agriculture

MDPI

Article

Drivable Agricultural Road Region Detection Based on Pixel-Level Segmentation with Contextual Representation Augmentation

Yefeng Sun [1], Liang Gong [1,2,*], Wei Zhang [1], Bishu Gao [1], Yanming Li [1] and Chengliang Liu [1,2]

[1] School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; ysun3019@sjtu.edu.cn (Y.S.); zhang_wei@sjtu.edu.cn (W.Z.); gaobishu2021@sjtu.edu.cn (B.G.); ymli@sjtu.edu.cn (Y.L.); chlliu@sjtu.edu.cn (C.L.)

[2] MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China

* Correspondence: gongliang_mi@sjtu.edu.cn

Abstract: Drivable area detection is crucial for the autonomous navigation of agricultural robots. However, semi-structured agricultural roads are generally not marked with lanes and their boundaries are ambiguous, which impedes the accurate segmentation of drivable areas and consequently paralyzes the robots. This paper proposes a deep learning network model for realizing high-resolution segmentation of agricultural roads by leveraging contextual representations to augment road objectness. The backbone adopts HRNet to extract high-resolution road features in parallel at multiple scales. To strengthen the relationship between pixels and corresponding object regions, we use object-contextual representations (OCR) to augment the feature representations of pixels. Finally, a differentiable binarization (DB) decision head is used to perform threshold-adaptive segmentation for road boundaries. To quantify the performance of our method, we used an agricultural semi-structured road dataset and conducted experiments. The experimental results show that the *mIoU* reaches 97.85%, and the *Boundary IoU* achieves 90.88%. Both the segmentation accuracy and the boundary quality outperform the existing methods, which shows the tailored segmentation networks with contextual representations are beneficial to improving the detection accuracy of the semi-structured drivable areas in agricultural scene.

Keywords: semi-structured road detection; contextual representation; pixel-level segmentation; agricultural robot

check for updates

Citation: Sun, Y.; Gong, L.; Zhang, W.; Gao, B.; Li, Y.; Liu, C. Drivable Agricultural Road Region Detection Based on Pixel-Level Segmentation with Contextual Representation Augmentation. *Agriculture* **2023**, *13*, 1736. https://doi.org/10.3390/agriculture13091736

Academic Editor: José Lima

Received: 12 July 2023
Revised: 24 August 2023
Accepted: 31 August 2023
Published: 1 September 2023

1. Introduction

Accurate detection of semi-structured roads and unstructured roads has a wide range of applications. Taking agriculture as an example, the agricultural environment is harsh, and there are various safety hazards [1]. With the ageing of the population and the reduction in the labor force, automation and intelligence will be the future development direction of agricultural equipment [2]. Drivable area detection using machine vision is crucial for the autonomous navigation of agricultural robots in natural environments. The road environment has three categories, structured roads, semi-structured roads and unstructured roads [3]. For structured roads, such as well-marked highways, lane detection is commonly used to find drivable areas. However, roads in the agricultural scene are mostly semi-structured with ambiguous boundaries and no artificial markings [4]. The road features are complex and easily disturbed by the environment. Compared with structured road detection, semi-structured road detection in agricultural scenes is more challenging [5].

Semi-structured road detection in complex environments has always been a research hotspot in autonomous navigation. The research on vision-based semi-structured road detection can be divided into feature-based road detection, model-based road detection, and image segmentation-based road detection.

Feature-based road detection methods mainly use salient visual characteristics to detect road regions, such as the vanishing point (VP), color, texture, boundary, and other features. Yang G. et al. [5] proposed a contourlet transform framework for VP detection in a single image for unstructured roads. Shi J. et al. [6] used Gabor filters and particle filters to estimate the road texture direction, and designed a noise-insensitive observation model to vote for vanishing points. Liu Y.B. et al. [7] proposed an unstructured road VP detection solution combining a convolutional neural network and heatmap regression. Hernandez D.E. et al. [8] designed a vision-based road boundary tracking system, which detects and tracks the boundary of unstructured roads based on color differences and texture lines. Liu G. et al. [9] estimated the probability distribution of lane parameters using multiple kernel densities, and presented a partitioned particle filter approach for lane detection and tracking. The advantage of feature-based methods is that they rely less on prior knowledge and are not sensitive to road shapes. The disadvantage is that the requirements for road characteristics are high, and misjudgment is prone to occur when there is environmental interference (e.g., water stains, cracks, and shadows).

Model-based methods construct parametric road boundary models with straight lines or curves, and use road model matching to solve the parameters. Perng J.W. et al. [10] adopted a hyperbolic model to fit the feature points of the lane lines and used a particle filter to track the lanes. Based on the third-order B-spline curve model, Cao J. et al. [11] used the random sample consensus algorithm to fit the lanes and evaluate the fitted curve. Wang K. et al. [12] used Bezier splines to construct a variable road template whose parameters could be solved by the improved RANSAC algorithm. Yuan Y. et al. [13] used a structural support vector machine to discriminate road boundary and non-boundary instances for fitting complete boundary lines. Wang X. et al. [14] proposed an unstructured road detection method based on contour selection, which uses Hough transform and contour detection function to detect lines and edges, and takes the line with the best coincidence as the edge contour. These feature-based methods are generally applied to roads with relatively regular shapes. The advantage is that the performance is robust, and the detected road area is relatively complete. The disadvantage is also apparent. It is difficult to accurately match roads with complex shapes, and the wrong road model selection will lead to detection failure [15].

