at each pixel, into a weighted average of all pixel coordinates weighted by their softmax-normalized intensities. In this section, we consider Cartesian coordinates as the spatial feature, and the spatial feature extraction using soft-argmax is shown in Figure 5.

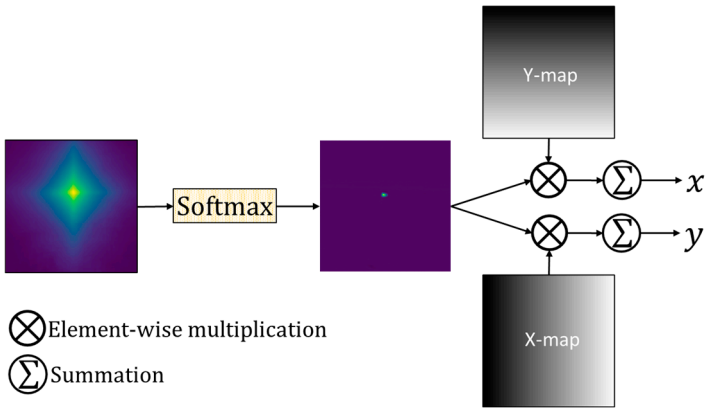

Figure 5. Soft-argmax function for the 2D heatmap with spatial coordinate maps (Cartesian). The softmax function normalizes the heatmap into a probability distribution, which is then element-wise multiplied by the X-map and Y-map to weight coordinates. The results are summed to pinpoint the landmark's precise location based on maximum probabilities.

Initially, the softmax function was applied to the heatmap to convert the intensity values into a probability distribution, which transformed the raw intensity values into a probability distribution, highlighting the region most likely to contain a boundary point while deactivating the less likely areas. This probability distribution was then multiplied element-wise using two coordinate maps: the X- and Y-maps. The final step involved summing these weighted coordinates across the entire map to calculate the x- and y-coordinates.

$$\text{SoftArgmax}\left(O_{h_i}^{(l)}\right) = \sum_i \frac{e^{O_{h_i}^{(l)}\alpha}}{\sum_j e^{O_{h_j}^{(l)}\alpha}} \cdot p_i \quad (2)$$

where $O_h^{(l)}$ represents the l-th input heatmap, $O_{h_i}^{(l)}$ denotes the individual elements of the heatmap, p_i is the corresponding index, and $e^{O_{h_i}^{(l)}}$ represents the exponential value $O_{h_i}^{(l)}$. α serves to control the sharpness of the softmax distribution; when α is increased, the softmax function becomes more selective, enhancing the impact of the highest values in the vector x. In this study, we empirically set α to 10. This process yielded the final coordinates, pinpointing the most salient positions of the tillage boundary with high precision, thereby translating the heatmap into precise spatial locations.

2.2.3. Context-Aware Learning Using a Transformer

In this section, we employ a transformer block to learn the contextual relationships between the tillage boundary points extracted from the soft-argmax function. Transformers [26] are a type of neural network architecture that has become fundamental in various fields of machine learning, including natural language processing and computer vision. They used an attention mechanism to weigh the influence of different parts of the input data, thereby enabling them to capture the relationships between distant elements within a feature space. Unlike CNNs, which focus primarily on local features, transformers are particularly useful for tasks requiring a comprehensive understanding of an entire scene. In our model, the tillage boundary points are considered as embedding vectors and used as inputs for the transformer to refine the detection results of the tillage boundaries. Initially, the embedding vector was normalized using layer normalization (LN) [28] to ensure consistency in the scale and variance.

$$\text{LN}(x) = \frac{x - \mu}{\sigma} \times \gamma + \beta \quad (3)$$

where μ and σ are the mean and standard deviation of the embeddings, and γ and β are learnable parameters of LN. Following normalization, these embeddings are subjected to the multi-head attention mechanism, which processes inputs as sets of queries (Q), keys (K), and values (V) across multiple attention heads.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where d_k is the dimension of the key for scale normalization. The expressions $Q = XW^Q, K = XW^K$, and $V = XW^V$ represent the results of the inner products with query, key, and value weights, respectively. We denote $X = \text{LN}\left(\text{SoftArgmax}\left(O_{h_i}^{(l)}\right)\right)_{l=1}^{n}$, where n is the number of heatmaps.

