ORIGINAL RESEARCH Open access

Weakly Supervised Weed Instance Segmentation in Paddy Fields Using Point Annotations and Adaptive Pyramid Mask Generation

Noah Williams[1*], Grace Collins[1], Ethan Brooks[2], Daniel Green[3]

Abstract

In paddy cultivation, yield and grain quality declines are largely attributable to weed damage. A prerequisite for achieving automated, precision weed removal in rice paddies is accurate, rapid weed identification. Recently, Vision Transformers (ViTs) have outperformed convolutional neural network (CNN)-based architectures across various computer vision tasks. Nonetheless, the full exploitation of deep learning models for weed identification is constrained by a shortage of comprehensively annotated weed datasets. To overcome these obstacles, the present work introduces a custom point-supervised instance segmentation network, termed PIS-Net, designed for the weakly supervised instance segmentation of paddy weeds. Concretely, a novel instance segmentation point-annotation approach is proposed, in which randomly generated annotation points are placed within each instance to reduce both annotation time and complexity. Moreover, to harness point labels most effectively, a mask generation technique that relies on adaptive pyramid-level selection is proposed. Accordingly, the network is empowered to autonomously select the pyramid level most likely to yield the optimal instance mask based on its own reliability score. Finally, a pseudo-label refinement network, named PLR-Net, is constructed to polish coarse instance masks. By employing only 13 randomly generated annotation points per instance, the presented PIS-Net still achieves an AP of 38.5 and an AP50 of 68.3, surpassing the baseline mask R-CNN, which records an AP of 8.2 and an AP50 of 6.9, and thereby attaining 90% of the performance achieved under full supervision. This methodology capitalizes on point labels, rendered with high annotation efficiency, as a potent weak supervision cue to surmount difficulties in weed data annotation and the subpar accuracy of prevailing weakly supervised models. Experimental evidence confirms that the point-annotation framework of PIS-Net outpaces full-object mask annotation in speed while delivering higher AP than the existing semi-supervised weed segmentation model, indicating strong applicability in real-world paddy environments.

Keywords Deep learning, Instance segmentation, Weakly supervised, Weed recognition

*Correspondence:
Noah Williams
noah.williams@sydney.edu.au
[1] Department of Artificial Intelligence for Cotton Boll and Fiber Quality Prediction, Faculty of Agriculture, University of Sydney, Sydney, Australia
[2] Department of Precision Spraying and Drone Application Models, Faculty of Agriculture, University of Queensland, Brisbane, Australia
[3] Department of Machine Learning for Harvest Timing, Faculty of Agriculture, University of New South Wales, Sydney, Australia

Introduction

As China's primary staple food crop, rice faces formidable weed-related challenges in the field. Paddy weeds possess a tremendous reproductive capacity and impose both immediate and secondary detrimental effects on rice output [1]. Their aggressive extraction of nutrients, their part in fostering crop ailments, and their role in harboring pests and diseases

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

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

collectively drive down yields [2]. Beyond this, weed infestations in rice fields are a chief culprit in the degradation of rice quality and the impoverishment of cropland fertility. Sound weed management thus represents a cornerstone practice for rice cultivation.

Conventional strategies for weed suppression encompass mechanical cultivation and chemical application [3, 4]. Without direction from intelligent mechanisms, mechanical weeding relies solely on differences in root structure between the crop and the unwanted plants, often resulting in higher crop injury rates [5]. Traditional broadcast chemical spraying makes no distinction between the crop and the weeds, thereby raising concerns about chemical residue contamination and excessive waste. Hence, there is a pressing need for high-fidelity automated discrimination between weeds and crops [6, 7]. Such capability is indispensable for reducing herbicide volumes while maintaining weed suppression efficacy and enabling precise, site-specific weed intervention.

During earlier phases, classical computer vision methodologies reliant on hand-engineered features, encompassing plant color [8], textural patterns [9], and morphological characteristics [10], found broad application in weed recognition and detection tasks. However, these conventional computer vision pipelines require manual image preprocessing and are highly dependent on expert human judgment, introducing a degree of subjectivity. These techniques are likewise tedious, demanding considerable manual workloads [11]. They further demonstrate frail robustness in complex background settings and are unsuitable for the heterogeneous, complex conditions characteristic of rice paddies. The development of an efficient, automated, and resilient algorithm is thus imperative for coping with such multifaceted and intricate situations.

Deep learning architectures, notably convolutional neural networks [12-14], when paired with Transformer-based designs, have consistently delivered state-of-the-art results on computer vision benchmarks compared to earlier machine vision paradigms [15]. Galvanized by this progress, a growing body of investigators has proposed weed-detection frameworks built on deep learning vision models. Ma *et al.* [16] engineered 101-dimensional feature vectors capturing color, shape, and texture from images of paddy weeds. They deployed a deep belief network with multi-feature fusion, achieving recognition accuracy of 91.13% across six prevalent paddy weed varieties. Huang *et al.* [17] adopted an FCN architecture for segmenting high-resolution imagery acquired by unmanned aerial vehicles (UAVs). The technique achieved 91.96% accuracy and an average Intersection over Union (IoU) of 84.73%. Peng *et al.* [18] designed a Retinanet-derived WeedDet model, achieving a high mAP of 94.1% at 24.3 fps on a paddy weed dataset spanning 8 weed categories. The effective training of deep-learning-driven computer vision models demands abundant, completely annotated weed data. Given that agricultural production cycles are distinctly seasonal, collecting substantial volumes of annotated paddy weed data imposes substantial temporal and labor costs. Accordingly, weakly supervised instance segmentation (WSIS) and semi-supervised instance segmentation (SSIS) paradigms have steadily gained traction as means to reduce annotation overhead.

In weed detection, the scarcity of well-annotated datasets acts as a bottleneck, limiting the deployment potential of deep learning models. To this end, and to mitigate annotation costs and the reliance on exhaustive labeling, weed detection strategies built on weak supervision (WSIS) and semi-supervision (SSIS) have attracted considerable interest. One branch of SSIS revolves around a semi-supervised teacher architecture, ordinarily trained on a modest set of labeled examples. This framework leverages the teacher to generate pseudo-labels, after which a student model is trained on a blended dataset that merges genuinely labeled and pseudo-labeled samples. Picon *et al.* [19] unveiled a semi-supervised dual classification and segmentation network designed to reduce manual labeling requirements while improving weed segmentation fidelity. The system yielded an average balanced accuracy of 75.71% in distinguishing seven weed species. Nong *et al.* [20] advanced a semi-supervised semantic segmentation model called Semi-Weed-Net. By integrating Selective Kernel Attention (SKA) with encoded representations, the model highlights defining attributes of target crops and weeds, thereby mitigating recognition degradation caused by inter-class resemblance and physical overlap between crops and weeds. The addition of Online Hard Example Mining (OHEM) further optimizes training with labeled data, enabling more effective detection of small-statured crops and weeds. The injection of auxiliary consistency constraints maximizes the information extracted from a vast pool of unlabeled imagery, yielding an Intersection over Union (IoU) of 69.8% on the Weed Map datasets.

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

Another stream of WSIS is built on weakly supervised learning models that use image-level annotations or bounding boxes. Such models provide coarse supervision for pixel-level prediction tasks via bounding boxes. In contrast, off-the-shelf proposal mechanisms such as MCG and GrabCut are used to construct pseudo-instance masks. Qiang *et al.* [21] harnessed weakly supervised learning grounded on box-level annotations and Excess Green (ExG) color similarity. They embedded a weakly supervised instance segmentation algorithm, rooted in deep convolutional neural networks, into the arena of plant phenotyping. They additionally designed a weed-filtering algorithm predicated on area thresholding, k-means clustering, and temporal contextual constraints to suppress weeds within a cultivated backdrop, achieving an F1 score of 0.95 and an sMOTSA of 76.3. Thomas *et al.* [22] formulated a rule-based approach to generate labeled data targeting perennial weed species. With fully annotated datasets comprising only 10%, the average IOU on the weed datasets using a U-Net architecture was 0.7180.

Even though the above SSIS and WSIS techniques successfully reduce labeling costs, they concurrently lead to a marked deterioration in detection precision. This arises chiefly from the disproportionate ratio of instances lacking markers (i.e., false negatives) to those with marker noise (i.e., false positives) in the weed annotations, which, in turn, compromises the fidelity of the resultant pseudo-labels.

Bearing in mind the shortcomings of earlier investigations, we are exploring the use of point-based supervisory training regimens to address persistent challenges in weakly supervised instance segmentation for paddy weeds. Chief among these difficulties are imprecise segmentation boundaries and substandard model accuracy. In existing weakly and semi-supervised weed segmentation frameworks, the precision of the segmentation mask is predominantly governed by the extraction of instance proposals. The segmentation network's overall capability hinges critically on the validity of the proposals emerging from the instance proposal branch. Simultaneously, when only a sparse, fully annotated dataset is available, the network often fails to produce adequate instance masks in the mask branch, producing rough, noise-contaminated mask outputs. Stated differently, the presence of true-positive proposals does not routinely translate into true-positive instance masks. To remedy these drawbacks, the present work proposes a weakly supervised instance segmentation model for weed segmentation guided by adaptive point-label guidance. It capitalizes on point-annotated weed data to achieve a segmentation that is both efficient and highly performant. A PLR-Net is introduced to polish the initially coarse weed instance masks, while jointly leveraging point labels as instance proposals to improve the performance of the proposal branch in the weed segmentation network. The ultimate objective is to sharpen the accuracy with which weed segmentation masks are generated.

To summarize, the contributions furnished by this study are:

(i) The conception of an innovative point-labeling scheme, for instance, segmentation. For each instance, only a single bounding box, along with a handful of randomly generated annotation points, is marked, striking a better balance between annotation speed and precision.

(ii) The formulation of a point-supervised training methodology, complemented by an adaptive pyramid level selection strategy, and the coherent fusion of point labels with sampled proposals to serve as guidance for producing weed segmentation masks.

