Pseudo-Label Refinement for Robust Wheat Head Segmentation via Two-Stage Hybrid Training

Jiahao Jiang* Zhangrui Yang* Xuanhan Wang Jingkuan Song
Tongji University
zaleni233@outlook.com, jerry.zryang@hotmail.com, wxuanhan@hotmail.com, jingkuan.song@gmail.com

1. Introduction

Accurate semantic segmentation of wheat heads is crucial for various agricultural applications, which involves many challenges [2, 6] such as limited labeled data and diverse field conditions. This significantly limits the generalization of fully supervised models and requires effective strategies that can leverage all available data, especially from those extensive unlabeled data.

Inspired by prior works [1, 4, 5, 7–10] for fine-grained segmentation, we develop a self-supervised training framework that iteratively refines pseudo-labels through a teacher–student loop. In particular, we propose a two-stage hybrid training strategy. This effectively combines pseudo-label pre-training with high-resolution fine-tuning on ground-truth data. This dual approach ensures both broad feature learning and fine detail capture.

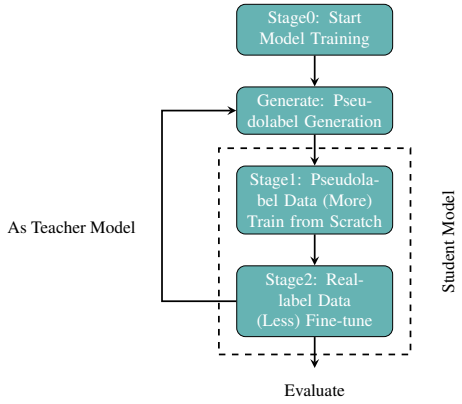

Figure 1. Overview of our self-training method. The process begins with a base model, which generates pseudo-labels for unlabeled data. The student model is then trained in two stages: first on pseudo-labeled data, and then fine-tuned on true-labeled data.

2. Approach

Our methodology is built upon a self-training mechanism, drawing inspiration from knowledge distillation concepts[3]. The core of our approach utilizes the SegFormer model [10], specifically with a Mix Transformer (MiT-B4) as its backbone. We use the same architecture—SegFormer with a MiT-B4 backbone—in all training stages. Each stage shares this structure. Only the training data and hyperparameters differ. This setup maintains consistency while allowing task-specific adaptation. The overall architecture and strategy are illustrated in Figure 1.

The process begins with "Stage 0: Start Model Training". We first train a model using only the limited labeled dataset. This model serves as the initial "teacher." It generates pseudo-labels for the unlabeled data. These labels form the basis for the next training stage. The "student" model then undergoes a two-stage training. In "Stage 1: Pseudolabel Data (More) Train from Scratch," the student model is trained from scratch on a larger set of pseudo-labeled data. Subsequently, in "Stage 2: Real-label Data (Less) Fine-tune," the student is fine-tuned using a smaller set of true-labeled data. The refined student model's performance is evaluated. A critical aspect of our approach is the iterative feedback loop: the refined student model from Stage 2 can itself become a new "teacher" model, generating improved pseudo-labels for subsequent training iterations, thus enabling a self-improving training loop.

2.1. Dataset Utilization

Our dataset strategy involved a phased expansion for robust training. We began by fine-tuning a model pre-trained on ImageNet-1k to generate pseudo-labels using the provided pre-train dataset. These pseudo-labels were meticulously filtered based on a set confidence threshold to ensure high quality for our training set. The provided 99 masked training data points were then used for further fine-tuning, progressively refining performance. This iterative process was repeated multiple times, leading to a gradual improvement in pseudo-label quality.

Data augmentation played a crucial role in preventing overfitting and enhancing model generalization. Our aug-

*Equal contribution to this work.

mentation pipeline included: random cropping, horizontal and vertical flipping, 90-degree rotations, scaling ($\pm$20%), translation ($\pm$6.25%), rotation ($\pm$30°), brightness/contrast adjustments ($\pm$30%), HSV transformations, and ImageNet normalization.

2.2. Novel Contributions

Our primary contribution is a systematic self-training framework specifically tailored for the wheat segmentation task. This framework optimizes data utilization and progressively enhances model accuracy through two key innovations:

1. **Iterative Teacher-Student Loop:** We employ a cyclical learning process where a refined "student" model transforms into the "teacher" for the next iteration. This mechanism continuously generates higher-quality pseudo-labels from unlabeled data, establishing a self-improving loop that learns from the entire dataset.
2. **Two-Stage Hybrid Training Strategy:** Within each cycle, we introduce a specialized two-stage training regimen. The model is initially pre-trained on a large set of high-confidence pseudo-labels at a lower resolution (512x512) to learn robust general features. Subsequently, it is fine-tuned on the high-resolution (1024x1024) ground-truth data to capture fine details.