This operation allowed the model to focus on different parts of the input sequence to capture various aspects of the spatial context. The outputs from all the attention heads are concatenated and linearly transformed to integrate the information processed by each head.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{h}_1, \ldots, \text{h}_\text{n})W^O, \text{where } h_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{5}$$

where W^O is a learnable weight matrix that combines the concatenated outputs from all individual attention heads. Finally, the transformer refines the tillage boundary points through a series of linear transformations based on the learned contextual relationships.

$$O_c = \text{Linear}(\text{MultiHead}(Q, K, V)) \tag{6}$$

where O_c is 3 × 2, corresponding to the three tillage boundary points. In the experiments, we empirically set the number of multi-heads to 3 and the number of nodes in the final linear layer to 64. Figure 6 illustrates the architecture of a transformer block and its multi-head attention mechanism.

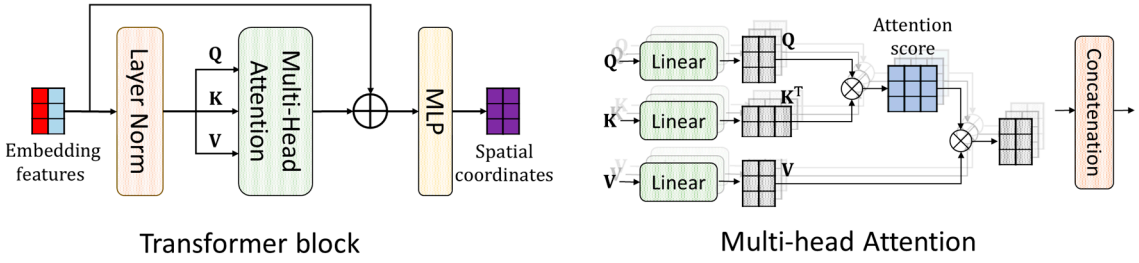

Figure 6. Illustration of a transformer block and its multi-head attention mechanism. The input embedding features undergo layer normalization, multi-head attention, and multilayer perceptron processing to yield spatial coordinates.

In the transformer block, the input embedding features are first normalized using LN. These normalized features are then fed into the multi-head attention module, where the attention mechanism is applied in parallel across the multiple attention heads. Each attention head computes the attention scores by taking the dot products between the query and key vectors, which are subsequently applied to the value vectors. The outputs from all the attention heads are concatenated and passed through a multilayer perceptron that produces the final spatial coordinates. This design allowed the model to focus simultaneously on different parts of the input sequence, thereby improving its ability to capture complex dependencies.

2.2.4. Implementation Detail

Evaluations were performed using a computer equipped with an Intel Core i7-7800X CPU running at 3.50 GHz, with 32 GB of RAM, and an Nvidia GeForce RTX 3090 GPU. We used Pytorch version 1.8.0, a widely recognized framework for deep learning, to train and test the network. Because the original images were captured at high resolution, they were resized to 320 × 240 pixels to streamline the processing and analysis phases without significant performance loss.

For training, the network underwent 500 epochs with the Adam optimizer, utilizing a mini-batch size of one. The initial learning rate was set to 1×10^4 and adjusted based on the cosine-annealing rate decay strategy [29] to optimize convergence. Our data augmentation process incorporates two types of transformations to enhance model robustness and performance: geometric and intensity modifications. Initially, the input images, along with their corresponding ground truth heatmaps, undergo random rotations between −15 and 15 degrees and are rescaled to vary between 80% and 120% of their original size. Subsequently, the image quality was adjusted by randomly altering the brightness, contrast, saturation, and hue to simulate different lighting conditions and camera settings.

For the loss function, we employed a joint loss function to effectively train both the heatmap regression and the coordinate regression derived from the transformer. The joint loss function is defined as:

$$\mathcal{L}_{joint} = \mathcal{L}2_{heat}(G_h, O_h) + \gamma \mathcal{L}2_{coord}(G_c, O_c) \quad (7)$$

where $\mathcal{L}2_{heat}$ is the loss calculated between the predicted heatmaps O_h and the ground truth heatmaps G_h, focusing on accurately capturing the presence and location of the tillage boundary points within the heatmap. Similarly, $\mathcal{L}2_{coord}$ is the loss function for coordinate regression. γ is a tunable hyperparameter that balances the importance of the heatmap regression loss against the coordinate regression loss. We set the weight hyperparameter γ to 1×10^{-3}. To obtain the final boundary points, we integrated the results of the heatmap and coordinate regressions. It is noteworthy that, prior to the ingestion process, the heatmaps were converted into coordinates using the soft-argmax method.