(iii) The introduction of a PLR-Net (Pseudo Label Refinement Network) to refine coarse instance masks.

Materials and Methods

Weed datasets

A purpose-built, mixed-species weed dataset was assembled for the segmentation experiments, comprising eight commonly encountered paddy weed types and rice specimens. The assortment spanned barnyard grass (*Echinochloa crus-galli*), Lysimachia (*Eclipta prostrata*), hollow lotus seed grass (*Alternanthera philoxeroides*), redroot amaranth

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

(*Cardamine occulta* Hornem), Oxalis corniculate (*Eleusine indica*), Smartweed (*Lepidium meyenii*), as well as rice entries. Field collections were conducted at trial plots located across Zhaoqing and Jiangmen, Guangdong, China (23° 2′ 48″ N, 112° 27′ 55.1″ E; 22° 34′ 43″ N, 113° 4′ 54.8″ E).

All weed imagery originated from rice parcels that underwent natural development throughout 2023, free from any husbandry interventions. Image recording was performed on iPhone 14 Pro Max and Honor 30 mobile devices from three distinct vantage points: vertically above, straight down, 45° relative to the ground, and 60° relative to the ground. The rationale for this diversified shooting geometry was to capture the paddy weeds' hallmark visual traits comprehensively. Representative photographs of the eight weed categories appear in **Figure 1**, captured on separate dates under varying lighting conditions and against dissimilar background settings. Cumulatively, 2000 weed and rice images were stockpiled.

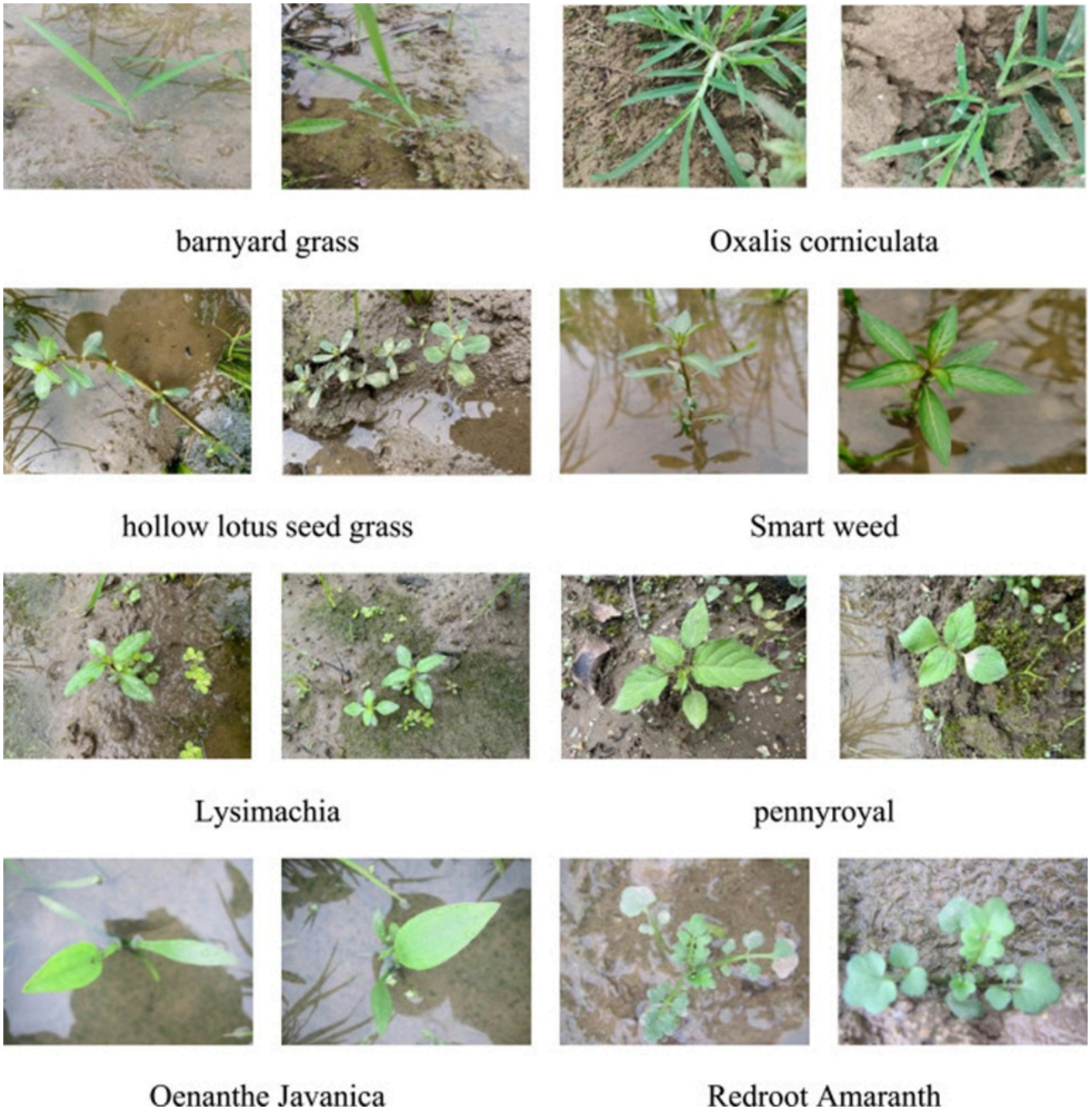


Figure 1. Rice field weed images from a self-constructed weed dataset.

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

Overview of the proposed approach

As shown in **Figure 2**, the suggested pipeline assembles point-wise feature descriptors at chosen coordinates by merging (e.g., via concatenation) two classes of representations: high-resolution detail features and coarser predictive features. The aspiration to secure trustworthy instance proposals that yield faithful mask reconstructions motivated the adoption of point labeling—a proposal paradigm that performs classification at the granularity of individual pixels. When an object is spotted, PLR-Net formulates tailor-made parameters destined for the point head. Functioning autonomously regardless of position, the point head ingests fine-resolution point features and the spatial coordinate displacement from the bounding box at that location, and from these outputs, produces dense, pixel-level masks. Furthermore, a restructuring of the Region of Interest (ROI) module within the incumbent framework was executed, as highlighted in **Figure 3**. Once the incoming feature map advances through the point head. The point intelligence is secured, and the Region of Interest (ROI) operation maps points and the pixels enclosed within the ROI to the exact coordinates of the feature map. This registration step yields raw point features that encode both spatial locations and offsets relative to the bounding box. In marked contrast to the orthodox rectangular or polygonal ROI-Align, ROI-Point-Align offers superior versatility and sharpness by capturing regions of interest as sets of scattered points or individual pixels. Producing high-quality weed segmentation masks is a pivotal prerequisite for weakly supervised instance segmentation-based weed detection pipelines. Reflecting this, point labels were designated to serve as the training objective for the weed segmentation network, fusing an adaptive, pyramid-level-driven mask construction strategy with the refinement apparatus PLR-Net. The intended outcome of this union is the generation of weed segmentation masks of elevated caliber.

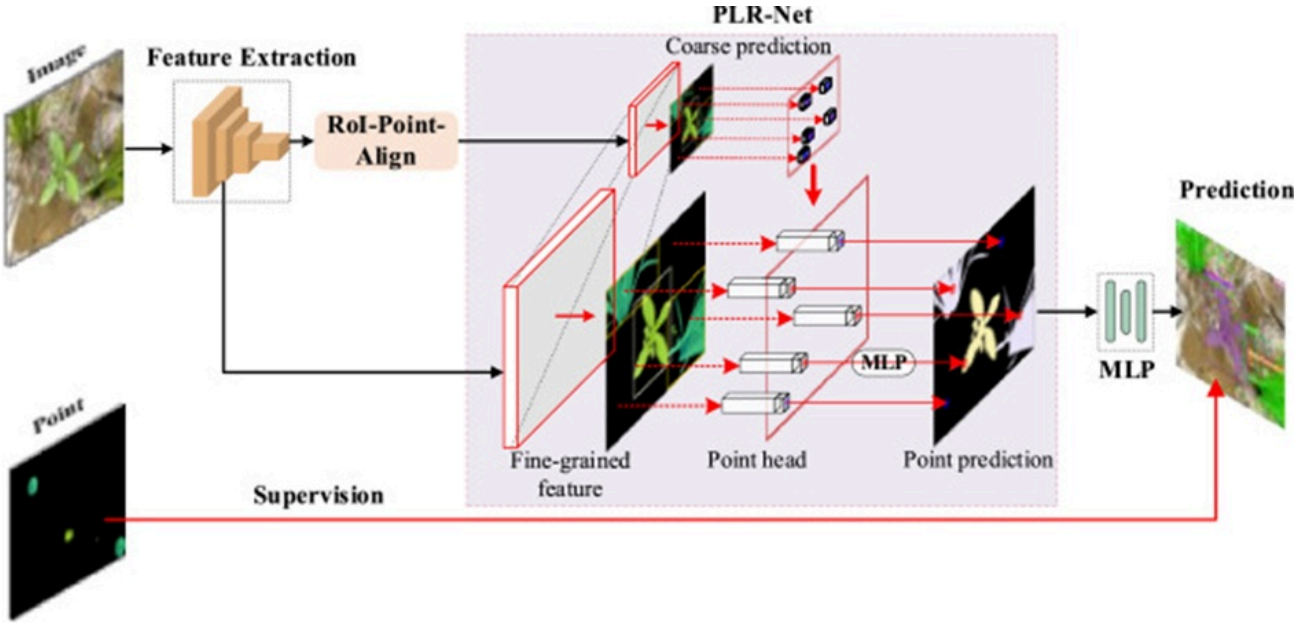


Figure 2. Overview of the proposed method. The fine-grained point representations extracted from the precise locations of random points in the image-level feature maps using bilinear interpolation distinguish PLR-Net from traditional point supervision, where only coarse mask predictions provide region-level context to differentiate objects.

