Semantic Segmentation of Extremely Small Defects in Sliced Apples

Yueying Shi and Oky Dicky Ardiansyah Prima

Graduate School of Software and Information Science, *Iwate Prefectural University*
152-52 Sugo, Takizawa, Iwate, Japan
E-mail: s236w002@s.iwate-pu.ac.jp, prima@iwate-pu.ac.jp

***Abstract*—Semantic segmentation of extremely small defect regions in food inspection remains a challenging task due to severe foreground–background class imbalance and the high cost of missed detections. In the inspection of sliced apples, remaining skin and core fragments often occupy only a few pixels, making them prone to being overlooked despite high overall segmentation accuracy. This study systematically investigates semantic segmentation strategies for detecting extremely small defects in ultraviolet (UV) images of sliced apples. A stepwise experimental framework is employed to isolate and evaluate the effects of loss functions, encoder architectures, and decoder designs under identical training conditions. Quantitative results demonstrate that region-based loss functions, particularly Tversky and Focal Tversky losses, provide superior spatial consistency and recall compared to pixel-wise reweighting approaches. Furthermore, lightweight encoders, such as MobileNetV2, combined with UNet-based decoders, achieve more stable and robust performance in preserving fine-grained defect regions across different random seeds. These findings provide practical design guidelines for recall-oriented semantic segmentation in food inspection tasks, where reliable detection of extremely small defects is critical for quality assurance and safety.**

Keywords*—*Machine vision; semantic segmentation; small defect regions; extreme class imbalance.

I. INTRODUCTION

Automated visual inspection is a key component of modern food processing industries, where high-throughput production requires consistent quality control and strict safety assurance. Apples are widely processed in sliced form in industrial food production [1][2]. During these mechanical operations, residual skin or core fragments may remain on sliced apples due to natural variations in fruit size, shape, and internal structure. Although such defects are typically small, their presence can negatively affect product appearance and may raise safety concerns, making reliable inspection an important requirement in apple processing lines.

In current industrial practice, inspection of sliced apples is still largely dependent on manual visual checking by human operators. Machine vision–based systems [3][4] and deep learning enable pixel-level defect localization through semantic segmentation. However, detecting defects in sliced apples presents a particularly challenging scenario. Remaining skin and core fragments often occupy less than one percent of the image area and appear as extremely small and sparse regions against a dominant background. Under such severe foreground–background class imbalance, segmentation models may achieve high overall accuracy while still failing to detect critical defect regions, where missed detections are especially problematic in food inspection.

UV imaging has emerged as an effective modality for enhancing the visibility of subtle surface features in food inspection. Under UV illumination, residual skin and core fragments on sliced apples exhibit higher contrast than surrounding flesh, enabling more reliable visual discrimination compared to conventional RGB imaging [5]. When combined with semantic segmentation, UV-based imaging provides a promising approach for detecting extremely small defects at the pixel level. Nevertheless, segmentation performance under such extreme class imbalance remains highly sensitive to both network architecture and optimization strategy, and inappropriate design choices can easily suppress rare defect regions during training.

Although previous studies have investigated apple defect detection using various imaging modalities and deep learning models, most have focused on object-level classification or segmentation of relatively large defects. Systematic evaluation of semantic segmentation strategies for extremely small defect regions remains limited, particularly in terms of isolating the effects of loss functions, encoder architectures, and decoder designs under identical experimental conditions. As a result, practical design guidelines for recall-oriented segmentation in sliced apple inspection are still insufficient.

To address this gap, this study systematically investigates semantic segmentation of extremely small defects in sliced apples using UV images. A stepwise experimental framework is adopted to independently evaluate loss functions, encoder architectures, and decoder designs under severe class imbalance. By emphasizing recall-oriented and overlap-based evaluation metrics, this work aims to identify practical design guidelines that minimize missed detections of small defect regions. The findings of this study provide useful insights for designing robust semantic segmentation models for food inspection tasks, where reliable detection of extremely small defects is critical for quality assurance and safety.