2.3. Evaluation Metrics

For the evaluation metrics, we employed the mean radial error (MRE) as a key metric for quantitatively measuring the accuracy of our predictions. MRE is particularly suited to tasks involving spatial coordinates because it directly quantifies the average distance between the predicted and true positions of boundary points. MREs were calculated using the following formula:

$$MRE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left(x_{i,pred} - x_{i,true}\right)^2 - \left(y_{i,pred} - y_{i,true}\right)^2} \quad (8)$$

where N denotes the total number of samples. $x_{i,pred}$ and $y_{i,pred}$ are the predicted coordinates, and $x_{i,true}$ and $y_{i,true}$ are the ground truth coordinates. This metric provides a straightforward and intuitive measure of the localization performance by averaging the pixel-wise Euclidean distances (radial errors) between the predicted and actual positions across all samples.

3. Results and Discussion

Table 1 compares the performance of U-Net, existing methods (Seo et al., 2021 [20]; Choi et al., 2022 [25]), and the proposed context-aware learning method using MRE under various conditions. MRE values were obtained through five individual training trials for each method to ensure the robustness of the comparison.

Table 1. Performance comparisons of basic learning and context-aware learning.

Weather Conditions	Method	MRE (Pixel)			
		Average	Point 1	Point 2	Point 3
All days	U-Net	6.72	6.39	6.62	7.16
	Seo et al., 2021 [20]	6.28	5.87	6.24	6.68
	Choi et al., 2022 [25]	5.86	5.66	5.81	6.23
	Proposed context-aware learning	**5.46**	**5.17**	**5.31**	**5.93**
Sunny day	U-Net	6.47	6.19	6.25	6.97
	Seo et al., 2021 [20]	6.05	5.68	5.92	6.38
	Choi et al., 2022 [25]	5.62	5.30	5.41	5.96
	Proposed context-aware learning	**5.30**	**5.15**	**5.01**	**5.75**
Cloudy day	U-Net	6.97	6.58	6.99	7.35
	Seo et al., 2021 [20]	6.52	6.05	6.56	6.97
	Choi et al., 2022 [25]	6.10	6.02	6.21	6.50
	Proposed context-aware learning	5.62	5.15	5.61	6.11
Sunny day (morning)	U-Net	6.75	6.14	6.57	7.55
	Seo et al., 2021 [20]	6.24	5.84	6.12	6.76
	Choi et al., 2022 [25]	5.68	**5.35**	5.47	6.16
	Proposed context-aware learning	**5.54**	5.41	**5.13**	**6.15**
Sunny day (afternoon)	U-Net	6.18	6.24	5.93	6.38
	Seo et al., 2021 [20]	5.85	5.52	5.72	6.00
	Choi et al., 2022 [25]	5.56	5.24	5.34	5.76
	Proposed context-aware learning	**5.06**	**4.95**	**4.88**	**5.35**

* **Bold** indicates the best performance.

The context-aware learning consistently reduced the MRE across all conditions, demonstrating its superior accuracy. On average, MRE decreases from 6.72 (U-Net) to 5.46 pixels in general conditions, from 6.47 (U-Net) to 5.30 pixels on sunny days, and from 6.97 (U-Net) to 5.62 pixels on cloudy days. This improvement was even more pronounced during specific times of the day; for instance, the MRE on sunny afternoons drops from 6.75 (U-Net) to 5.06 pixels. Existing methods (Seo et al., 2021 [20]; Choi et al., 2022 [25]) demonstrate performance improvements that are not as significant as those achieved by the proposed context-aware learning. These findings validate the effectiveness of the proposed method, particularly under variable lighting and weather conditions, and emphasize its robustness and adaptability. Figure 7 presents box plots of MRE across different weather conditions. The results show that the proposed method consistently outperformed the other methods. The box plots highlight the variability in performance across the different methods, with the proposed method showing less spread and lower median MRE values. This consistent performance underscores the effectiveness of the proposed context-aware learning approach for handling different weather conditions, making it a reliable choice for practical applications.

Table 2 provides an insightful analysis of how the number of training images affects the performance of different learning approaches, including U-Net, existing methods (Seo et al., 2021 [20]; Choi et al., 2022 [25]), and the proposed context-aware learning method.

Table 2. Performance comparisons of basic learning and context-aware learning based on the number of training images.

Number of Training Images	Method	MRE (Pixels)			
		Average	Point 1	Point 2	Point 3
2590 (100%)	U-Net	6.72	6.39	6.62	7.16
	Seo et al., 2021 [20]	6.28	5.87	6.24	6.68
	Choi et al., 2022 [25]	5.86	5.66	5.81	6.23
	Context-aware learning	**5.46**	**5.15**	**5.31**	**5.93**
1480 (≈50%)	U-Net	7.13	6.84	7.11	7.45
	Seo et al., 2021 [20]	6.69	6.46	6.72	6.88
	Choi et al., 2022 [25]	6.59	6.12	6.66	6.99
	Context-aware learning	**5.68**	**5.67**	**5.44**	**5.93**
740 (≈25%)	U-Net	7.98	7.61	7.77	8.55
	Seo et al., 2021 [20]	7.45	7.02	7.22	8.11
	Choi et al., 2022 [25]	7.27	6.58	7.33	7.91
	Context-aware learning	**6.08**	**5.84**	**5.99**	**6.41**

* **Bold** indicates the best performance.