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

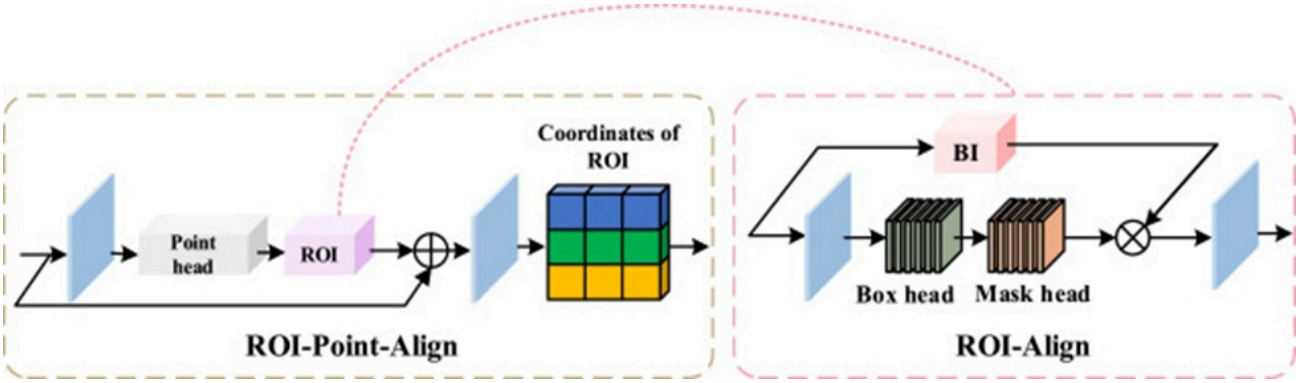


Figure 3. Architecture of the ROI-Point-Align. ⊕ indicates element-wise addition and ⊗ indicates element-wise multiplication.

Point-wise representation and point head

The annotation blueprint underpinning the novel scheme is elegantly uncomplicated. Points are stochastically drawn inside each bounding box and subsequently categorized as either object (illustrated in red) or background (illustrated in blue). Going a step beyond conventional bounding box labeling, the working assumption holds that N points situated within each object bounding box receive a dual object-background label, as visualized in **Figure 4**. Those points found inside the bounding box are collectively referred to as PN. Rather than persisting with the prior practice of manually positioned points—a tactic constrained by low variability—randomly synthesized points are elected over manually curated ones for weeds. Bearman *et al.* [23] provided evidence that training segmentation models on randomly generated ground truth points yields better results, owing to enhanced data robustness relative to models conditioned on human-clicked points. Hence, only bounding boxes are compiled for each weed instance, with annotation points randomly sampled within every object bounding box and individually classified as object or background. Concretely, bilinear interpolation is brought to bear on forecasting the exact loci of ground truth points (portrayed in red and yellow for object and background ground truth points, respectively). As depicted in **Figure 2**, PLR-Net incorporates a point-head that undertakes point-wise segmentation inference. Architecturally, this point-head boils down to an unadorned Multi-Layer Perceptron (MLP) whose learned weights are shared globally across all points (and regions), paralleling the design philosophy of graph convolutional networks [24] and PointNet [25]. Since the point-head emits a segmentation label per point, it can be optimized via customary, task-oriented segmentation losses. By replicating the annotation style embedded in our proposed framework, the annotation procedure can be made more efficient, and results can be validated on our self-built weed datasets, obviating the need for onerous data gathering exercises. This squarely addresses the bottleneck posed by the arduous task of annotating weed data.

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

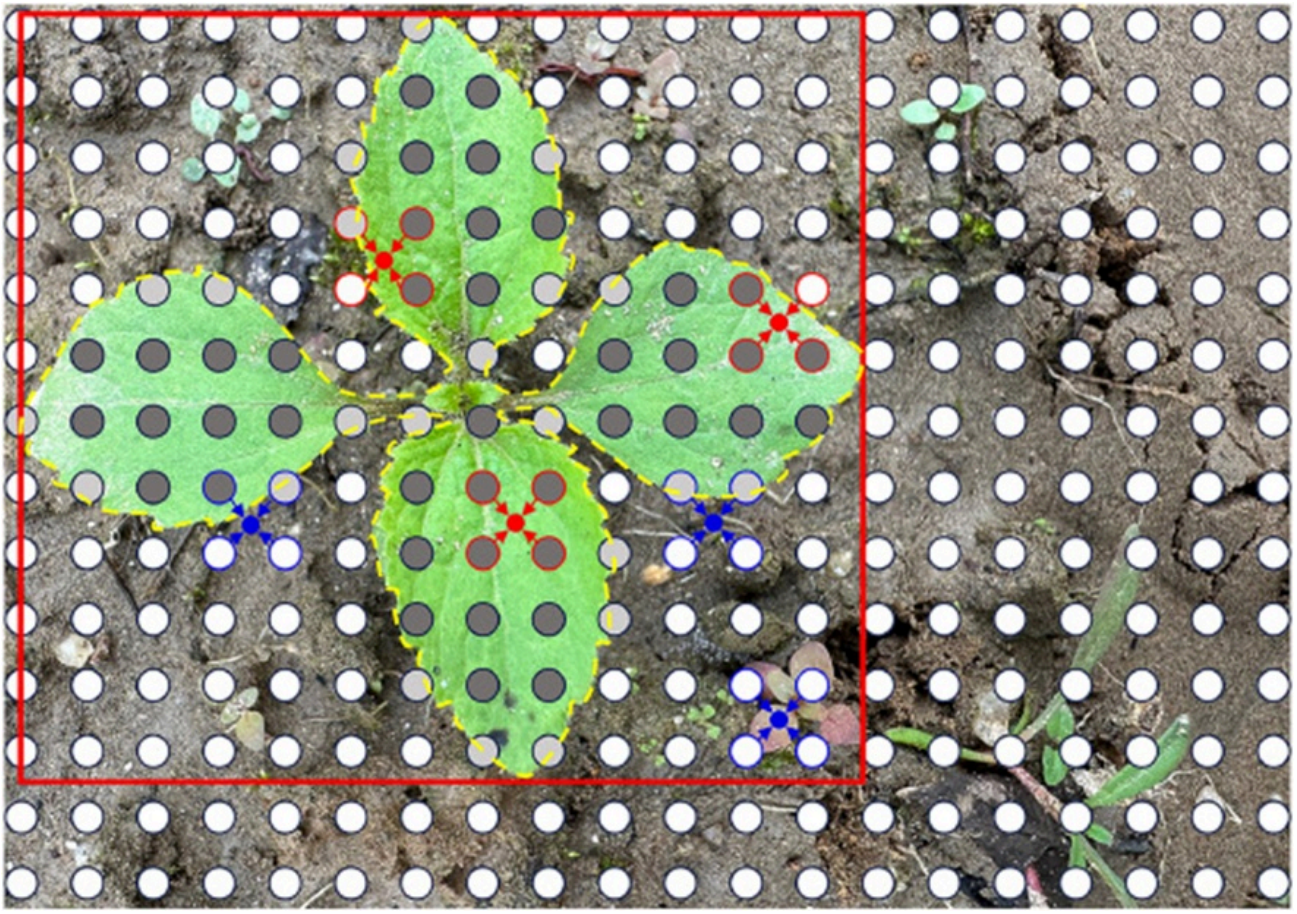


Figure 4. Illustration of point supervision. For a 12 × 17 prediction mask on the regular grid (darker color indicates foreground prediction), red and blue indicate foreground and background ground truth points, respectively. Note that the object boundary line (yellow) is only for illustration.

Weed segmentation mask generation

Seeking a well-reasoned exploitation of point labels, the present inquiry entertains the deployment of a Feature Pyramid Network (FPN [26]), assembled from a hierarchy of feature maps engineered expressly for recognizing instances across a continuum of scales. Each pyramid stratum is calibrated to respond to instances falling within particular size brackets. Drawing instance masks from mismatched strata, however, regularly produces defective weed segmentation masks. As a corrective, a strategy christened Adaptive Pyramid-Level Selection is proposed, in which the pyramid stratum forecasted to deliver the most fitting weed instance mask is automatically selected, guided by the network's self-assessed confidence. In operational terms, point label coordinates are rescaled in step with the spatial granularity of every pyramid stratum, after which confidence scores are queried across the entire set of strata. Strata registering confidence scores shy of the maximum are eliminated, and an instance mask is manufactured strictly from the one pyramid stratum commanding the topmost confidence score, as diagrammed in **Figure 5**. Expressed more formally, N proposal branches {FLi}Ni are in play, with the FPN arrangement retained at N = 5. Given any point label (x, y, z), where z denotes the weed category identifier, an instance mask generation proposal paired with a confidence score is extracted as (Pi, si) = FLi(x, y, z). That confidence score is adopted as a barometer of mask accuracy. This permits adaptive selection of a pyramid tier distinguished by the maximal confidence score k, k = argmaxk∈{1,2,…, L}sk. At the concluding mask branch Fm, a weed instance mask is synthesized as M = σ(FmPk), where σ symbolizes the sigmoid nonlinearity.