The remainder of this paper is organized as follows. Section 2 reviews related work and outlines key challenges. Section 3 presents the proposed framework. Section 4 describes the experimental setup, followed by results in Section 5. Section 6 concludes the paper and discusses future work.

II. Related Work

Early studies on apple inspection relied on handcrafted features extracted from RGB images to identify surface defects, while recent deep learning–based approaches have significantly improved detection accuracy through object detection and semantic segmentation models [5].

To enhance defect visibility beyond RGB imaging, various modalities, such as near-infrared, short-wave infrared, and X-ray imaging have been explored [6]. These approaches are effective for detecting internal or early-stage defects but often focus on object-level detection or relatively large defect regions and require specialized hardware. In contrast, inspection of sliced apples introduces additional challenges, as defects, such as remaining skin and core fragments, are extremely small, irregularly shaped, and sparsely distributed, making pixel-level semantic segmentation more suitable than object detection.

UV imaging has recently attracted attention for sliced apple inspection because it enhances the contrast of residual skin and core fragments relative to surrounding flesh. Previous studies have shown that UV illumination, when combined with deep learning–based segmentation, enables more reliable detection of such subtle defects. However, semantic segmentation of extremely small defect regions remains difficult due to severe foreground–background class imbalance, where standard pixel-wise loss functions tend to bias optimization toward the dominant background class [5].

To address this issue, imbalance-aware loss functions, such as Dice [7], Tversky [8], and focal variants [9] have been proposed to improve sensitivity to minority regions by optimizing spatial overlap. In addition, network architecture plays a significant role in preserving fine-grained spatial information, as excessive down-sampling can easily suppress small target regions. While these techniques have been extensively studied in medical image segmentation, systematic evaluation of their combined effects in industrial food inspection, particularly for UV-based inspection of sliced apples, remains limited.

Overall, existing studies demonstrate the potential of deep learning and advanced imaging modalities for apple defect detection, yet practical design guidelines for recall-oriented semantic segmentation of extremely small defects in sliced apples are still insufficient. This study addresses this gap by systematically evaluating loss functions and network architectures under identical experimental conditions.

III. Materials and Method

A. Overview of the Framework

This study addresses semantic segmentation of extremely small defect regions in sliced apples using UV images, where remaining skin and core fragments appear as small dark regions occupying only a few pixels.

To systematically analyze this problem, a stepwise experimental framework was adopted (Figure 1). Loss functions, which are computed by comparing the decoder output with the corresponding ground truth and used to optimize the network via backpropagation, encoder architectures, which extract hierarchical features from the input image and compress them into latent representations, and decoder designs, which progressively restore spatial resolution and generate pixel-wise predictions, were evaluated independently under identical training conditions. This framework isolates the impact of each component on segmentation performance.

B. Image Acquisition

UV image acquisition was conducted in a controlled environment to ensure stable and reproducible illumination conditions. Sliced apple samples were placed on a flat platform inside a black enclosure to suppress ambient light and external reflections. Multiple UV light sources were arranged around the enclosure to uniformly illuminate the apple slices and enhance the contrast between defect regions and surrounding flesh.

Images were captured from multiple views to account for variations in defect appearance. For samples containing remaining skin, images were acquired from four directions, while samples containing remaining core were captured from three directions due to limited contrast under bottom-view illumination (Figure 2). This multi-view acquisition strategy increases dataset diversity and improves robustness against viewpoint-dependent variations.

C. Dataset and Processing

Ground-truth annotation was performed manually at the pixel level using a dedicated annotation tool. Three semantic classes were defined: background, remaining core, and remaining skin. The background class includes apple flesh and non-defect regions, while the remaining core and skin classes correspond to defect regions observed under UV illumination.

The annotated dataset was divided into training, validation, and test sets following a fixed ratio to ensure fair evaluation. To mitigate overfitting caused by limited data, data augmentation techniques, such as random rotation and horizontal or vertical flipping were applied to the training set. These augmentations preserve defect characteristics while increasing the effective size of the dataset.