3. Experiments

3.1. Experimental Settings

Training was conducted on a single NVIDIA RTX 4090 D 24GB GPU (CUDA 12.4) running Ubuntu 20.04.6 LTS. We utilized a two-stage training strategy, with the second stage incorporating 10-fold cross-validation.

- **Stage 1 (Pseudo-label Pre-training):** The model was trained for 40 epochs on 512x512 images with a batch size of 8 and a learning rate of 6×10^{-5}.
- **Stage 2 (Fine-tuning):** The model was fine-tuned for 25 epochs per fold on 1024x1024 images with a batch size of 1. We employed the AdamW optimizer with an initial learning rate of 1×10^{-5}.

Data augmentation served as our primary regularization technique throughout the training process.

3.2. Inference & Post-processing

Our inference pipeline is designed for high accuracy and robustness. We implement a model ensembling strategy by averaging predictions from multiple well-performing models trained during the 10-fold cross-validation. To further enhance performance, extensive Test-Time Augmentation (TTA) is applied to each model. This includes processing the original image, horizontal and vertical flips, 90-degree rotations, and multi-scale inputs (0.75x, 1.25x). The logits from each augmented view are transformed back to their original orientation and then averaged. These TTA-enhanced predictions are subsequently averaged across all ensembled models. The final segmentation mask is obtained by upsampling the ensembled logits to the input resolution (1024x1024), performing an argmax operation, and then resizing the result to the required 512x512 output format using nearest-neighbor interpolation before saving as a CSV file.

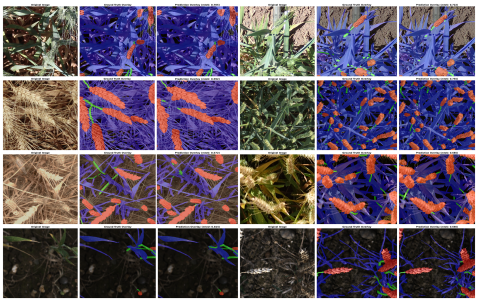

Figure 2. Qualitative results on eight random test samples, including input images, ground-truth, and predictions

3.3. Results

Our model's performance was rigorously evaluated on both the Development Phase and Testing Phase datasets, using the competition-provided metrics. The key results are summarized in Table 1.

Dataset	mIoU
Development Phase	0.7480
Testing Phase	0.7099

Table 1. Performance results on competition datasets

Figure 2 shows that the model effectively segments wheat heads, demonstrating robustness across challenging scenarios.

4. Conclusion

We presented a comprehensive solution for the Global Wheat Full Semantic Segmentation Competition, leveraging a self-training framework with an iterative teacher-student loop and a two-stage hybrid training strategy. Our approach effectively utilized both labeled and unlabeled data through high-confidence pseudo-labeling and progressive refinement. The integration of SegFormer as the backbone, coupled with extensive data augmentation and an advanced inference pipeline incorporating model ensembling and Test-Time Augmentation, contributed to our competitive results. Future work could explore more investigate the impact of different backbone architectures within this iterative framework.

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1

[2] Alireza Ghanbari, Gholam Hassan Shirdel, and Farhad Maleki. Semi-self-supervised domain adaptation: developing deep learning models with limited annotated data for wheat head segmentation. *Algorithms*, 17(6):267, 2024. 1

[3] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1

[5] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[6] Keyhan Najafian, Alireza Ghanbari, Mahdi Sabet Kish, Mark Eramian, Gholam Hassan Shirdel, Ian Stavness, Lingling Jin, and Farhad Maleki. Semi-self-supervised learning for semantic segmentation in images with dense patterns. *Plant Phenomics*, 5:0025, 2023. 1

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[8] Xuanhan Wang, Xiaojia Chen, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Cpi-parser: Integrating causal properties into multiple human parsing. *IEEE Transactions on Image Processing*, 33:5771–5782, 2024.

[9] Huisi Wu, Chongxin Liang, Mengshu Liu, and Zhenkun Wen. Optimized hrnet for image semantic segmentation. *Expert Systems with Applications*, 174:114532, 2021.

[10] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 1