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

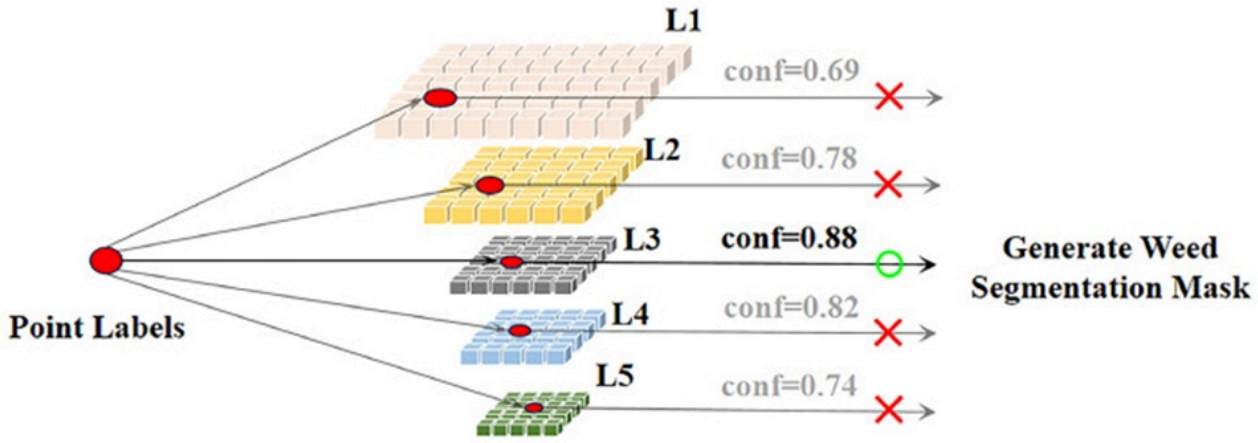


Figure 5. Adaptive Pyramid-Level Selection with FPN, generating an instance mask solely from the pyramid level with the highest confidence score.

PLR-Net

The point-based annotation convention springing from our method meshes readily with the PLR Head module. This module transforms point-wise predictions into object detection masks. For each identified weed instance, breaking from the mold of traditional coarse mask estimation, the PLR Head deduces the parameterization of a point-wise function. That function computes per-point target-mask estimates at any queried position, along with the associated image features. Such a construction equips the model to trace instance boundaries and subtle morphological details with greater verisimilitude. In the interest of computational tractability, the PLR Head module settles on a single-point level mask loss, expressed as the cross-entropy penalty summed over all pixels in conjunction with the Dice Loss:

$$Diceloss = 1 - \frac{2*\sum_{i=1}^{N} Intersection_i}{\sum_{i=1}^{N} Cardinality_i + \in} \quad (1)$$

Where $\sum_{i=1}^{N}$ Intersectioni tallies the overlap magnitude between the predicted segmentation output and the ground truth segmentation output, $\sum_{i=1}^{N}$ Cardinality represents the aggregate extent of the predicted segmentation output and the true segmentation output, and N denotes the count of weed categories.

The corrective lift provided by PLR-Net stems from a tandem of explanations: (1) the prior knowledge locked within annotated bounding boxes is put to work; since PLR-Net receives as its input the coarse mask predictions emitted by the segmentation network, it progressively learns to remediate the characteristic blunders present in those predictions. (2) By folding in the incoming point label intelligence, PLR-Net forecasts the target mask for each point at a designated position together with a corresponding image feature. This provides sharp instance-level clues that help disentangle overlapping instances and identify pixels erroneously labeled by the segmentation network. Hence, steered by point-label signals, PLR-Net mends omissions and quells noise artifacts resident in the coarse mask, and additionally partitions individual target instances within the coarse mask relying on the furnished point labels, as portrayed in **Figure 2**.

To endow PLR-Net with the capacity to render intricate segmentation nuances, feature vectors are drawn from CNN feature maps at every sampling point. Cognizant that a point is characterized by real-valued 2D coordinates, standard convention [27] is observed by conducting bilinear interpolation over the feature maps to obtain the feature vectors. In this investigation, features are extracted from a single feature map (res2 in ResNet [28]), and fine-resolution point-wise features are aggregated at selected points. While fine-resolution features are adept at handling detail-oriented

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

challenges, they also have two salient shortcomings. In the first place, region-specific contextual awareness is missing, meaning identical fine-resolution features would be mapped to points that overlap across the bounding boxes of two separate instances. Yet, any given point may legitimately constitute foreground for only a single instance. Thus, in instance segmentation contexts—where distinct regions may assign distinct labels to the same point—supplementary region-specific context becomes necessary. In the second place, depending on which feature map is tapped for fine-resolution feature extraction, those features might encapsulate only relatively elementary information (e.g., res2 in ResNet) [29, 30]. In such scenarios, feature repositories bearing richer semantic and contextual freight could prove valuable. This consideration extends to both instance segmentation and semantic segmentation problems. Taking these facets into account, the second type of feature derives from the network's coarse segmentation predictions. By architectural design, coarse predictions supply a more panoramic, semantically informed outlook, albeit at a coarser spatial scale. The refinement progression is laid out in **Algorithm 1**. For any given region, the output mask materializes iteratively, progressing from coarse to fine granularity. Initially, predictions are formed at the coarsest resolution, using points sampled on a regular grid (i.e., with a standard coarse segmentation prediction head). In each successive round, PLR-Net up-samples the previously inferred segmentation via bilinear interpolation and subsequently homes in on the N points exhibiting the greatest uncertainty (for instance, those whose estimated probabilities dwell nearest to 0.5 in the binary-mask scenario) spread over this now-denser grid. PLR-Net then derives a point-wise feature representation for each of the N points and issues label predictions for them. This cyclical refinement persists until the segmentation attains the intended resolution.

1: **Input:** Training image samples, point labels
2: **Initialization:** Coarsest granularity and resolution definition. Coarse segmentation head for preliminary point predictions.
3: Set the iteration condition as T.
4: **while** $t = 1$ *to* T **do**
5: Upsampling: Apply bilinear interpolation to upsample the segmentation mask from the previous iteration.
6: Uncertain Point Identification: Identify N points with probabilities closest to 0.5 on the denser grid.
7: Feature Representation and Label Prediction: Compute feature representations for these points and predict their labels.
8: Mask Update: Update the segmentation mask at the current resolution based on new predictions.
9: **end while**
10:Set the resolution of the predicted mask as R.
11: **for** $R <$ target resolution **do**
12: Increase the iteration counter and repeat step 3.
13: **end for**
14: **Output:** The final high-resolution segmentation mask.

Algorithm 1. The procedure of progressive high-resolution mask generation has been refined.

A single step of this procedure is illustrated in a toy case in **Figure 6**. This comprehensive approach, combining bounding-box priors and point-label information, enhances PLR-Net's performance in segmenting paddy field weeds.

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

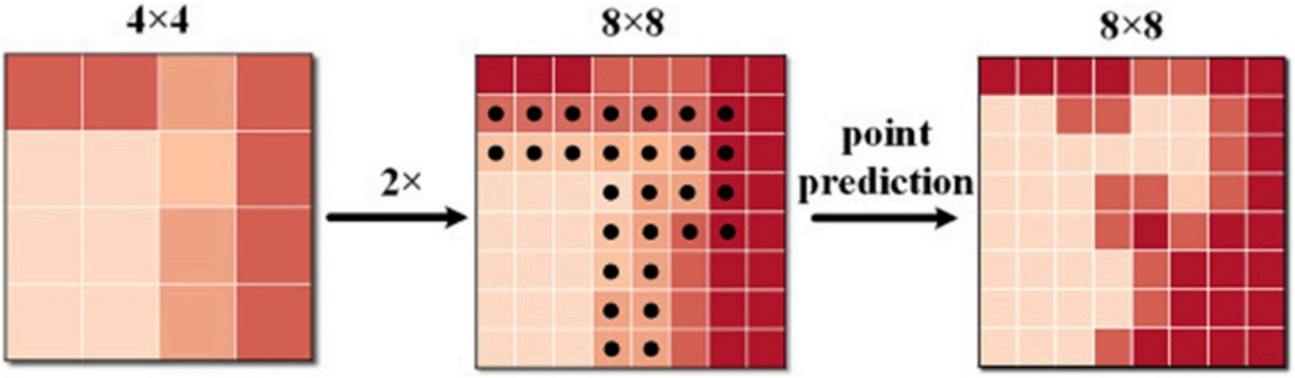


Figure 6. Illustration of an adaptive subdivision step: Predictions on a 4 × 4 grid are upsampled by a factor of 2 × using bilinear interpolation. Subsequently, PLR-Net predicts for the N most uncertain points (depicted as black dots) to restore detail on the finer grid.

Experiments

Experimental datasets construction and network training

To substantiate the effectiveness of the advanced PIS-Net point-supervised instance segmentation methodology for paddy weed segmentation, experiments were conducted on an independently curated rice-field weed dataset. The point-label annotation workflow was kept deliberately simple. For every object present in the image, once the bounding box together with the object class label had been furnished, N random point coordinates falling inside the bounding box were disclosed sequentially to the human annotator, who delivered a binary object/background verdict, as exhibited in **Figure 7**. The self-constructed weed dataset was procured from experimental rice paddies spanning Zhaoqing and Jiangmen, Guangdong, China. The dataset combined a mixture of regionally characteristic weed flora alongside the rice crop itself. **Table 1** presents an exhaustive characterization of these self-constructed datasets.