The dataset consists of 339 images collected under a multi-view acquisition setup, where each apple is captured from multiple viewpoints. Images of the same apple are grouped and assigned to a single split (train/validation/test) to prevent data leakage. Each of the remaining core and skin classes

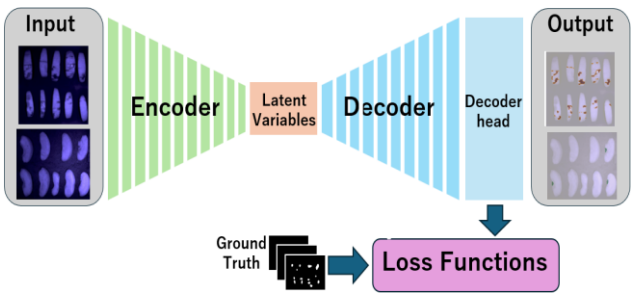

Figure 1. Overall architecture of the encoder-decoder segmentation framework.

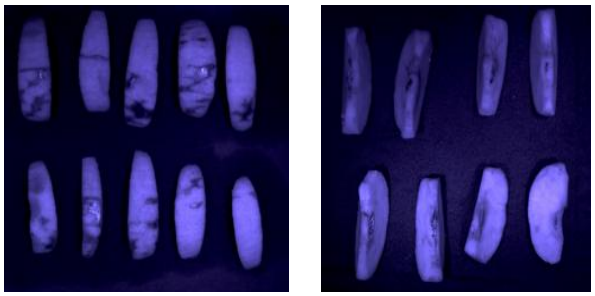

Figure 2. UV images of apple samples containing remaining skin (left) and remaining core (right).

occupies less than 1% of the total image area, confirming the extreme class imbalance in this task.

D. Data Processing

All images were resized to a fixed spatial resolution to standardize input dimensions across experiments. Input images were normalized using the ImageNet mean and standard deviation to ensure compatibility with ImageNet-pretrained encoders. This normalization stabilizes optimization and accelerates convergence during training.

Segmentation masks were resized using nearest-neighbor interpolation to avoid label distortion at class boundaries. This extreme imbalance highlights the necessity of carefully designed loss functions and architectures capable of preserving fine-grained spatial information.

E. Model Architecture

A multi-class semantic segmentation model based on an encoder–decoder architecture was employed. The encoder is responsible for extracting hierarchical feature representations, while the decoder reconstructs spatial resolution and generates dense pixel-wise predictions. ImageNet-pretrained encoders were used in all experiments to improve feature generalization under limited training data.

Multiple encoders (ResNet-34, ResNet-50, MobileNetV2, EfficientNet-B0) [10]-[12] with different depths and computational complexities were evaluated to analyze their ability to preserve small defect features. Lightweight encoders are expected to retain fine-grained spatial details, whereas deeper encoders provide richer contextual information. Several decoder designs were also compared, focusing on their ability to fuse multi-scale features and accurately reconstruct extremely small and fragmented defect regions.

F. Decoder Design

The decoder plays a crucial role in semantic segmentation of extremely small defect regions, where minor spatial reconstruction errors can easily result in missed detections. Its primary function is to restore spatial resolution from encoder feature maps and accurately reconstruct fine-grained defect structures under severe foreground–background class imbalance.

In this study, several representative decoder architectures were evaluated, including UNet, UNet++, DeepLabV3+, and Pyramid Attention Network (PAN)–based decoders [13]-[16]. These architectures differ in their approaches to multi-scale feature fusion and spatial detail recovery. Among them, a UNet-based decoder was adopted as the primary design in the proposed model.

The UNet decoder employs skip connections that directly transfer high-resolution feature maps from the encoder to corresponding decoder stages, effectively mitigating information loss caused by down-sampling. This property is particularly important for preserving boundary and location information of extremely small and sparse defect regions. All decoder variants were integrated with the same encoder backbone and trained under identical conditions to ensure fair comparison.

Pre-experimental results showed that the UNet-based decoder provided more stable and accurate reconstruction of fine-grained defect regions than other decoder designs. Therefore, it was selected as the final decoder configuration used in this study.

G. Loss Function and Training Strategy

Semantic segmentation of extremely small defect regions is strongly affected by severe foreground–background class imbalance, where defect pixels account for only a small fraction of the image. To address this issue, both pixel-wise and region-based loss functions were evaluated in this study, with particular emphasis on reducing false negatives.