Segmentation-based methods describe road detection as a classification problem of road regions and background regions. Different segmentation algorithms are used to separate different classes of regions with salient features. Traditional road semantic segmentation algorithms are based on handcrafted features (e.g., SIFT, SURF, HoG, etc.) and classical classifiers (e.g., multiple nearest neighbors [16], random forest classifiers [4], conditional random fields [17], etc.) [18]. Alam A. et al. [16] proposed a road detection system for classifying unstructured roads into road and non-road regions using a multiple nearest neighbor (NN) classifier and a soft voting aggregation approach. Xiao L. et al. [4] proposed a structural random forest-based road detection algorithm that exploits the context of image patches and the structural information of labels to obtain consistent segmentation results. Wang Q. et al. [19] combined depth cues with traditional RGB colors to predict road and non-road regions using context-aware label transfer. Geng L. et al. [20] used the Markov random field to optimize the road detection results according to the relationship between super-pixel neighborhoods. These traditional methods explicitly clustered or grouped the pixels while classifying the patch category. In recent years, semantic segmentation algorithms have witnessed tremendous progress with the development of deep learning. Many deep learning-based semantic segmentation networks have been proposed, such as U-net [21], SegNet [22] and DeepLab series [23–26]. These methods have a wide range of applications and improvements in road segmentation. Eff-UNet [18] used Efficient Net as the encoder feature extractor and decoder of UNet, combining high-level features and low-level spatial information to achieve accurate segmentation. Lane-DeepLab [27] added the ASPP to optimize the encoder–decoder structure, and employed the Semantic Embedding Branch and the Single Stage Headless modules to obtain multi-level semantic features.

The segmentation method based on deep learning has a strong learning ability and good adaptability to the environment, and shows unique advantages in road detection tasks. However, a large number of training samples are required in the early stage. Unfortunately, there are few public datasets of semi-structured roads, which brings difficulties to the model's training.

The semi-structured roads, such as roads and fixed paths in greenhouses, mainly have the following characteristics: (1) there are neither standardized lane lines nor identifiable manual signs in the drivable area. (2) The environment is complex and changeable; water stains, shadows, and changing lighting will interfere with road features. (3) Road boundaries are difficult to distinguish because soil and water stains may blur road boundaries. This paper proposes a high-resolution detection network for semi-structured roads to address the challenge of semi-structured drivable area detection in agriculture. Most encoder–decoder architectures concatenate high-to-low encoders to extract road features gradually, then recover high-resolution from low-resolution feature representations. However, as the network layer deepens, the size of the feature map gradually decreases. This operation inevitably produces blurred feature maps after multiple convolutions and loses some crucial details, which is fatal for semi-structured roads in greenhouse boundary extraction. The backbone of our method adopts HRNet [28] to extract high-resolution road features in parallel at multiple scales. HRNet changes the connection between high-resolution and low-resolution resolutions from series to parallel, thus the high-resolution could representationed throughout the whole network structure. To strengthen the relationship between pixels and their corresponding object regions, we use OCR [29] to enhance the feature representations of pixels. Finally, a DB module [30] is used as a decision head to perform adaptive threshold segmentation of road boundaries. The loss function includes two parts: segmentation loss and threshold map loss. To quantify the performance of the proposed method, we make a dataset of semi-structured roads in agricultural scenes and conduct experiments in agricultural scenes.

To summarize, the contributions of this study are as follows:

(1) This paper proposes a high-resolution detection network for semi-structured roads in facility agriculture. The network uses HRNet to extract high-resolution road features in parallel at multiple scales, then uses OCR to enhance the feature representations, and finally uses a DB decision head to segment road boundaries adaptively.

(2) A loss function is designed, including segmentation and threshold map losses.

(3) A dataset of semi-structured agricultural roads for greenhouses is produced, and the method in this paper is validated.

The rest of the paper is organized as follows: Section 2 presents the high-resolution detection method for semi-structured roads in detail. Then, Section 3 discusses the experimental results. Finally, conclusions and future work are presented in Section 4.

2. Theory and Method

2.1. Architecture of Agricultural Semi-Structured Road Detection Network

The agricultural semi-structured road detection network consists of three main components: high-resolution feature extraction, road-contextual representations, and threshold-adaptive boundary segmentation. To extract high-resolution features of semi-structured roads, we must adopt the HRNet as the backbone to preserve high-resolution features during downsampling. To distinguish the features of drivable and non-drivable regions more clearly, we can use contextual representations to augment road objectness. Finally, an adaptive threshold-learned DB decision head is used to segment road boundaries. The pipeline of the semi-structured road detection network is shown in Figure 1.

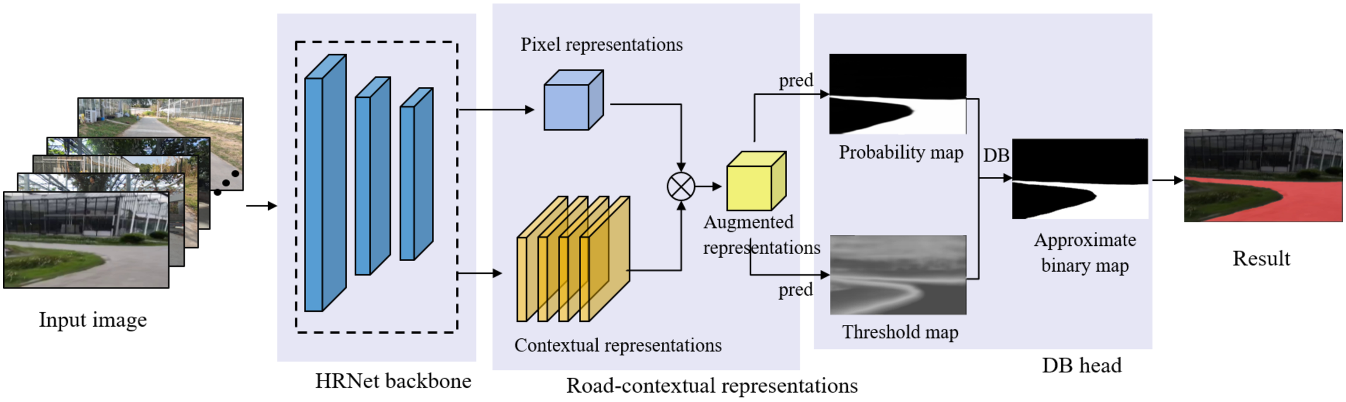

Figure 1. The structure of the agricultural semi-structured road detection network.