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

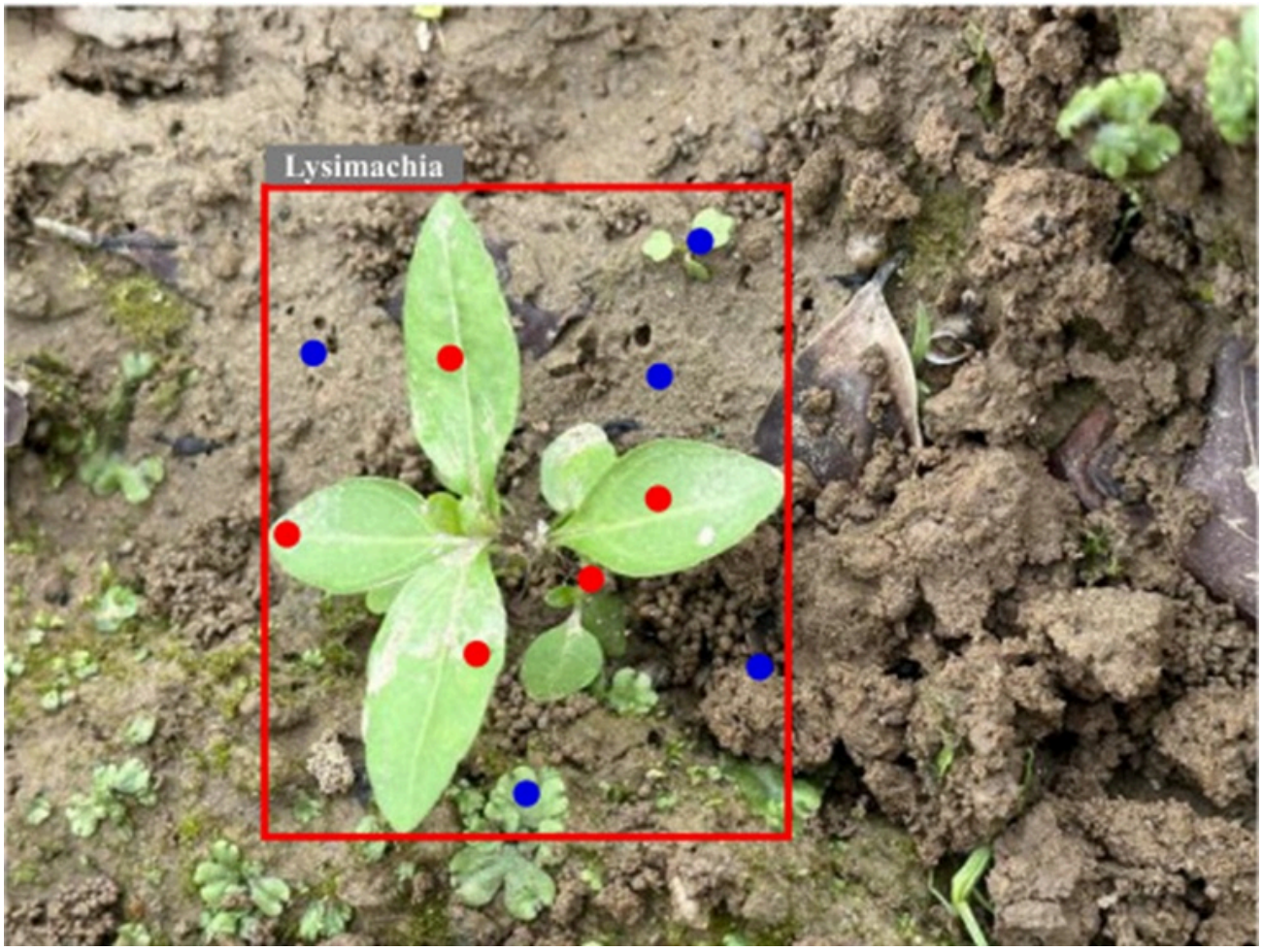


Figure 7. Example of point-based instance annotation scheme, showing a collect object bounding box with points randomly sampled inside the box and annotated as the object (red) or background (blue).

Table 1. Description of the datasets used in the experiments.

Dataset	Total Images	Split Details
Barnyard grass	881	Training: 695; Testing: 186
Hollow Lotus Seed Grass	441	Training: 353; Testing: 88
Lepidium meyenii	119	Training: 94; Testing: 25
Glehnia littoralis	167	Training: 128; Testing: 39
Oenanthe javanica	582	Training: 457; Testing: 125
Redroot Amaranth	200	Training: 158; Testing: 42
Lysimachia	538	Training: 446; Testing: 92
Oxalis corniculata	98	Training: 80; Testing: 18
Cyperaceae	441	Training: 358; Testing: 83

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

Network training parameters

The entire experimental work was conducted on a workstation equipped with an NVIDIA RTX 6000 Ada Generation/PCIe/SSE2 and an Intel(R) Xeon(R) Gold 6226 CPU @ 2.70 GHz. To safeguard the fairness of comparisons, each model was instantiated within the PyTorch ecosystem. At the same time, the input imagery dimensions, the number of training epochs, the learning rate, the momentum coefficient, and the batch size were fixed to identical values across all networks examined. The segmentation network, most prominently the PIS-Net, was optimized on a weed dataset annotated using random points. The input resolution for every network was standardized to 224 × 224 × 3. During training on the self-constructed rice field weed datasets, the epoch budget was set to 4000, the batch size was fixed at 16, momentum was set to 0.9, and the learning rate was set to 0.0002. Any remaining hyperparameters of each network were retained at their original default specifications.

Performance evaluation

Under the COCO evaluation framework, all Average Precision (AP) values were treated as equivalent to the mean Average Precision (mAP). Three widely adopted indicators were employed to quantify model capability: AP, AP50, and AP75. The AP metric was derived by sweeping across Intersection over Union (IoU) thresholds from 0.5 to 0.95, a computation that captures the algorithm's knack for balancing precision and recall across different confidence thresholds, thereby delivering a holistic picture of detection performance under varying cutoff stringencies. Precision reflected the model's fidelity in correctly identifying targets, whereas recall reflected its thoroughness in uncovering them. In the AP computation, IoU served as the standard metric for measuring the spatial overlap between predicted and reference bounding boxes. AP50 specifically measured average precision at an IoU cutoff of 50%, effectively testing the algorithm under moderately permissive overlap constraints.

Meanwhile, AP75 was typically used to assess model performance when overlap criteria were considerably stricter. Higher numerical readings across the three metrics directly translated into stronger model forecasting performance. Their mathematical specifications are as follows:

$$AP = \frac{1}{11} \sum_{r \in \{0, 0.1, 0.2, \ldots, 1.0\}} P_{\text{intep}}(r) \quad (2)$$

Results and Discussion

Comparison with other methods

To corroborate the efficacy of the introduced point-supervised instance segmentation framework, PIS-Net, for paddy weed delineation, head-to-head evaluations were conducted against contemporary convolutional neural network benchmarks, specifically BoxInst [31], PointRend [32], and Mask R-CNN [27]. For these comparisons, the same Mask R-CNN [27] backbone was used in both full-mask and point-supervision settings. For bounding-box supervision, the recently unveiled BoxInst—an instance segmentation model built atop CondInst [31]—was selected, given its status as a top performer among weakly supervised techniques that leverage bounding-box annotations on the COCO benchmarks. All models were uniformly equipped with a ResNet-50-FPN backbone [28]. A synopsis of the comparative performance figures on the self-curated rice field weed datasets is collated in **Table 2**.

Table 2. Comparison of average precision across various methods on weed datasets.

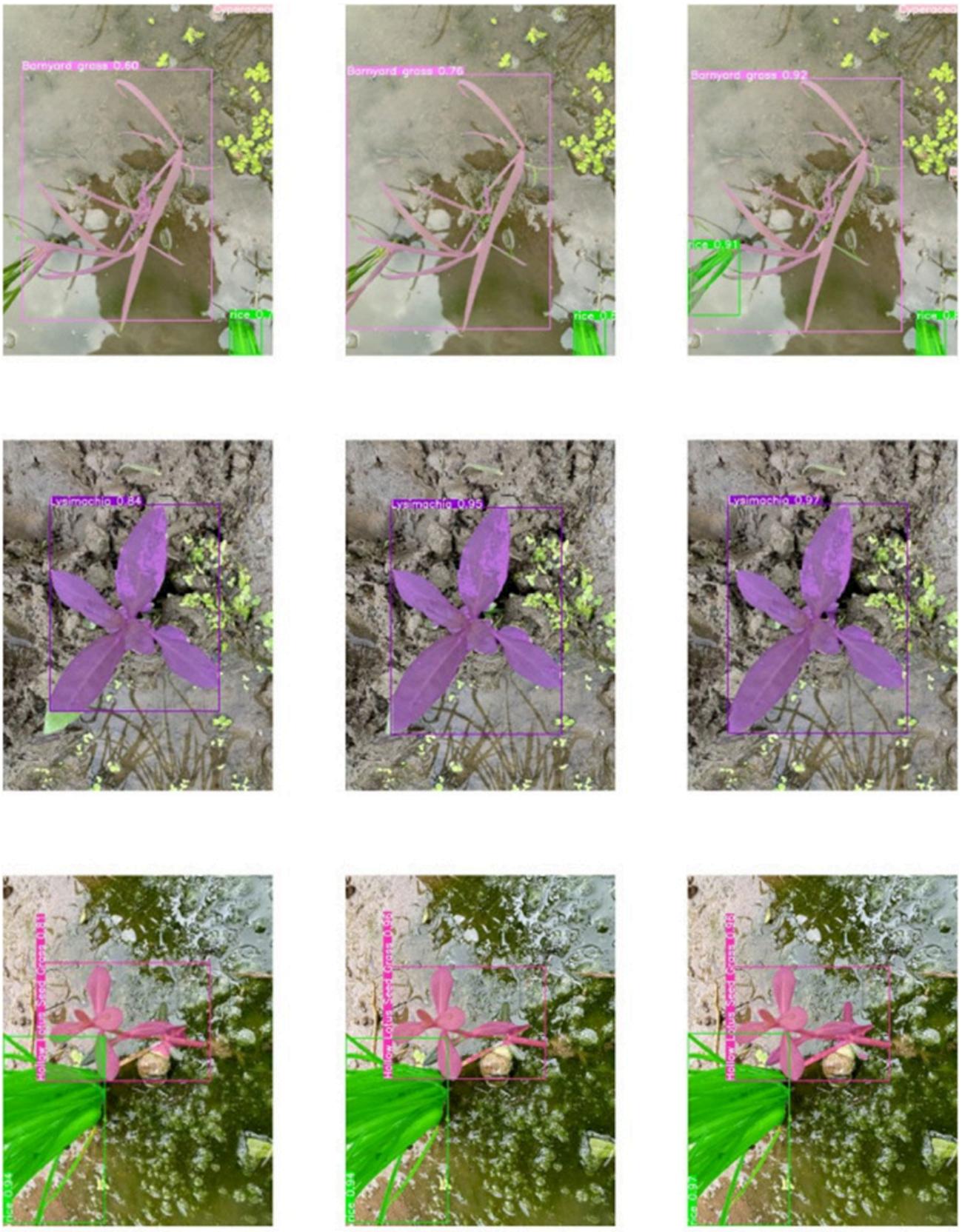
Method	Supervision	Backbone	AP75	AP50	AP
BoxInst	Box	R50	26.3	65.0	33.7

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

BoxInst	Box	R101	32.5	66.7	35.9
PointRend	Point	R50	21.6	43.6	21.9
PointRend	Point	R101	22.8	49.3	24.5
Mask-R-CNN	Point	R50	28.4	61.4	30.3
Mask-R-CNN	Point	R101	33.9	63.1	34.9
PIS-Net	Point	R50	35.8	68.2	38.5

Figure 8 presents side-by-side segmentation renderings from PointRend, BoxInst, and our own methodology. Weeds are overlooked in the outputs of **Figures 8a and 8b** were correctly delineated by our approach, as borne out in **Figure 8c**. Beyond that, in cases where PointRend and BoxInst were unable to tease apart individual objects within coarse mask predictions, our framework further enriched the mask portrayal, suggesting a superior capacity for segmenting paddy weeds relative to both PointRend and BoxInst.