Specifically, the following loss functions were examined: Weighted Cross-Entropy loss, Dice loss, Tversky loss, and Focal Tversky loss. Weighted Cross-Entropy assigns larger weights to minority classes based on class frequency and serves as a representative pixel-wise reweighting approach. Dice loss and Tversky loss are region-based overlap losses that directly optimize spatial agreement between predictions and ground-truth masks, making them more suitable for highly imbalanced segmentation tasks. In Tversky loss, the weighting parameters were set to emphasize false negatives, reflecting the importance of recall in defect inspection. Focal Tversky loss further introduces a focusing parameter to emphasize hard-to-segment regions and improve sensitivity to extremely small defects.

All models were trained under identical conditions to ensure fair comparison across loss functions and architectures. The same dataset split, optimizer, learning rate schedule, batch size, and number of training epochs were used throughout the experiments. Training samples were shuffled at each epoch to stabilize optimization, and all networks were trained end-to-end using backpropagation.

This strategy enables systematic analysis of loss functions under controlled conditions. By combining recall-oriented loss functions with controlled training conditions, the proposed framework aims to reliably detect extremely small defect regions in sliced apple inspection.

H. Evaluation Metrics

Segmentation performance was evaluated using recall-oriented and overlap-based metrics that are suitable for extreme class imbalance. Recall measures the ability to detect defect pixels, while Intersection-over-Union (IoU) evaluates spatial consistency between predictions and ground truth. The F2-score was adopted to place greater emphasis on recall than

precision, reflecting the importance of minimizing missed detections in food inspection tasks [17].

These metrics provide a complementary evaluation of segmentation performance for extremely small defects.

IV. Experimental Setup

All experiments were conducted to evaluate semantic segmentation performance for extremely small defect regions in sliced apples under severe foreground–background class imbalance. UV images and corresponding pixel-level annotations were divided into training, validation, and test sets using a fixed split. Data augmentation, including rotation and flipping, was applied only to the training set. All input images were resized to a fixed resolution of 224 × 224 pixels and normalized using ImageNet mean and standard deviation. Segmentation masks were resized using nearest-neighbor interpolation.

A stepwise experimental design was adopted to ensure fair comparison. Loss functions, encoder architectures, and decoder designs were evaluated independently, with only one component varied at a time while all other settings were fixed. ImageNet-pretrained weights were used for all encoders to ensure consistent initialization.

All models were trained for 200 epochs with a batch size of 8 using the same optimizer and learning rate schedule across all experiments. Training samples were shuffled at each epoch, and model parameters were optimized end-to-end using backpropagation. Experiments were implemented using PyTorch and executed on a system equipped with an NVIDIA GeForce RTX 4080 Laptop GPU.

V. Results and Discussion

Following the experimental protocol described above, quantitative and qualitative results are presented in a stepwise manner to clarify the effects of loss functions, encoder architectures, and decoder designs on segmentation performance. Unless otherwise specified, results reported in Tables 1–4 correspond to single training runs, while multi-seed evaluation is conducted for representative configurations to assess robustness.

A. Loss Function Comparison

The first set of experiments examined the influence of loss function design under a fixed encoder–decoder configuration (Table 1). Among the evaluated loss functions, Tversky loss achieved the highest F2-score for remaining core defects (0.82), while Focal Tversky loss yielded the highest F2-score for remaining skin defects (0.578). Dice loss demonstrated relatively balanced performance across both defect types, whereas Weighted Cross-Entropy achieved high recall but substantially lower IoU and F2-scores, indicating degraded spatial consistency.

Pixel-wise reweighting approaches show a fundamental limitation under extreme class imbalance. Although Weighted Cross-Entropy increases the contribution of minority-class pixels, it does not explicitly enforce spatial coherence, leading to fragmented predictions and reduced overlap with ground-truth regions. In contrast, region-based overlap losses directly optimize spatial agreement, making them inherently more suitable for segmenting extremely small and sparse defect regions.