The performance was evaluated using the MRE, in pixels, under conditions of full data availability (2590 images at 100%) and limited data availability (1480 images at about 50% and 740 images at 25%). The results highlighted a clear trend: as the number of training images decreased, the MRE increased for all methods across all points, indicating a decline in performance. The performance disparity became more pronounced when the data availability was reduced to about 25% (740 images). Under this condition, the context-aware learning approach showed an increase in the average MRE from 5.46 to 6.08, reflecting an error increase of only 0.62 pixels. In contrast, U-Net experienced a more significant degradation, with the MRE rising from 6.72 to 7.98 (an error increase of 1.26 pixels). Existing methods also exhibit intermediate levels of performance degradation but do not match the robustness of the context-aware learning approach. These results indicate that, although all methods suffer from limited data availability, the context-aware learning approach is more robust, exhibiting a smaller increment in error. The stronger resilience of the context-aware learning model in scenarios with limited data can be attributed to its ability to leverage spatial regularization effects. This method integrates contextual information, which effectively constrains the learning process, thereby allowing more accurate information to be inferred from fewer data points. Table 3 presents the results of the ablation study conducted to evaluate the impact of varying the number of heads in the transformer within the proposed framework. The optimal performance was achieved with three heads, where the MRE was at its lowest at 5.46. Using one or two heads resulted in MREs of 5.56 and 5.58, respectively, suggesting that fewer heads may limit the model's ability to capture sufficient data relationships. Conversely, increasing the number of heads beyond three, such as four or five, caused their MREs to slightly rise to 5.50 and 5.48, likely due to inefficiencies such as overfitting or a redundancy in processing additional attention mechanisms. These findings indicate that while additional heads can capture more complex data relationships, an excessive number of heads may introduce inefficiencies, ultimately reducing performance.

Table 3. Ablation study on transformers with different numbers of heads.

Head Number	1	2	3	4	5
MRE	5.56	5.58	5.46	5.50	5.48

Table 4 evaluates the impact of varying the soft-argmax parameter β using MRE. The results show that increasing β from 1 to 10 leads to a significant improvement in MRE, decreasing from 5.56 to 5.46. However, further increasing β to 15 and 20 leads to an increase in the MRE to 5.93 and 7.18, respectively. This suggests that too high a value of β may lead to over-sharpening, where the model might be overfit to noise or specific features, losing the ability to generalize well across different scenarios.

Table 4. The impact of soft-argmax parameter β on the performance.