2.2. High-Resolution Road Feature Extraction

As mentioned above, the encoder–decoder architectures lead to the insufficient resolution of feature maps after multiple convolutions, thus losing some important shape and boundary details, which is fatal for semi-structured road detection. In contrast, HRNet connects high-to-low resolution convolution streams in parallel rather than in series, which maintains high-resolution representations throughout the whole process.

Therefore, high-resolution feature extraction uses HRNet-W48 as the backbone to extract the semi-structured road features of different levels. The backbone contains four stages with four parallel convolution streams. The first stage includes 4 residual units, each containing a 3 × 3 convolution and two 1 × 1 convolutions. The 1 × 1 convolution is mainly used to change the dimension of feature maps, and the 3 × 3 convolution is used to extract features. The second, third, and fourth stages include 1, 4, and 3 modularized blocks, respectively. Each branch of the modularized block contains 4 residual units. The modularized block performs upsampling of low-resolution features, downsampling of high-resolution features, and fusion of feature maps at the same level. The multi-resolution fusion of feature maps at different levels can obtain rich road features. For an image, by adding a downsampling branch from high resolution to low resolution, the corresponding number of channels will be doubled, and the resolution will be half. In the fourth stage, the channels of the convolutions of the four resolutions are 48, 96, 192, and 384, and the resolutions are 1/4, 1/8, 1/16, and 1/32. In the final feature fusion stage, we adopt HRNetV2 as the representation head, which uses bilinear upsampling to merge low-resolution and high-resolution representations to obtain pixel representations of u semi-structured roads.

2.3. Road-Contextual Representations

The agricultural environment is complex and changeable. To distinguish the features of drivable and non-drivable regions more clearly, we must enhance the pixel representations with the contextual relationship between pixels and corresponding object regions [28]. The network details are shown in Figure 2. The backbone outputs two parts: pixel representations and soft object regions. The pixel representations are the feature representations of semi-structured roads outputted by the HRNetV2 representation head. In this section, we define two classes: road and non-road. The soft object regions are regions of road and non-road roughly segmented from the backbone through supervised training. Each position of each layer represents the probability that the corresponding pixel belongs to the class represented by this layer. During training, we learn the object region generator under the supervision of the ground-truth segmentation using the binary cross-entropy loss.

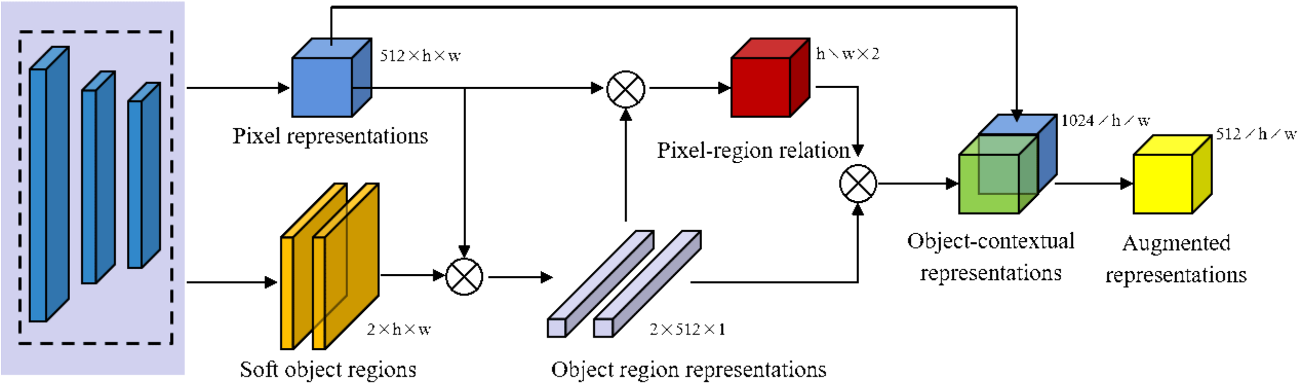

Figure 2. Pixel representations enhanced by object-contextual features.

Object region representations are defined as the aggregation of the feature representations of pixels weighted by their degrees belonging to the *k*-th object region:

$$\mathrm{f}_k = \sum_{i \in I} \widetilde{m}_{ki} \mathrm{x}_i$$

where f_k is the *k*-th object region representation and x_i is the representation of pixel p_i of image *I*. $\widetilde{m}_{ki}$ is the normalized degree for pixel p_i belonging to the *k*-th object region.

We use self-attention [31] to combine contextual information and pixel representations to obtain object-contextual representations:

$$\mathrm{y}_i = \rho\left(\sum_{k=1}^{2} w_{ik} \delta(\mathrm{f}_k)\right)$$

where y_i is the object-contextual representation of pixel p_i, $w_{ik} = \frac{e^{\kappa(\mathrm{x}_i, \mathrm{f}_k)}}{\sum_{j=1}^{2} e^{\kappa(\mathrm{x}_i, \mathrm{f}_j)}}$ is the relationship between pixel and object regions, $\kappa(\mathrm{x}, \mathrm{f}) = \phi(\mathrm{x})^T \psi(\mathrm{f})$ is the unnormalized relation function and $\phi(\cdot)$, $\psi(\cdot)$, $\rho(\cdot)$ and $\delta(\cdot)$ are transformation functions implemented by 1 × 1 conv → BN → ReLU.

Augmented representations are a fusion of the object-contextual representations and pixel representations, as demonstrated below.

$$\mathrm{z}_i = g\left(\left[\mathrm{x}_i^T \quad \mathrm{y}_i^T\right]^T\right)$$

where z_i is the augmented representation of pixel p_i and $g(\cdot)$ is a transform function implemented by 1 × 1 conv → BN→ ReLU.