The difference between Tversky and Focal Tversky losses further suggests that defect morphology influences optimal loss design. Remaining core defects tend to form compact regions, benefiting from the false-negative suppression emphasized by Tversky loss, whereas remaining skin defects are often irregular and fragmented, where the focusing mechanism of Focal Tversky loss improves sensitivity to hard-to-segment pixels. This observation underscores the importance of aligning loss function design with defect characteristics rather than relying on a single generic objective.

B. Encoder Comparison under Fixed Loss Functions

1) Core Defect Segmentation under Fixed Tversky Loss

Table 2 compares encoder performance for core defect segmentation under a fixed Tversky loss. MobileNetV2 achieved the highest core F2-score (0.868) and IoU (0.811), outperforming deeper encoders, such as ResNet-34 and ResNet-50. EfficientNet-B0 exhibited the lowest performance among the evaluated encoders.

Lightweight encoders are more effective than deeper architectures when target regions are extremely small. Excessive network depth and repeated downsampling may suppress fine-grained spatial cues associated with small core defects, even if high-level contextual features are well captured. In contrast, MobileNetV2 preserves spatial detail through its compact architecture and reduced parameterization, which appears advantageous under severe class imbalance.

2) Skin Defect Segmentation under Fixed Focal Tversky Loss

Table 3 presents encoder comparison results for skin defect segmentation under a fixed Focal Tversky loss. Again, MobileNetV2 achieved the highest skin F2-score (0.707) and IoU (0.628). Although its core F2-score (0.829) was slightly lower than that obtained under fixed Tversky loss, overall performance remained competitive. ResNet-34 showed moderate performance, while EfficientNet-B0 consistently yielded the lowest scores for both defect types.

Encoder effectiveness depends on preserving spatial resolution rather than representational depth. Lightweight encoders, particularly MobileNetV2, demonstrate strong robustness across defect types and loss configurations, making them well suited for recall-oriented segmentation of extremely small defects [11].

C. Decoder Comparison with Fixed Encoder and Loss Function

Decoder architectures were compared under a fixed MobileNetV2 encoder and Tversky loss, as shown in Table 4. UNet and UNet++ achieved the highest overall segmentation performance among the evaluated models, with UNet achieving a core F2-score of 0.879 and a skin F2-score of 0.678. UNet++ slightly improved skin defect segmentation (F2-score of 0.711) at the cost of a marginal reduction in core performance.

TABLE 1. LOSS FUNCTION COMPARISON UNDER A FIXED ENCODER AND DECODER

Loss Function	Core Recall	Core IoU	Core F1	**Core F2**	Skin Recall	Skin IoU	Skin F1	**Skin F2**
Tversky	0.895	0.75	0.811	**0.828**	0.772	0.403	0.492	0.469
Focal Tversky	0.868	0.65	0.707	0.708	0.773	0.516	0.605	**0.578**
Dice	0.851	0.752	0.806	0.801	0.724	0.488	0.577	0.535
Weighted CE	0.902	0.546	0.609	0.651	0.817	0.396	0.489	0.523

TABLE 2. CORE FOCUSING ENCODER COMPARISON UNDER FIXED TVERSKY LOSS AND DECODER

Encoder	Core Recall	Core IoU	Core F1	**Core F2**	Skin Recall	Skin IoU	Skin F1	**Skin F2**
MobileNetV2	0.866	**0.811**	0.872	**0.868**	0.758	**0.636**	0.724	**0.711**
ResNet34	0.895	0.75	0.811	0.828	0.772	0.403	0.492	0.469
ResNet50	0.905	0.735	0.788	0.795	0.792	0.53	0.617	0.593
EfficientNet-B0	0.87	0.694	0.753	0.759	0.751	0.363	0.454	0.441

TABLE 3. SKIN FOCUSING ENCODER COMPARISON UNDER FIXED FOCAL TVERSKY AND DECODER

Encoder	Core Recall	Core IoU	Core F1	**Core F2**	Skin Recall	Skin IoU	Skin F1	**Skin F2**
MobileNetV2	0.89	0.756	0.821	0.829	0.756	**0.628**	0.717	**0.707**
ResNet50	0.895	0.832	0.889	0.892	0.767	0.511	0.601	0.572
ResNet34	0.868	0.65	0.707	0.708	0.773	0.516	0.605	0.578
EfficientNet-B0	0.85	0.58	0.636	0.637	0.722	0.424	0.513	0.477