β	1	5	10	15	20
MRE	5.66	5.53	5.46	5.93	7.18

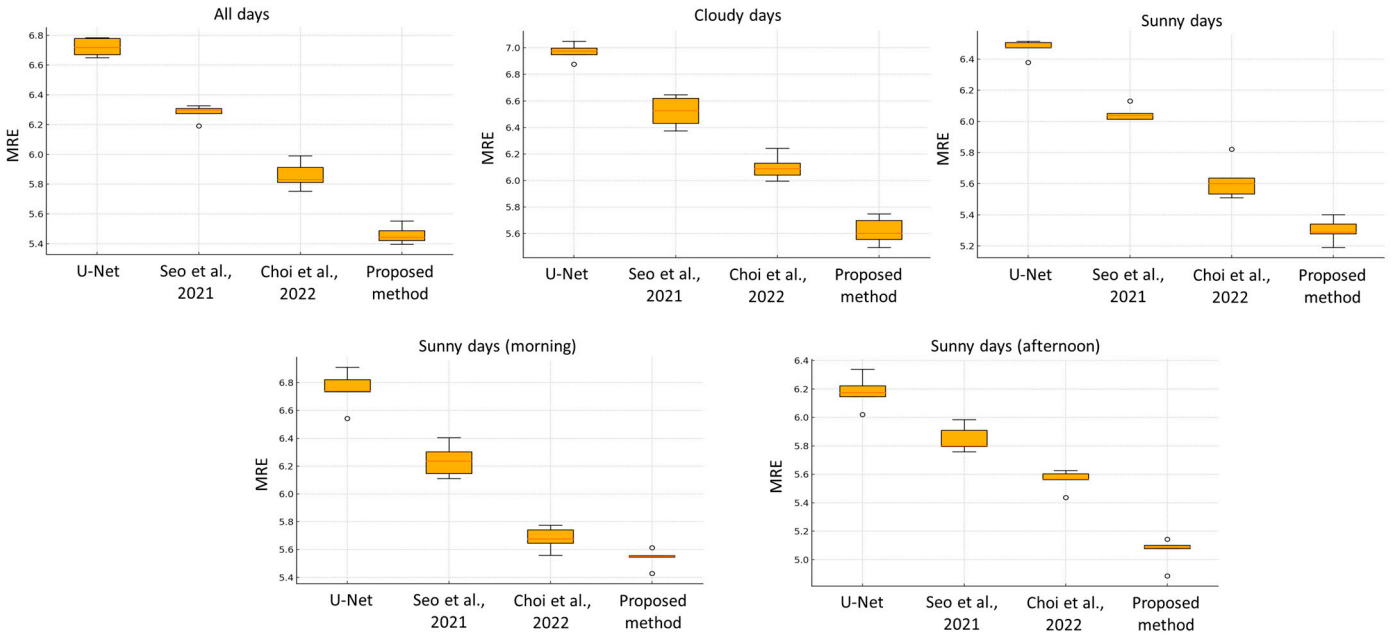

Figure 7. Box plots of MREs for different methods under various weather conditions. The MRE scores were obtained from five different training sessions. The methods compared are U-Net, Seo et al., 2021 [20], Choi et al., 2022 [25], and the proposed method. The weather conditions include all days, sunny days, cloudy days, sunny days (morning), and sunny days (afternoon). The dots in the box plots represent outliers, which are data points that fall outside the typical range of the distribution.

Visualization Results

A visual comparison of the tillage boundary detection under varying weather conditions and times of day is shown in Figure 8.

Figure 8. Visual comparison of tillage boundary detection under different lighting conditions using the basic heatmap regression (U-Net) and context-aware learning. Yellow dashed boxes indicate localization errors.

Each image has distinct tillage boundary points marked by blue-, green-, and red-colored dots for the first, second, and third points, respectively. These points were strategically placed to represent a significant area for accurately guiding agricultural machinery. The provided image collage compares the results of basic heatmap regression using U-Net versus a context-aware learning approach that integrates U-Net with a transformer under different weather conditions. On sunny days, inaccuracies in tillage point detection were visible in the vanilla heatmap regression approach, as indicated by the yellow dashed boxes. However, the context-aware learning approach of combining U-Net with a transformer significantly improved the results. Similarly, on cloudy days, basic heatmap regression struggled with mislocalization, whereas the context-aware learning method demonstrated enhanced results and robustness against varying environmental conditions. Figure 9 illustrates the self-attention matrices of the trained transformer model, specifically visualizing the relationships among three boundary points. Each subplot comprises an image and a corresponding attention matrix. In the images, lines connect the three points, and the thickness of the lines represents the attention scores. Thicker lines denote higher attention scores, indicating stronger attention or relevance between the points. The matrices on the right display attention scores in red. These scores quantify the importance of each boundary point in the context of the others, as learned by the

transformer model. Our findings indicate that boundary points exhibit uniformly low attention scores when they reference themselves, and tend to allocate more focus to other points. In addition, no noticeable changes in the patterns of the attention score metrics were observed with variations in the environment or amount of training data.