2.4. Threshold-Adaptive Boundary Segmentation

Threshold-adaptive segmentation models road and non-road regions in two channels, calculating and comparing the scores of each pixel belonging to the different regions. The pavement characteristics of semi-structured roads are inconsistent, and the boundaries are not obvious, so it is difficult to set boundary thresholds to segment the road regions. Traditional segmentation methods set a fixed threshold to convert the probability map generated by the segmentation network into a binary image. However, this method is not differentiable and cannot be optimized in the network. Inspired by DBNet++ [30] in scene text detection, we adopt the DB module with an adaptive boundary threshold. Combining the DB module to optimize the segmentation network can obtain highly robust segmentation results.

We perform supervised training on the augmented representations in Section 2.2 to predict the probability map and the threshold map. Combining the probability map and the threshold map, we use an approximate step function to perform differentiable binarization on each pixel to obtain an approximate binary map.

$$B_{i,j} = \frac{1}{1 + e^{-50(P_{i,j} - T_{i,j})}}$$

where B is the approximate binary map, T is the adaptive threshold map, and P is the adaptive threshold map learned from the network. The function quickly changes from 0 to 1 when the probability is close to the threshold, and this process is derivable. Therefore, the network can adaptively predict the segmentation threshold of each pixel of the road boundary in the image, making the semi-structured road boundary detection more robust.

2.5. Loss Function

This paper has three road segmentation outputs: the soft object region in the backbone, the probability map and the approximate binary map in the DB Head. All three parts can generate road segmentation results, but the fineness of segmentation is different. The classification of road and non-road is a binary classification problem. The segmentation loss of each output applies the binary cross-entropy (BCE) loss, which is defined as

$$L_s = L_p = L_b = -\frac{1}{N}\sum_{i=1}^{N} y_i log x_i + (1 - y_i)\log(1 - x_i)$$

where L_s is the segmentation loss for the soft object region, L_p is the segmentation loss for the probability map, and L_b is the segmentation loss for the approximate binary map, N is the number of pixels, $y_i \in \{0, 1\}$ is the label of the pixel, and $x_i \in [0, 1]$ is the probability of the pixel.

The threshold map of the DB Head can be obtained by learning or supervised training. The value of the threshold map is between 0 and 1, and the network needs to regress to continuous values. We use the L1 loss function as the threshold map loss function, denoted as L_t

$$L_t = -\frac{1}{N}\sum_{i=1}^{N}|y_i - x_i|$$

The complete loss function L consists of two parts: the segmentation loss and the threshold map loss.

$$L = \omega_s L_s + \omega_p L_p + \omega_b L_b + \omega_t L_t$$

where ω_i is the weight parameter of different parts, set to 1 in this paper. The threshold map can be supervised or unsupervised. If the threshold map is not supervised, then ω_t is 0.

3. Experiments

3.1. Dataset

The semi-structured facility agriculture scenario is complex and changeable without the support of open-source datasets. We need to make an agricultural semi-structured road dataset, because network training requires a lot of labeled data. We collected 560 semi-structured road images of agricultural scenes using ZED2 in the glass greenhouses from Shanghai Jiao Tong University and Langxia Town, Jinshan District, Shanghai. We use EISeg [32] to annotate road and non-road regions of the image interactively. To increase the diversity of the samples, we augment the dataset of semi-structured roads with the Road Traversing Knowledge (RTK) dataset [33]. The RTK dataset contains data for different road conditions, and we selected 700 images containing semi-structured roads. The 12 semantic categories of the RTK dataset are mapped into two categories, the drivable region and the non-drivable region. The drivable region includes roads, lane lines, and sidewalks, which correspond to roads in agricultural environments.

The threshold map label of the DB Head needs to be calculated according to the edge of the drivable region. Specifically, the Vatti clipping algorithm [34] is used to offset the road edge polygons inside and outside by D, respectively.

$$D = \frac{A\left(1 - r^2\right)}{L}$$

where A is the area of the road region polygon, L is the perimeter of the polygon, and r = 0.4 is the shrink ratio. The threshold map gradually decreases from a maximum threshold of 0.7 at the edge to 0.3.

In the end, the dataset has 560 labeled samples and 700 RTK samples, of which 1200 are in the training set and 60 in the validation set, as shown in Figure 3.