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



(a) **(b)** **(c)**

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

Figure 8. Comparison of segmentation visualization between existing semi-supervised instance segmentation models and our model: (a) segmentation visualization of PointRend, (b) segmentation visualization of BoxInst, and (c) segmentation visualization of our method.

Performance of PIS-Net

Performance statistics for the proposed point-supervised instance segmentation technique, PIS-Net, evaluated against the self-curated rice field weed dataset, are tabulated in **Table 2**. Operating without architectural surgery or tampering with the ingrained training recipe and hyperparameters native to PyTorch, the PIS-Net formulation achieved an AP of 38.5%. In an equivalent setup with full supervision, Mask R-CNN achieved an AP of 40.8%. What stands out is that, when schooled exclusively on point annotations, PIS-Net scaled to 94% of its fully supervised counterpart's achievement, all while the architecture and PyTorch's default training configuration remained entirely untouched. Pitted against leading-edge point-supervised instance segmentation competitors, our methodology came out ahead (38.5% versus 24.5%).

Meanwhile, stacked against premier bounding-box-supervised methods sharing an identical backbone, our approach achieved superior segmentation scores (38.5% versus 33.7%) through an annotation-efficient pipeline that requires no more than point-level labeling. **Figure 9** exhibits prediction snapshots generated by the point-supervised PIS-Net model. The evidence pointed to a substantial easing of the weed data annotation burden thanks to the point-supervised instance segmentation paradigm (PIS-Net). Relying exclusively on point labels, it achieved high-performance weed segmentation and dramatically reduced the gap between fully supervised and point-supervised instance segmentation. This outcome spotlights the considerable promise that point-supervised instance segmentation holds for deployment in agricultural practice.

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



Figure 9. Recognition results of the proposed PIS-Net trained with 13 points per object on a self-constructed weed dataset in rice fields.

Results from Ablation

Ablation analyses targeting the proposed method were executed on the self-curated weed dataset. To ensure experimental fairness across ablations, every model was implemented using the detectron2 library within the PyTorch environment. The input image dimensionality, batch size, backbone architecture, and learning rate were kept identical across the board.

Ablation of the point number

The quantity of annotated points exerted a pronounced influence on the quality of the weakly supervised segmentation network. **Figure 10** traces the behavior of PIS-Net as a function of the number of labeled points per instance. Ramping the annotated point count from 5 up to 13 triggered a swift performance surge of roughly 3.5 AP. Yet, pushing the annotation tally even higher failed to yield commensurate performance gains; rather, a clear pattern of tapering returns emerged. Although enlisting 40 points for supervision can secure a 0.8 AP edge relative to 13 points, the supplementary temporal cost of fabricating the extra annotation points swells considerably. To broker a satisfactory compromise between

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

performance uplift and time expenditure, to assure functional viability in practical contexts, and to derive maximal utility from point labels under finite computational means, every downstream experiment was carried out on a self-curated weed dataset with 13 annotated points.

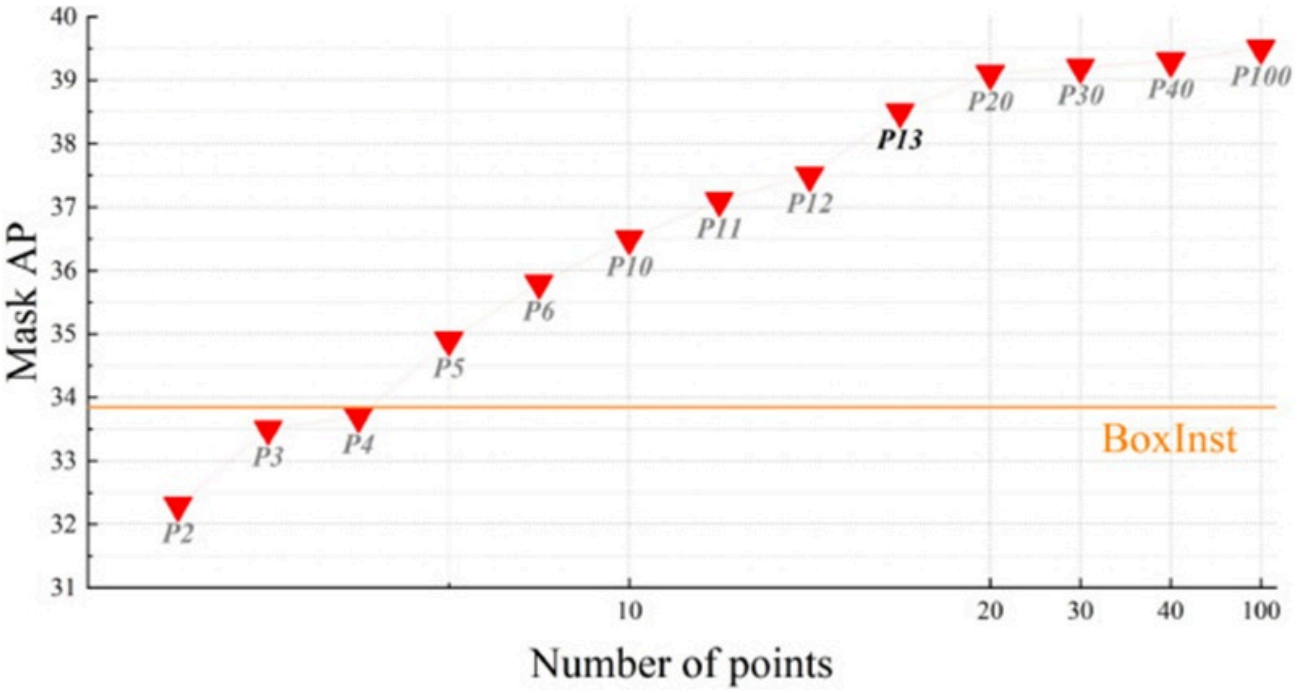


Figure 10. Training with a different number of points (P). The proposed method, trained on a self-constructed weed dataset with as few as 13 labeled points per instance (P13), achieves 38.5 mask AP with diminishing returns from more labeled points.

Ablation of the PLR-Head

The bearing of PLR-Net on segmentation network output is charted in **Figure 11**. Stripped of PLR-Net, segmentation capability on weeds languished at a meagre 30.3 AP. Bereft of the refinements afforded by PLR-Net, the network failed to capture diminutive weed targets scattered across the imagery fully. Upon grafting PLR-Net onto the architecture, weed segmentation quality vaulted dramatically to 38.508 AP. The boost is attributable to the use of annotated points as navigational hints for target instances, paving the way for more precise carving of occluded weed specimens, as rendered in **Figure 12**. Qualitative snapshots under the restriction of employing exclusively annotated point-labeled data: Annotated point intelligence was harnessed as instance-level beacons, with the PLR-Head managing to segment both minuscule weed targets and partially obscured weed instances (1st and 2nd rows) directed by the given point labels. The PLR-Head also further enriches the resultant mask representations (3rd row). On the precision dimension, our technique chalked up substantial gains across both AP50 and AP75, hitting 68.292 and 35.870, respectively, compared with Mask R-CNN, which mustered 61.412 and 28.374. This testifies to a more robust performance spanning varied Intersection over Union (IoU) cutoffs. The inference is that, when operating strictly with annotated point labels, PLR-Head emerges as an exceptionally potent lever for lifting segmentation exactitude.