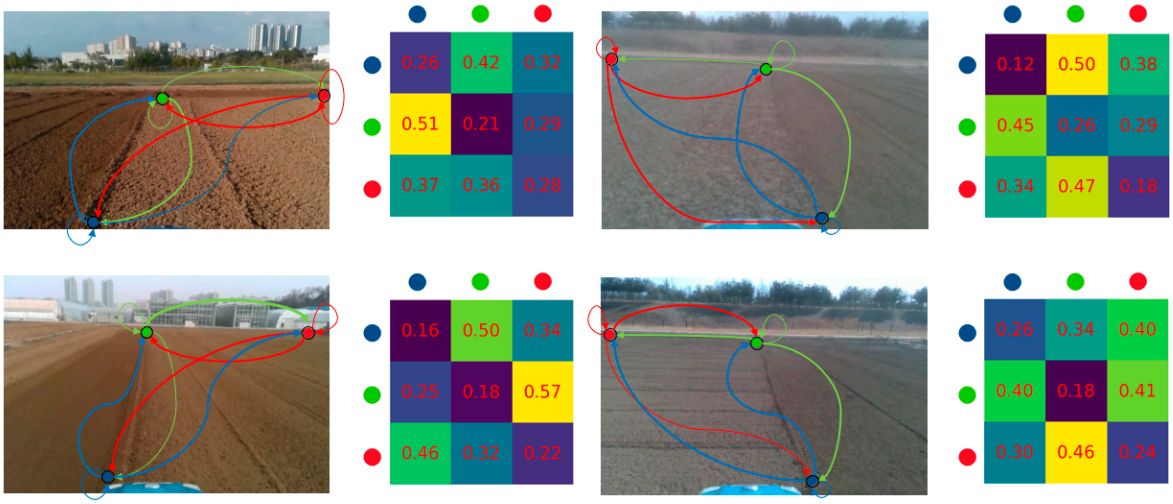

Figure 9. Visualization of self-attention matrices within the trained transformer model, demonstrating the relationships among three boundary points. The attention scores are indicated in red within the matrices. The lines in the images represent the attention scores, where thicker lines indicate higher attention scores.

Figure 10 illustrates the loss curves for the training and validation datasets using 100%, 50%, and 25% of the available images. The loss curves show that the model performance declined as the amount of training data decreased. With the full dataset (2590 images), both the training and validation losses decreased steadily, indicating good generalization. With half of the data (1480 images), the validation loss fluctuated more and was higher, indicating reduced generalization. Using only 25% of the dataset (740 images), the validation loss became more erratic but stabilized after early epochs without overfitting.

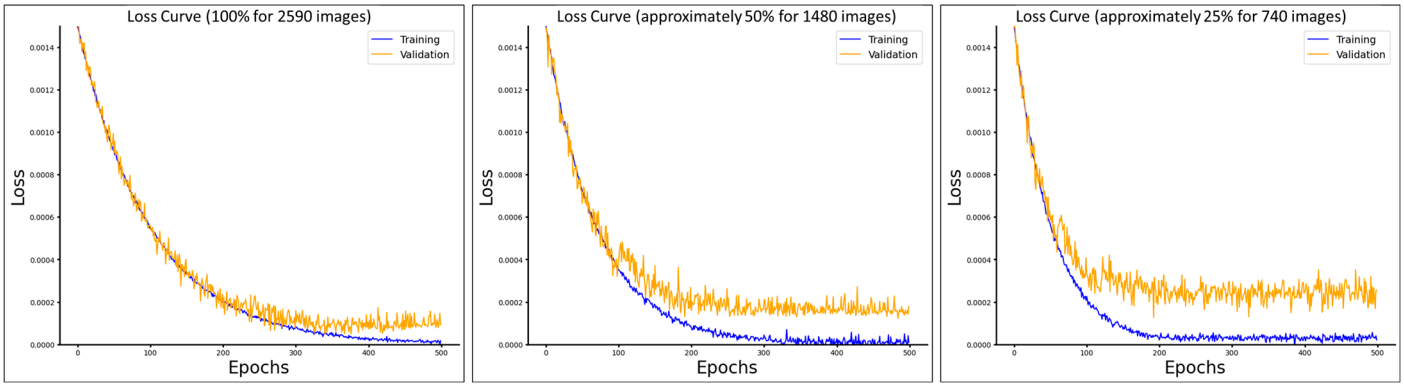

Figure 10. Loss curves for training and validation with varying amounts of data: (left) 100% of the data (2590 images), (middle) approximately 50% of the data (1480 images), and (right) approximately 25% of the data (740 images).

4. Conclusions

This study proposed a heatmap regression-based context-aware network for tillage boundary detection. The proposed method comprises both the heatmap regression and transformer methods, and operates in an end-to-end manner, simplifying the detection

process and making it suitable for real-time applications in agriculture. The proposed approach was tested under various conditions, demonstrating a consistent reduction in the MRE across all scenarios, particularly when data were limited.

The context-aware learning model excels because of its ability to leverage spatial contextual information, allowing it to maintain high accuracy even with reduced datasets. This advantage is particularly pronounced in scenarios with varying environmental conditions, such as changes in lighting, soil textures, or weather patterns. The model achieves this robustness by utilizing the transformer's capability to capture broader contextual cues, effectively bridging the gap that conventional methods, such as the basic U-Net model, may overlook. Consequently, the context-aware learning approach exhibits remarkable robustness and adaptability, effectively handling the challenges posed by varying environmental conditions and data scarcity. By integrating transformers, the model balances local and global relationships, enhancing its ability to generalize across diverse conditions. This adaptability is critical for agricultural tasks, where environmental variability and data scarcity often present significant challenges. As highlighted in Table 3, our model consistently outperforms existing methods in these scenarios, demonstrating lower MRE values and greater stability, further validating its robustness and adaptability.