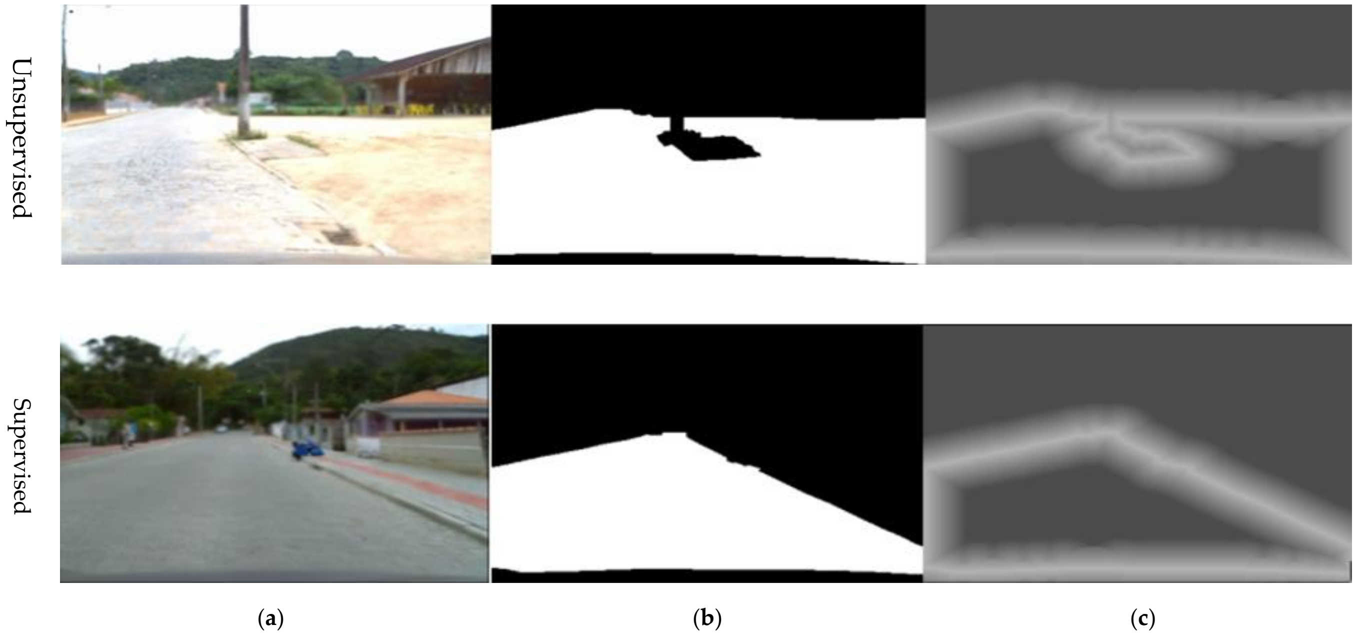

Figure 3. Examples of the agricultural semi-structured road dataset. (**a**) shows the input figure; (**b**) shows the mask for the road region; (**c**) shows the binary mask for the threshold map.

3.2. Evaluation Metrics

In this paper, the mean intersection of union (*mIoU*) of segmentation quality and *Boundary IoU* of boundary quality are selected as the primary evaluation criteria, and pixel accuracy (*PA*) is supplemented as the basis for comparison.

The *mIoU* is a common metric to measure the image overlap rate in object detection. The larger the average cross-union ratio, the more accurate the target prediction is. The calculation formula for *mIoU* is

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}$$

where k is the number of classifications in the segmentation task, p_{ii} is the i class label predicted to be the i class pixel, p_{ij} is the i class label predicted to be the j class pixel.

Compared with a large number of simple samples, the proportion of boundary samples prone to ambiguity is small. *mIoU* cannot accurately reflect the boundary quality of the segmentation algorithm. We adopt *Boundary IoU* [35] to measure the boundary segmentation accuracy of the segmentation algorithm.

$$Boundary\ IoU = \frac{|a|}{|(G_d \cap G) \cup (P_d \cap P)|}$$

where G is the ground truth binary mask, P is the prediction binary mask, G_d and P_d are the set of pixels in the boundary region of the binary mask.

The PA is the ratio of the number of correctly predicted pixels to the total number of pixels, which is calculated as:

$$PA = \frac{\sum_{i=0}^{k} p_{ii}}{\sum_{i=0}^{k} \sum_{j=0}^{k} p_{ij}}$$

where k is the number of classifications in the segmentation task. This paper includes the drivable region class and non-drivable region class, where the drivable region is class 1, and the non-drivable region class is class 0.

3.3. Implementation and Training Details

We conduct experiments using PyTorch 2.0 on a personal workstation with two NVIDIA GTX 2080Ti. To fully exploit the expressive ability of the model, we use a series of data enhancements, including mirror flip, top-down flip, z-axis flip, and random scaling in the range of [0.5, 2]. We used AdamW as the optimizer for the training strategy, setting the learning rate to 0.0001.

4. Results and Discussion

4.1. Experimental Results and Comparison

4.1.1. Detection Results on the Dataset

We compare our method with DeepLabV3+ [26] and HRNet+OCR [29], which are state-of-the-art methods. DeepLabV3+ is used as the baseline for quantification. The recognition performance of different algorithms for agricultural semi-structured roads is shown in Table 1. The *mIoU* of our method is 97.85%, and the *Boundary IoU* is 90.88%. Compared with DeepLabV3+, the *mIoU* is improved by 2.73%, and the *Boundary IoU* is improved by 5.52%. Compared with HRNet+OCR, the *mIoU* is improved by 0.72%, and the *Boundary IoU* is improved by 2.92%. This result shows that our method has good performance on the dataset.

Table 1. Segmentation results of different methods for agricultural semi-structured roads.

Method	***mIoU*** **(%)**	***Boundary IoU*** **(%)**	**Comparison of *mIoU*** **(%)**	**Comparison of *Boundary IoU*** **(%)**
DeepLabV3+ (Baseline)	95.12	85.36	0	0
HRNet+OCR	97.13	87.96	2.01%	2.6
Ours	97.85	90.88	2.73%	5.52

4.1.2. Detection Results in Agricultural Scenarios

To further test the practical performance of the method, we conduct real-time semi-structured road region detection experiments on the farm. Test sites include indoor and outdoor sites. The indoor test site is a corridor inside a greenhouse. Lush crops are planted around the corridors that obscure the borders of the road. Meanwhile, crops block the sun, leaving dappled shadows on the road. The outdoor experimental site is a cement road in the lawn outside the greenhouse. Lawn grows in clutter, with no clear boundaries between roads and lawns. Figure 4 visually presents the segmentation effect of different methods on agricultural semi-structured roads. Although DeepLabV3+, HRNet+OCR and our method can detect drivable regions, the detection accuracy is different. The detection performance of our method outperforms the other two methods both indoors and outdoors. Especially in the presence of disturbances, our method is more robust. Specifically, in scene 1, a crack on the road surface affects the detection accuracy of DeeplabV3+ and HRNet+OCR. In scene 2, DeeplabV3+ and HRNet+OCR cannot segment distant roads because the grass and road features are similar. The large amount of shadows in scene 3 destroys the visual characteristics of the road, making the traditional segmentation method invalid. In contrast, our method can achieve robust detection of semi-structured roads in complex scenes. The field test results are consistent with the results on the dataset.