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

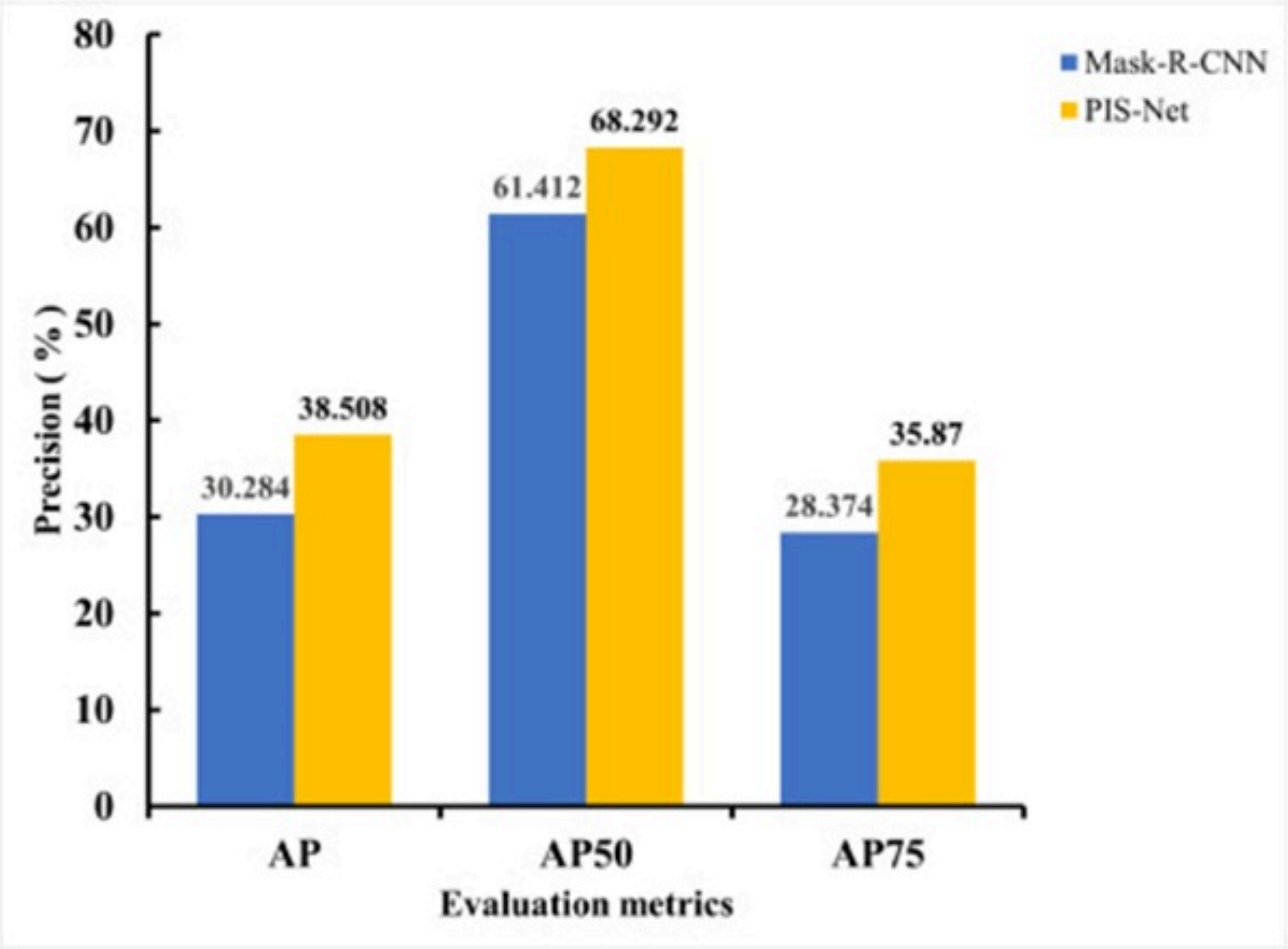
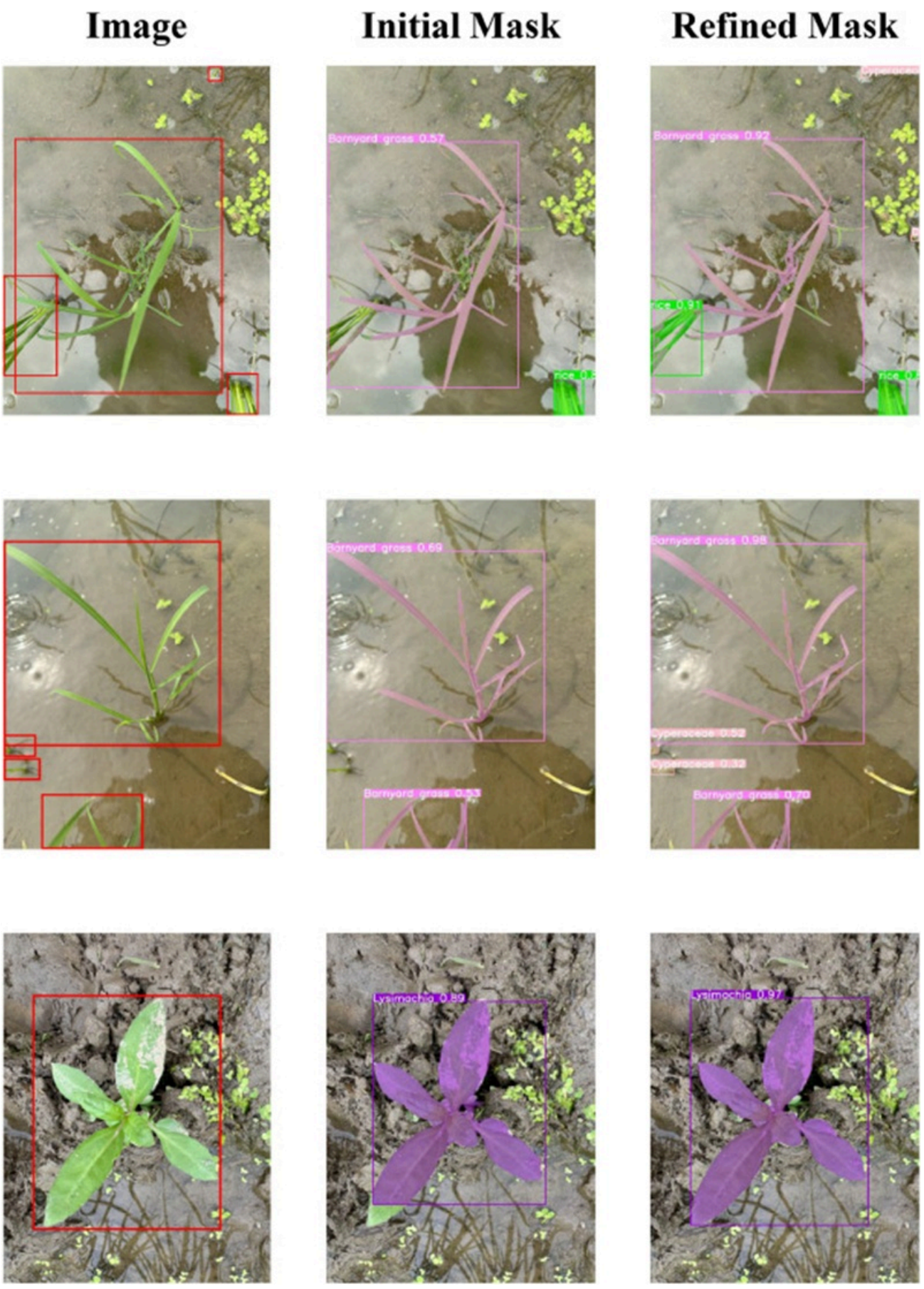


Figure 11. Average precision comparison between the original point-supervised training Mask-R-CNN model and the Mask-R-CNN model trained after adding PLR-Head.

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



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

Figure 12. Effect of PLR-Head. The qualitative results using only annotated point labelling data conditions.

Ablation of choosing features adaptively

The functioning of the FPN was dissected quantitatively in **Figure 13**. When weed masks were extracted from annotation point intelligence utilizing a lone feature map tier (L2), segmentation quality proved lackluster, scraping 27.791 AP, as demarcated by the pink swath in **Figure 13**. In contrast, synthesizing weed masks from the totality of pyramid feature map tiers (L2 ∼ L5) pushed segmentation quality to 37.830 AP. The explanation behind this jump rests in the fact that annotation point intelligence sourced from mismatched pyramid tiers injects substantial noise, degrading the fidelity of weed segmentation masks, as illustrated in **Figure 4**. Through our Adaptive Pyramid-Level Selection mechanism, a single well-matched pyramid tier was handpicked based on the network's internal confidence signal, culminating in an improved quality of 38.508 mask AP. At the 50% Intersection over Union (IoU) cutoff (AP50), the figures posted by Adaptive (68.845) and L2-L5 (68.292) stood neck and neck, with Adaptive holding a slim advantage. Moreover, in circumstances that place a premium on segmentation exactitude (AP75), Adaptive (35.870) outperformed L2-L5 (34.495) by a considerable margin. These findings confirmed that the proposed Adaptive FPN strategy demonstrated marked proficiency in leveraging the Feature Pyramid Network (FPN) to mass-produce high-quality weed segmentation masks.