In the future, we plan to conduct long-term studies to assess the performance and reliability of these models across multiple growing seasons. Such studies would offer deeper insights into the practical viability and long-term benefits of these models. However, one limitation of the current study is the focus on specific environmental and dataset conditions, which may not fully represent all agricultural scenarios. To address this, we plan to optimize the proposed methodology to ensure its effectiveness across a broader range of environmental conditions. This includes refining the model to better handle diverse agricultural scenarios, such as varying soil textures, crop types, climatic conditions, and even overnight operations, to enhance its robustness and accuracy in real-world applications. Although this study focused on tillage boundary detection, future research could extend the application of context-aware learning to other agricultural tasks.

This study provides valuable insights for manufacturers of autonomous tractors and navigation systems. The proposed landmark regression approach, unlike traditional tillage segmentation methods, accurately detects tillage boundaries and adapts to diverse conditions, significantly enhancing the precision and efficiency of autonomous operations. Its robustness and adaptability offer a practical solution for real-world challenges, driving innovation in agricultural automation and contributing to sustainable farming practices.

Author Contributions: G.-S.H.: writing—original draft, writing—review and editing, investigation, and validation. K.O.: writing—original draft, writing—review and editing, conceptualization, data curation, formal analysis, methodology, project administration, supervision, visualization, and software. All authors have read and agreed to the published version of the manuscript.

Funding: This study did not receive any specific grants from funding agencies in the public, commercial, or non-profit sectors.

Data Availability Statement: The datasets used and analyzed in the current study are available from the corresponding author upon reasonable request.

Acknowledgments: This paper was supported by Wonkwang University in 2022.

Conflicts of Interest: The authors declare that they have no competing financial interests or personal relationships that may have influenced the work reported in this study.