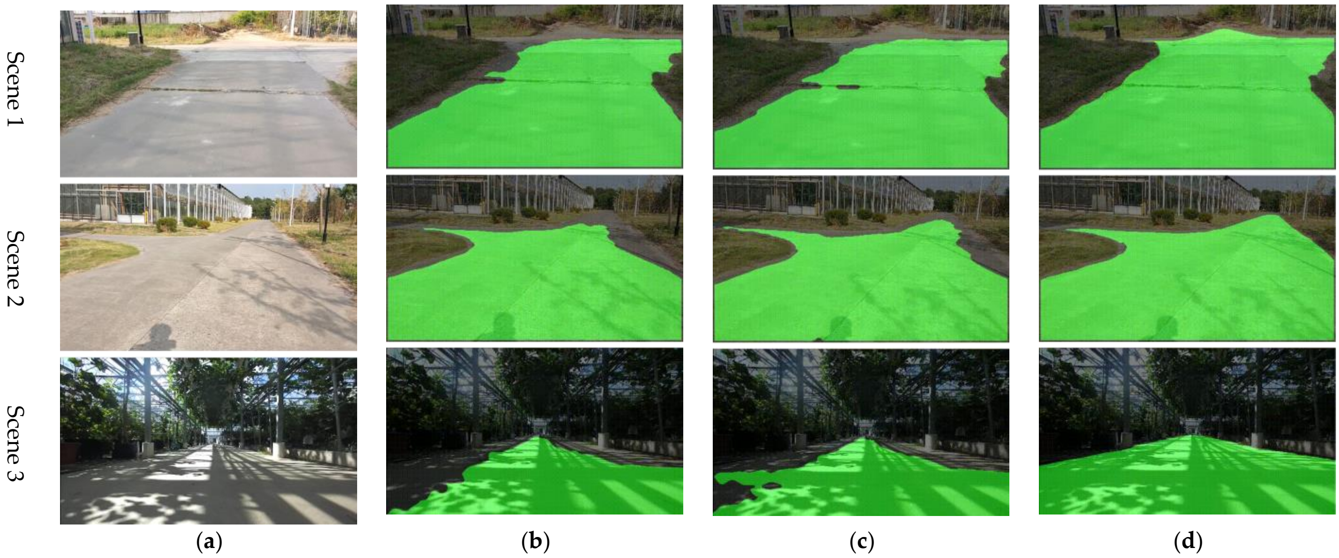

Figure 4. Examples of segmentation results for agricultural semi-structured roads. (**a**) shows the input image; (**b**) shows the result by the DeepLab V3+ (baseline); (**c**) shows the result by HRNet+OCR; (**d**) shows the result by this thesis.

4.2. Discussion

4.2.1. The Effect of Threshold Map Supervision on Road Detection

We first discuss the effect of adding threshold map supervision in the DB module on road detection performance. We compare road detection accuracy with and without threshold map supervision on the dataset. The experimental results are shown in Table 2. The results demonstrate that using threshold map supervision, the *mIoU* of the method is improved by 0.48%, the *PA* is improved by 0.17%, and the *Boundary IoU* is improved by 2.17%. Supervising the threshold map can significantly improve the quality of boundary detection.

Table 2. Comparison results of whether the threshold map is supervised.

Method	*mIou* (%)	*PA* (%)	*Boundary IoU* (%)
Unsupervised	97.16	98.64	88.71
Supervised	97.64	98.81	90.88
Comparison of results	0.48	0.17	2.17

Figure 5 shows the boundary quality improvement by the threshold map supervision. Compared with the unsupervised case, the threshold map with supervision can learn more accurate road boundaries. Especially when there are disturbances at the boundaries, the threshold map without supervision easily identifies disturbances as the road boundaries. For example, Figure 5b misidentifies water spots and steps as road boundaries, which damages the segmentation accuracy of the road in Figure 5c. In contrast, the threshold map with supervision can learn correct boundary thresholds on similar image features, resulting in sharp boundaries.

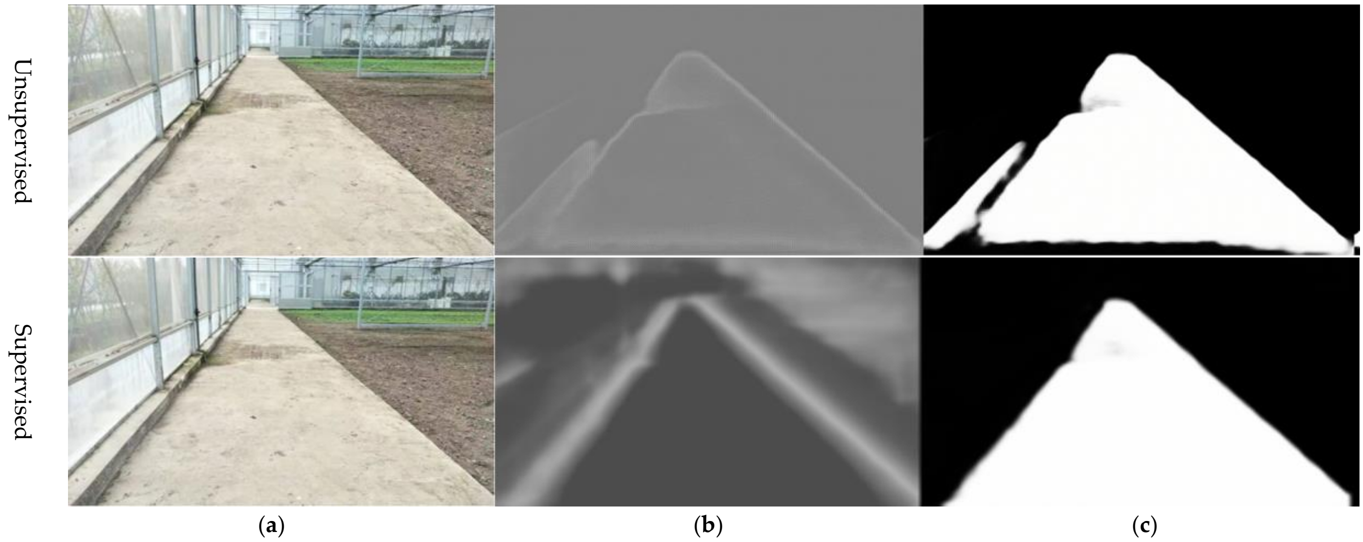

Figure 5. Comparison of segmentation results with threshold map supervision. (**a**) shows the input image; (**b**) shows the result by the boundary threshold; (**c**) shows the result by the segmentation result.

