Supplementary Information: A Task-Adapted Structurally Reparameterized Diverse-Branch YOLO Network for Mulberry Leaf Disease Detection

Nazeer Haider[1], **Khasru Alam**[2,*], **Jiaul Hoque Paik**[3,*]

[1]Centre for Computational and Data Sciences, IIT Kharagpur, India
[2]Central Sericultural Research and Training Institute, Berhampore, India
[3]Department of Artificial Intelligence, IIT Kharagpur, India

*Corresponding authors: khasru.csb@gov.in, jiaul@ai.iitkgp.ac.in

---

Supplementary Note 1: Image Classification Baseline Comparisons

In the initial manuscript submission, several standard image classification architectures (ResNet, EfficientNetV2, Vision Transformers, and classification-oriented YOLO variants) were evaluated alongside object detection models to benchmark representation-learning capacity under a pure image-classification paradigm.

In the revised manuscript, the primary paper focuses exclusively on the object detection task (using dedicated detection frameworks such as Faster R-CNN, RT-DETR, YOLOv8, YOLOv9, YOLOv10, YOLOv11, and YOLOv26) to maintain consistency with the whole-leaf detection objective and standard COCO evaluation metrics (mAP@0.5 and mAP@0.5:0.95).

To preserve the comparative studies conducted in the initial submission and provide a complete view of the representation-learning baselines, we present the image classification results in this Supplementary Information document (Supplementary Tables S1 and S2).

Competing Image Classification Methods

ResNet: ResNet (Residual Networks)[1], a fundamental deep learning architecture, was developed to address the vanishing gradient problem in deep neural networks. By introducing shortcut connections, ResNet enables the training of networks up to hundreds of layers. We evaluate five standard variants (ResNet18, ResNet34, ResNet50, ResNet101, and ResNet152) to benchmark traditional convolutional representations.

EfficientNet: EfficientNet[2] balances computational efficiency and accuracy through compound scaling, which uniformly scales network depth, width, and resolution. We include EfficientNetV2 variants (EfficientNetV2-S, EfficientNetV2-M, and EfficientNetV2-L) to represent highly optimized CNN architectures.

ViT (Vision Transformer): The Vision Transformer (ViT)[3] applies self-attention mechanisms to image classification by splitting images into patch sequences. ViT's ability to capture global context makes it a strong baseline for feature extraction. We evaluate four configurations: ViT-B/16, ViT-B/32, ViT-L/16, and ViT-L/32.

YOLO Classification Variants: The YOLO architecture can support pure image classification by replacing the multi-branch detection head with a single classification head evaluated via Binary Cross-Entropy loss. We evaluate classification-oriented variants of YOLOv11 (YOLOv11s-cls, YOLOv11m-cls, YOLOv11l-cls) and YOLOv26 (YOLOv26s-cls, YOLOv26m-cls, YOLOv26l-cls).

Image Classification Performance Results

Supplementary Table S1 summarizes the classification performance of the baseline models on our in-house mulberry dataset. All models were trained under identical classification protocols. Among the standard CNN and ViT baselines, ViT-L/16 achieves the highest accuracy of 92.14%, but requires 304.3M parameters, which limits its utility for edge deployment and further motivates the use of lightweight detection architectures in the main manuscript. The YOLO classification variants achieve substantially higher accuracy with fewer parameters (e.g., YOLOv11m-cls achieves 97.28% accuracy with 11.6M parameters, and YOLOv26m-cls achieves 95.85% accuracy).

Supplementary Table S2 presents the image classification performance on the public dataset (external dataset validation). Standard CNNs and ViT models achieve accuracies ranging from 95.12% to 95.34% (e.g., ResNet50 at 95.32% and ViT-L/16 at 95.29%). YOLO classification baselines achieve higher performance, reaching 96.33% accuracy for both YOLOv11m-cls and YOLOv26m-cls.

Table S1. Image classification performance of various baseline models and classification variants on our in-house dataset. All models are trained under a classification paradigm.

Model	Params (M)	Prec. (%)	Rec. (%)	F1-score (%)	Acc. (%)
ResNet18	11.7	84.87	92.10	88.34	92.10
ResNet34	21.8	84.87	92.10	88.34	92.10
ResNet50	25.5	85.40	92.12	88.37	92.12
ResNet101	44.5	84.85	92.05	88.30	92.05
ResNet152	60.2	85.87	91.90	88.39	91.90
EfficientNetV2-S	21.5	85.97	91.99	88.44	91.99
EfficientNetV2-M	54.1	86.46	92.01	88.36	92.01
EfficientNetV2-L	118.5	85.47	91.90	88.34	91.90
ViT-B/16	86.6	85.22	92.10	88.36	92.10
ViT-B/32	88.2	84.86	92.12	88.34	92.12
ViT-L/16	304.3	86.47	92.14	88.45	92.14
ViT-L/32	306.5	84.85	92.05	88.30	92.05
YOLOv11s-cls	6.7	94.50	96.60	95.45	97.96
YOLOv11m-cls	11.6	97.20	97.80	97.49	97.28
YOLOv11l-cls	17.5	96.00	96.40	96.14	98.64
YOLOv26s-cls	6.7	94.90	96.30	95.57	93.92
YOLOv26m-cls	11.6	97.50	99.30	98.38	95.85
YOLOv26l-cls	14.1	96.10	99.50	97.66	96.44

Table S2. Image classification performance of various baseline models and classification variants on the Public Dataset (external dataset validation).

Model	Params (M)	Prec. (%)	Rec. (%)	F1-score (%)	Acc. (%)
ResNet18	11.7	92.87	95.34	93.32	95.34
ResNet34	21.8	92.87	95.27	93.64	95.27
ResNet50	25.5	93.15	95.32	93.82	95.32
ResNet101	44.5	93.38	95.25	93.97	95.25
ResNet152	60.2	93.00	95.16	93.74	95.16
EfficientNetV2-S	21.5	91.27	95.25	92.99	95.25
EfficientNetV2-M	54.1	90.27	95.25	92.93	95.25
EfficientNetV2-L	118.5	92.57	95.12	93.63	95.25
ViT-B/16	86.6	92.59	95.34	93.19	95.12
ViT-B/32	88.2	91.79	95.18	93.02	95.18
ViT-L/16	304.3	93.64	95.29	94.23	95.29
ViT-L/32	306.5	90.72	95.25	94.23	95.25
YOLOv11m-cls	11.6	93.76	95.81	94.68	96.33
YOLOv26m-cls	11.6	94.94	95.83	95.34	96.33

References

1. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778, DOI: 10.1109/CVPR.2016.90 (2016).

2. Tan, M. & Le, Q. Efficientnetv2: Smaller models and faster training. In Meila, M. & Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 of *Proceedings of Machine Learning Research*, 10096–10106 (PMLR, 2021). https://proceedings.mlr.press/v139/tan21a.html.

3. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* DOI: 10.48550/arXiv.2010.11929 (2020).