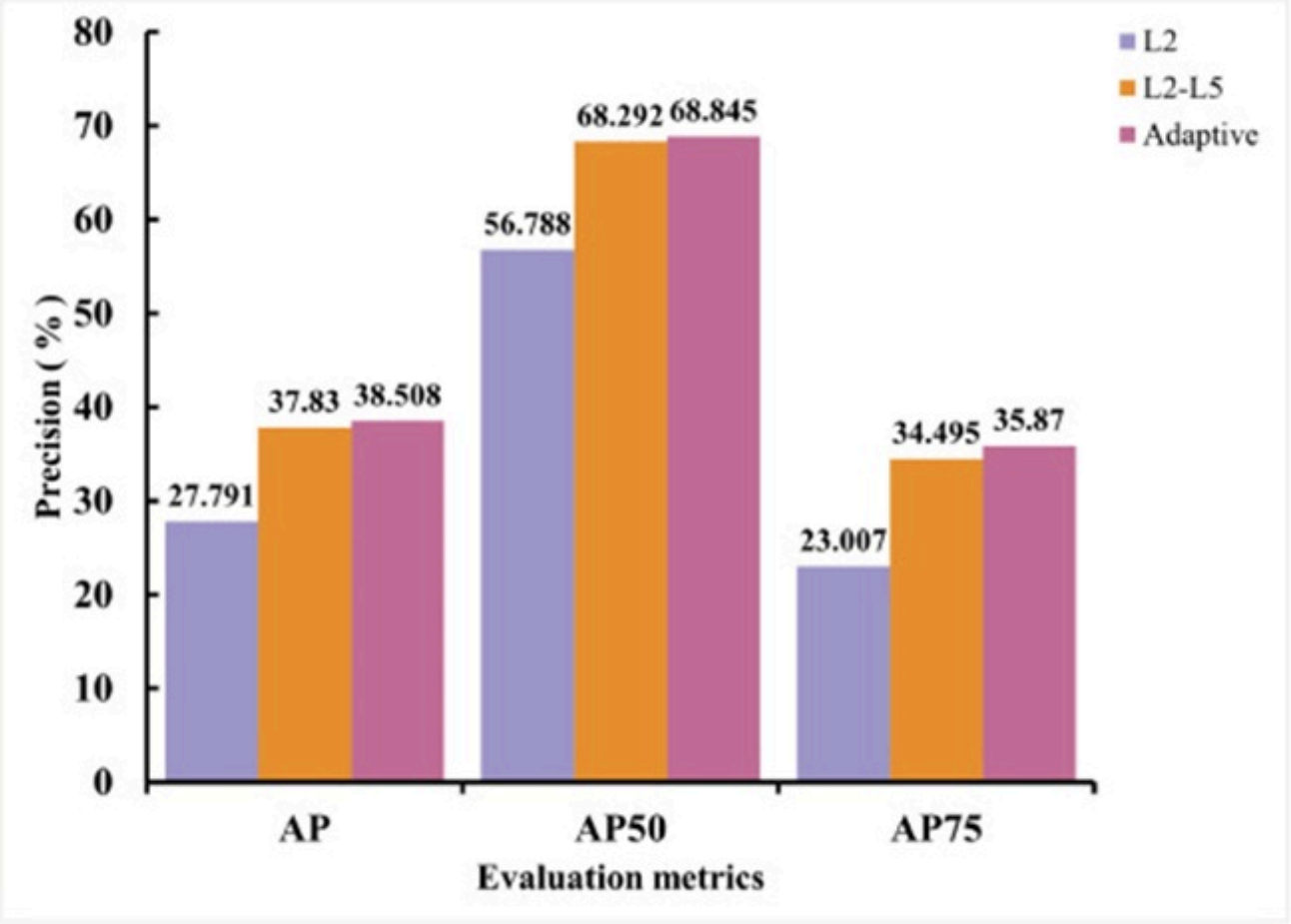


Figure 13. Comparison of average accuracy of selection strategy models with different pyramid levels.

Conclusion

Accurate weed pinpointing is indispensable to the realization of precise, spatially targeted weed suppression. This body of work advances an original, pragmatic weed-segmentation paradigm. The paradigm hinges on marking only a modest

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

number of stochastically generated points per instance, with those point labels subsequently serving as a weak supervisory signal to execute paddy weed segmentation efficiently. This blueprint navigates around the difficulty of annotating weed data while rectifying the substandard accuracy that characterizes presently available weakly supervised weed detection approaches. The body of evidence confirms that the proposed weakly supervised weed segmentation model, PIS-Net, which is anchored in point annotations, outstrips both BoxInst and PointRend. Measured against fully supervised regimes, the performance attained under an annotation mode that expends merely 13 annotation points per weed instance achieves 90% of that realized under the fully supervised training mode. The approach tangibly reduces the annotation overhead associated with agricultural data and has significant potential for real-world deployment to distinguish weeds from rice plants in paddy environments. Although the results, driven by frugal point labeling, are heartening, the model has yet to be compressed into a lightweight variant or subjected to field experiments on mobile devices to assess its real-world accuracy. Moreover, the datasets currently are confined to common paddy weeds such as barnyard grass, hollow lotus seed grass, and Lysimachia. The weed imagery housed in these datasets consists of either single-species stands or pairwise weed mixtures, constituting comparatively idealized weed portrayals. The road ahead may include pruning the model to a leaner form, deploying it for experimental validation on portable platforms, and folding in supplementary multiclass weed community data representative of real-field diversity. This would sweep in a broader spectrum of weed types encountered across varied natural environments—overcast conditions, rainy episodes, and weed captures influenced by shifting natural illumination—thereby extending the framework's reach to weakly supervised learning scenarios for weed detection amid the complex backdrop of rice paddies.

Acknowledgements

None

Conflict of interest

None

Financial support

This research was funded by the National Key Research and Development Program of China (File No 2023YFD1401100), the National Natural Scientific Foundation of China (File No U23A20174), the Earmarked Fund for Modern Agro-industry Technology Research System in China (File No CARS-01), and the Laboratory of Lingnan Modern Agriculture Project (File No NT2021009).

Ethics statement

None

Received: 06 Nov 2022 Revised: 11 Jan 2023 Accepted: 27 Jan 2023
Published online: 25 May 2024

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and

distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Zhu J, Wang J, DiTommaso A, Zhang C, Zheng G, Liang W, et al. Weed research status, challenges, and opportunities in China. Crop Prot. 2020;134.

2. Olsen A, Konovalov DA, Philippa B, Ridd P, Wood JC, Johns J, et al. Deepweeds: a multiclass weed species image dataset for deep learning. Sci Rep. 2019;9:2058.

3. Melander B, Lattanzi B, Pannacci E. Intelligent versus non-intelligent mechanical intra-row weed control in transplanted onion and cabbage. Crop Prot. 2015;72:1-8.

4. Van Der Weide R, Bleeker P, Achten V, Lotz L, Fogelberg F, Melander B. Innovation in mechanical weed control in crop rows. Weed Res. 2008;48:215-24.

5. Liu C, Yang K, Chen Y, Gong H, Feng X, Tang Z, et al. Benefits of mechanical weeding for weed control, rice growth characteristics and yield in paddy fields. Field Crops Res. 2023;293:108852.

6. Carballido J, Rodríguez-Lizana A, Agüera J, P´erez-Ruiz M. Field sprayer for inter- and intra-row weed control: performance and labor savings. Span J Agric Res. 2013;11(3):642-51.

7. Midtiby HS, Mathiassen SK, Andersson KJ, Jørgensen RN. Performance evaluation of a crop/weed discriminating microsprayer. Comput Electron Agric. 2011;77(1):35-40.

8. Zheng Y, Zhu Q, Huang M, Guo Y, Qin J. Maize and weed classification using color indices with support vector data description in outdoor fields. Comput Electron Agric. 2017;141:215-22.

9. Dongjian H, Yongliang Q, Pan L, Zhan G. Weed recognition based on SVM-DS multi-feature fusion. Trans Chinese Soc Agric Mach. 2013;44:182-7.

10. Bakhshipour A, JAFARI A. Evaluation of support vector machine and artificial neural networks in weed detection using shape features. Comput Electron Agric. 2018;145:153-60.

11. Bah M, Hafiane A, Canals R. Deep learning with unsupervised data labeling for weed detection in line crops in UAV images. Remote Sens. 2018;10:1690.

12. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. Proc IEEE/CVF Conf Comput Vis Pattern Recogn. 2022.

13. Ma R, Hu H, Xing S, Li Z. Efficient and fast real-world noisy image denoising by combining pyramid neural network and two-pathway unscented Kalman filter. IEEE Trans Image Process. 2020;29:3927-40.

14. Ma R, Li S, Zhang B, Hu H. Meta PID attention network for flexible and efficient real-world noisy image denoising. IEEE Trans Image Process. 2022;31:2053-66.

15. Dong X, Bao J, Chen D, Zhang W, Yu N, Yuan L, et al. Cswin transformer: A general vision transformer backbone with cross-shaped windows. Proc IEEE/CVF Conf Comput Vis Pattern Recogn. 2022. p. 12124-34.

16. Ma X, Deng X, Qi L, Jiang Y, Li H, Wang Y, et al. Fully convolutional network for rice seedling and weed image segmentation at the seedling stage in paddy fields. PLoS One. 2019;14(4):e0215676.

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

17. Huang H, Deng J, Lan Y, Yang A, Deng X, Zhang L. A fully convolutional network for weed mapping of unmanned aerial vehicle (UAV) imagery. PLoS One. 2018;13(4):e0196302.

18. Peng H, Li Z, Zhou Z, Shao Y. Weed detection in paddy field using an improved RetinaNet network. Comput Electron Agric. 2022;199:107179.

19. Picon A, San-Emeterio MG, Bereciartua-Perez A, Klukas C, Eggers T, Navarra-Mestre R. Deep learning-based segmentation of multiple species of weeds and corn crop using synthetic and real image datasets. Comput Electron Agric. 2022;194:106719.

20. Nong C, Fan X. Semi-supervised learning for weed and crop segmentation using UAV imagery. Front Plant Sci. 2022;13:927368.

21. Qiang Z, Shi J, Shi F. Phenotype tracking of leafy greens based on weakly supervised instance segmentation and data association. Agronomy. 2022;12(7):1567.

22. Thomas LF, Änäkkälä M, Lajunen A. Weakly supervised perennial weed detection in a barley field. Remote Sens. 2023;15(11):2877.

23. Bearman A, Russakovsky O, Ferrari V, Fei-Fei L. What's the point: semantic segmentation with point supervision. Eur Conf Comput Vis. Cham: Springer International Publishing; 2016:549-65.

24. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. Int Conf Learn Represent. 2017.

25. Qi RQ, Su H, Mo K, Guibas LJ. PointNet: deep learning on point sets for 3D classification and segmentation. IEEE Conf Comput Vis Pattern Recogn. 2017.

26. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. IEEE Conf Comput Vis Pattern Recogn. 2017.

27. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. Proc IEEE Int Conf Comput Vis. 2017:2961-9.

28. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE Conf Comput Vis Pattern Recogn. 2016:770-8.

29. Ma R, Li S, Zhang B, Fang L, Li Z. Flexible and generalized real photograph denoising exploiting dual meta attention. IEEE Trans Cybern. 2023;53(10):6395-407.

30. Ma R, Zhang Y, Zhang B, Fang L, Huang D, Qi L. Learning attention in the frequency domain for flexible real photograph denoising. IEEE Trans Image Process. 2024;33:3707-21.

31. Tian Z, Shen C, Chen H. Conditional convolutions for instance segmentation. Comput Vis ECCV. 2020;16:282-98.

32. Kirillov A, Wu Y, He K, Girshick R. PointRend: image segmentation as rendering. Proc IEEE/CVF Conf Comput Vis Pattern Recogn. 2020:9799-808.

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