4.2.2. Comparative Analysis of Different Loss Functions

Based on threshold supervision, we compare loss functions that balance between positive and negative samples as well as easy and hard samples. In supervising coarse features, this study utilizes the BCE loss function rather than the hard negative mining method. This ensures that coarse features retain a normal distribution, allowing the entire OCR module to play a role in feature fine-tuning. For the probability map and the approximate binary map, we employ the hard negative mining approach.

Table 3 shows the impact of different hard negative mining methods on the network accuracy. The data show that the performance of WCE is superior to that of BCE. This is attributed to the fact that WCE balances the loss between positive and negative samples by assigning weights based on the dataset's statistical distribution of these samples. Meanwhile, focal loss achieves better results due to its consideration of balancing between easy and hard samples. Pixels in the boundary regions are given higher weights during the training process. In addition, OHEM delivers the best performance. It effectively balances between easy and hard samples by sorting and filtering the scores of the negative samples. Compared to the baseline model, there was an improvement of 0.63% in *Road IoU* and a 1.56% increase in *Boundary IoU*. This indicates its efficacy on facility agricultural roads with ambiguous boundaries, enhancing the quality of road boundary segmentation. This is crucial for subsequent tasks such as edge localization [36,37].

Table 3. The effect of different difficult sample mining methods on accuracy.

Loss Function	*Road IoU* (%)	*Boundary IoU* (%)
BCE (Baseline)	97.22	89.32
WCE	97.38	89.60
Focal loss	97.50	89.92
OHEM	97.85	90.88

4.2.3. Discussion of Different Deployment Methods

We tested the performance of different deployment methods on the dataset. These deployment methods are single-precision floating point (FP32), half-precision floating point (FP16), and data enhancements (including mirror flip, top-down flip, z-axis flip, and random scaling). Table 4 shows the results obtained on the dataset with different deployment methods.

Table 4. Experimental results of different deployment methods.

Deployment Method	*mIoU* (%)	*PA* (%)	*Boundary IoU* (%)
FP32 (Baseline)	98.73	98.81	90.88
FP16	98.72	98.80	90.74
Data enhancements	97.78	98.87	91.30

When the network parameters use the half-precision floating point, the *mIoU* is reduced by 0.01% compared with using the single-precision floating point, and the *Boundary IoU* is only reduced by 0.14%. This means that using the half-precision floating point for network parameters does not affect road detection accuracy, while reducing inference time by nearly 33%. Therefore, selecting the half-precision floating point for network parameters is recommended during training. Using data enhancements can slightly improve the accuracy of road detection, but it will increase the training time.

5. Conclusions

This paper proposes a high-resolution detection network for semi-structured road segmentation for autonomous navigation of agricultural robots. Since the ambiguous road boundaries cannot be semantically identified with the contrast against their neighborhoods, the pixel-level contextual representations are used at different scales to augment the road objectness. To extract high-resolution features of semi-structured roads, the HRNet is adopted as the backbone to preserve high-resolution features during downsampling. To distinguish the features of drivable and non-drivable regions more clearly, we can use a relational context-based OCR module to enhance the feature representations. Finally, an adaptive threshold-learned DB decision head is used to segment road boundaries.

To quantify the performance of our method, we used an agricultural semi-structured road dataset and conducted experiments. The experimental results show that the *mIoU* reaches 97.85%, and the *Boundary IoU* achieves 90.88%. The segmentation accuracy and boundary quality are better than the existing methods, which can meet the needs for the accurate detection of semi-structured drivable areas.

Author Contributions: Conceptualization, Y.S., L.G. and W.Z.; data curation, L.G. and W.Z.; formal analysis, C.L.; funding acquisition, L.G.; investigation, L.G.; methodology, Y.S., B.G. and Y.L.; project administration, L.G.; software, Y.S. and W.Z.; supervision, C.L.; validation, B.G.; visualization, Y.S. and B.G.; writing—original draft, Y.S., L.G. and W.Z.; writing—review and editing, Y.S. and L.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Shanghai Agriculture Applied Technology Development Program, China (Grant No. T2022-03-01) and Shanghai Science and Technology Committee (STCSM) Science and Technology Innovation Program (Grant No. 20392000900).

Institutional Review Board Statement: The study did not require ethical approval.

Data Availability Statement: The data cannot be made publicly available upon publication because they are owned by a third party and the terms of use prevent public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

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

References

1. Zhang, W.; Gong, L.; Sun, Y.; Gao, B.; Yu, C.; Liu, C. Precise visual positioning of agricultural mobile robots with a fiducial marker reprojection approach. *Meas. Sci. Technol.* **2023**, *34*, 115110. [CrossRef]
2. Bechar, A.; Vigneault, C. Agricultural robots for field operations: Concepts and components. *Biosyst. Eng.* **2016**, *149*, 94–111. [CrossRef]
3. Qi, N.; Yang, X.; Chuanxiang, L.; Lu, R.; He, C.; Cao, L. Unstructured Road Detection via Combining the Model-based and Feature-based Methods. *IET Intell. Transp. Syst.* **2019**, *13*, 1533–1544. [CrossRef]
4. Xiao, L.; Dai, B.; Liu, D.; Zhao, D.; Wu, T. Monocular Road Detection Using Structured Random Forest. *Int. J. Adv. Robot. Syst.* **2016**, *13*, 101. [CrossRef]