References

1. Kim, W.-S.; Lee, D.-H.; Kim, G.; Sim, T.; Kim, Y.-J. One-shot classification-based tilled soil region segmentation for boundary guidance in autonomous tillage. *Comput. Electron. Agric.* **2021**, *189*, 106371. https://doi.org/10.1016/j.compag.2021.106371.
2. Liu, Y.; Ma, X.; Shu, L.; Hancke, G.-P.; Abu-Mahfouz, A.-M. From industry 4.0 to agriculture 4.0: Current status, enabling technologies and research challenges. *IEEE Trans. Ind. Inform.* **2020**, *17*, 4322–4334. https://doi.org/10.1109/TII.2020.3003910.
3. Abdalla, A.; Cen, H.; Wan, L.; Mehmood, K.; He, Y. Nutrient status diagnosis of infield oilseed rape via deep learning-enabled dynamic model. *IEEE Trans. Ind. Inform.* **2020**, *17*, 4379–4389. https://doi.org/10.1109/TII.2020.3009736.
4. Wang, D.; Zhang, D.; Yang, G.; Xu, B.; Luo, Y.; Yang, X. SSRNet: In-field counting wheat ears using multi-stage convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 4403311. https://doi.org/10.1109/tgrs.2021.3093041.
5. Kim, G.-H.; Seo, D.-S.; Kim, K.-C.; Hong, Y.; Lee, M.; Lee, S.; Kim, H.; Ryu, H.-S.; Kim, Y.-J.; Chung, S.-O.; et al. Tillage boundary detection based on RGB imagery classification for an autonomous tractor. *Korean J. Agric. Sci.* **2020**, *47*, 205–217. https://doi.org/10.7744/kjoas.20200006.
6. Kim, W.-S.; Lee, D.-H.; Kim, Y.-J.; Kim, T.; Hwang, R.-Y.; Lee, H.-J. Path detection for autonomous traveling in orchards using patch-based CNN. *Comput. Electron. Agric.* **2020**, *175*, 105620. https://doi.org/10.1016/j.compag.2020.105620.
7. Bah, M.D.; Hafiane, A.; Canals, R. Deep learning with unsupervised data labeling for weed detection in line crops in UAV images. *Remote Sens.* **2018**, *10*, 1690. https://doi.org/10.3390/rs10111690.
8. Winterhalter, W.; Fleckenstein, F.V.; Dornhege, C.; Burgard, W. Crop row detection on tiny plants with the pattern Hough transform. *IEEE Robot. Autom. Lett.* **2018**, *3*, 3394–3401. https://doi.org/10.1109/lra.2018.2852841.
9. Su, J.; Yi, D.; Su, B.; Mi, Z.; Liu, C.; Hu, X.; Xu, X.; Guo, L.; Chen, W.-H. Aerial visual perception in smart farming: Field study of wheat yellow rust monitoring. *IEEE Trans. Ind. Inform.* **2021**, *17*, 2242–2249. https://doi.org/10.1109/tii.2020.2979237.
10. Lu, E.; Xue, J.; Chen, T.; Jiang, S. Robust Trajectory Tracking Control of an Autonomous Tractor-Trailer Considering Model Parameter Uncertainties and Disturbances. *Agriculture* **2023**, *13*, 869. https://doi.org/10.3390/agriculture13040869.
11. Bakker, T.; van Asselt, K.; Bontsema, J.; Müller, J.; van Straten, G. Autonomous navigation using a robot platform in a sugar beet field. *Biosyst. Eng.* **2011**, *109*, 357–368. https://doi.org/10.1016/j.biosystemseng.2011.05.001.
12. Lenain, R.; Thuilot, B.; Cariou, C.; Martinet, P. Mixed kinematic and dynamic sideslip angle observer for accurate control of fast off-road mobile robots. *J. Field Robot.* **2010**, *27*, 181–196. https://doi.org/10.1002/rob.20319.
13. Si, J.; Niu, Y.; Lu, J.; Zhang, H. High-precision estimation of steering angle of agricultural tractors using GPS and low-accuracy MEMS. *IEEE Trans. Veh. Technol.* **2019**, *68*, 11738–11745. https://doi.org/10.1109/tvt.2019.2949298.
14. Malavazi, F.B.P.; Guyonneau, R.; Fasquel, J.B.; Lagrange, S.; Mercier, F. LiDAR-only based navigation algorithm for an autonomous agricultural robot. *Comput. Electron. Agric.* **2018**, *154*, 71–79. https://doi.org/10.1016/j.compag.2018.08.034.
15. Shalal, N.; Low, T.; McCarthy, C.; Hancock, N. Orchard mapping and mobile robot localisation using on-board camera and laser scanner data fusion—Part A: Tree detection. *Comput. Electron. Agric.* **2015**, *119*, 254–266. https://doi.org/10.1016/j.compag.2015.09.025.
16. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. https://doi.org/10.1038/nature14539.
17. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 2921–2929. https://doi.org/10.1109/cvpr.2016.319.
18. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Proceedings of the 18th International Conference, Munich, Germany, 5–9 October 2015, Proceedings, Part III*; Springer: Berlin, Germany, 2015; pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
19. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Springer: Berlin, Germany, 2018; pp. 801–818. https://doi.org/10.1007/978-3-030-01234-2_49.
20. Seo, D.-S.; Won, J.-H.; Yang, C.-Y.; Kim, G.-K.; Kwon, K.-D.; Kim, K.-C.; Hong, Y.-K.; Ryu, H.-S. Development of boundary detection methods based on images for path following of autonomous tractor. *J. Korean Inst. Commun. Inf. Sci.* **2021**, *46*, 2078–2087. https://doi.org/10.7840/kics.2021.46.11.2078.
21. Kounalakis, T.; Triantafyllidis, G.A.; Nalpantidis, L. Deep learning-based visual recognition of rumex for robotic precision farming. *Comput. Electron. Agric.* **2019**, *165*, 104973. https://doi.org/10.1016/j.compag.2019.104973.
22. Ferentinos, K.P. Deep learning models for plant disease detection and diagnosis. *Comput. Electron. Agric.* **2018**, *145*, 311–318. https://doi.org/10.1016/j.compag.2018.01.019.

23. Zhang, Z.; Kayacan, E.; Thompson, B.; Chowdhary, G. High precision control and deep learning-based corn stand counting algorithms for agricultural robots. *Auton. Robots* **2020**, *44*, 1289–1302. https://doi.org/10.1007/s10514-020-09915-y.
24. Shalal, N.; Low, T.; McCarthy, C.; Hancock, N. Orchard mapping and mobile robot localisation using on-board camera and laser scanner data fusion—Part B: Mapping and localisation. *Comput. Electron. Agric.* **2015**, *119*, 267–278. https://doi.org/10.1016/j.compag.2015.09.026.
25. Choi, D.-H.; Oh, K.-H. Tillage boundary detection using heat map regression for autonomous tractors. In Proceedings of the 11th International Conference on Smart Media & Application (SMA2022), Saipan, Northern Mariana Islands, 19–22 October 2022.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
27. Luvizon, D.C.; Tabia, H.; Picard, D. Human pose regression by combining indirect part detection and contextual information. *Comput. Graph.* **2019**, *85*, 15–22. https://doi.org/10.1016/j.cag.2019.09.002.
28. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450. https://doi.org/10.48550/arXiv.1607.06450.
29. Loshchilov, I.; Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017. https://doi.org/10.48550/arXiv.1608.03983.