5. Yang, G.; Wang, Y.; Yang, J.; Lu, Z. Fast and Robust Vanishing Point Detection Using Contourlet Texture Detector for Unstructured Road. *IEEE Access* **2019**, *7*, 139358–139367. [CrossRef]
6. Shi, J.; Wang, J.; Fu, F. Fast and Robust Vanishing Point Detection for Unstructured Road Following. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 970–979. [CrossRef]
7. Liu, Y.B.; Zeng, M.; Meng, Q.H. Unstructured Road Vanishing Point Detection Using Convolutional Neural Networks and Heatmap Regression. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–8. [CrossRef]
8. Hernandez, D.E.; Blumenthal, S.; Prassler, E.; Bo, S.; Haojie, Z. Vision-based road boundary tracking system for unstructured roads. In Proceedings of the 2017 IEEE International Conference on Unmanned Systems (ICUS), Beijing, China, 27–29 October 2017; pp. 66–71.
9. Liu, G.; Worgotter, F.; Markelic, I. Stochastic Lane Shape Estimation Using Local Image Descriptors. *IEEE Trans. Intell. Transp. Syst.* **2013**, *14*, 13–21. [CrossRef]
10. Perng, J.W.; Hsu, Y.W.; Yang, Y.Z.; Chen, C.Y.; Yin, T.K. Development of an embedded road boundary detection system based on deep learning. *Image Vis. Comput.* **2020**, *100*, 103935. [CrossRef]
11. Cao, J.; Song, C.; Song, S.; Xiao, F.; Peng, S. Lane Detection Algorithm for Intelligent Vehicles in Complex Road Conditions and Dynamic Environments. *Sensors* **2019**, *19*, 3166. [CrossRef]
12. Wang, K.; Huang, Z.; Zhong, Z. Algorithm for urban road detection based on uncertain Bezier deformable template. *Jixie Gongcheng Xuebao Chin. J. Mech. Eng.* **2013**, *49*, 143–150. [CrossRef]
13. Yuan, Y.; Jiang, Z.; Wang, Q. Video-based road detection via online structural learning. *Neurocomputing* **2015**, *168*, 336–347. [CrossRef]
14. Xiang, W.; Juan, Z.; Zhijun, F. Unstructured road detection based on contour selection. In Proceedings of the 4th International Conference on Smart and Sustainable City (ICSSC 2017), Shanghai, China, 5–6 June 2017; pp. 1–5.
15. Li, J.; Liu, C. Research on Unstructured Road Boundary Detection. In Proceedings of the 2021 IEEE International Conference on Unmanned Systems (ICUS), Beijing, China, 15–17 October 2021; pp. 614–617.
16. Alam, A.; Singh, L.; Jaffery, Z.A.; Verma, Y.K.; Diwakar, M. Distance-based confidence generation and aggregation of classifier for unstructured road detection. *J. King Saud Univ.—Comput. Inf. Sci.* **2022**, *34*, 8727–8738. [CrossRef]
17. Sturgess, P.; Alahari, K.; Ladicky, L.; Torr, P. Combining Appearance and Structure from Motion Features for Road Scene Understanding. In Proceedings of the British Machine Vision Conference, BMVC 2009, London, UK, 7–10 September 2009.
18. Baheti, B.; Innani, S.; Gajre, S.; Talbar, S. Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 1473–1481.
19. Wang, Q.; Fang, J.; Yuan, Y. Adaptive road detection via context-aware label transfer. *Neurocomputing* **2015**, *158*, 174–183. [CrossRef]
20. Geng, L.; Sun, J.; Xiao, Z.; Zhang, F.; Wu, J. Combining CNN and MRF for road detection. *Comput. Electr. Eng.* **2018**, *70*, 895–903. [CrossRef]
21. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18, 2015. pp. 234–241.
22. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
23. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef]
24. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv* **2014**, arXiv:1412.7062.
25. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
26. 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; pp. 801–818, 833–851.
27. Li, J.; Jiang, F.; Yang, J.; Kong, B.; Gogate, M.; Dashtipour, K.; Hussain, A. Lane-DeepLab: Lane semantic segmentation in automatic driving scenarios for high-definition maps. *Neurocomputing* **2021**, *465*, 15–25. [CrossRef]
28. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. [CrossRef]
29. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part VI 16, 2020. pp. 173–190.
30. Liao, M.; Zou, Z.; Wan, Z.; Yao, C.; Bai, X. Real-Time Scene Text Detection With Differentiable Binarization and Adaptive Scale Fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 919–931. [CrossRef]
31. 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*. [CrossRef]

32. Hao, Y.; Liu, Y.; Wu, Z.; Han, L.; Chen, Y.; Chen, G.; Chu, L.; Tang, S.; Yu, Z.; Chen, Z. Edgeflow: Achieving practical interactive segmentation with edge-guided flow. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 1551–1560.
33. Rateke, T.; Justen, K.A.; Von Wangenheim, A. Road surface classification with images captured from low-cost camera-road traversing knowledge (rtk) dataset. *Rev. Inf. Teórica Apl.* **2019**, *26*, 50–64. [CrossRef]
34. Vatti, B.R. A generic solution to polygon clipping. *Commun. ACM* **1992**, *35*, 56–63. [CrossRef]
35. Cheng, B.; Girshick, R.; Dollár, P.; Berg, A.C.; Kirillov, A. Boundary IoU: Improving object-centric image segmentation evaluation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 15334–15342.
36. Sun, Y. A Comparative Study on the Monte Carlo Localization and the Odometry Localization. In Proceedings of the 2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA), Changchun, China, 25–27 February 2022; pp. 1074–1077.
37. Zhang, W.; Gong, L.; Huang, S.; Wu, S.; Liu, C. Factor graph-based high-precision visual positioning for agricultural robots with fiducial markers. *Comput. Electron. Agric.* **2022**, *201*, 107295. [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.