TABLE 4. DECODER COMPARISON UNDER A FIXED LOSS FUNCTION AND ENCODER

Decoder	Core Recall	Core IoU	Core F1	**Core F2**	Skin Recall	Skin IoU	Skin F1	**Skin F2**
UNet	0.881	0.815	0.878	**0.879**	0.715	0.614	0.708	0.678
UNet++	0.866	0.811	0.872	0.868	0.758	0.636	0.724	**0.711**
PAN	0.879	0.773	0.841	0.862	0.677	0.597	0.682	0.677
DeepLabV3+	0.87	0.712	0.78	0.8	0.703	0.566	0.651	0.649

PAN demonstrated slightly lower performance across metrics, while DeepLabV3+ yielded the lowest IoU and F2-scores for both defect types. Decoders relying on deep, low-resolution features are less effective for reconstructing extremely small and fragmented regions.

The superior performance of UNet-based architectures highlights the importance of skip connections that directly transfer high-resolution spatial features from the encoder to the decoder. Such connections mitigate information loss caused by down-sampling and are particularly critical when target regions consist of only a few pixels. These findings confirm that decoder design is a decisive factor in small defect segmentation under extreme class imbalance [14].

D. Reproducibility Analysis across Random Seeds

To assess robustness, reproducibility analysis was conducted using three different random seeds for representative model configurations. The MobileNetV2–UNet model with Tversky loss achieved mean core and skin F2-scores of 0.870 ± 0.022 and 0.734 ± 0.035, respectively, indicating both high performance and low variance. In contrast, the UNet++ model with the same configuration showed lower mean performance and higher variance.

Furthermore, the UNet model with a ResNet-50 encoder and Focal Tversky loss exhibited larger fluctuations, particularly for skin defects. Although full multi-seed evaluation was not performed for all configurations due to computational constraints, these results suggest that lightweight encoders combined with skip-connected decoders tend to achieve both higher accuracy and more stable optimization. The observed performance differences are larger than the corresponding variances, indicating that the overall trends are consistent across random initializations.

E. Qualitative Results

Qualitative analysis revealed that false positive predictions primarily appeared as small, isolated regions distributed across background areas. In addition to background noise caused by illumination artifacts or water droplets, this behavior is an expected consequence of optimizing recall-oriented metrics, such as the F2-score, where reducing false negatives is prioritized over suppressing minor false positives under extreme class imbalance [18].

As illustrated in Figure 3, the proposed models successfully localized most remaining core and skin defect regions. Occasional false negatives were observed near ambiguous boundaries or low-contrast regions, reflecting the inherent difficulty of segmenting extremely small defects. From a practical perspective, this trade-off is acceptable, as missing defects are more critical than minor false positives, which can be handled in downstream inspection.

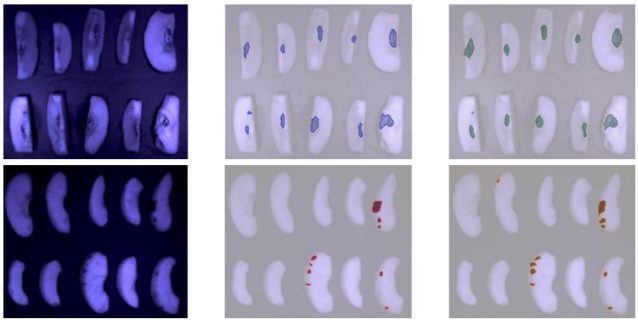

Figure 3. Qualitative segmentation results under UV imaging. From left to right: input images, ground truth masks, and model predictions. The top row corresponds to remaining core, and the bottom row corresponds to remaining skin.

While this study focuses on widely used segmentation architectures and loss functions under controlled conditions, more specialized techniques such as boundary-aware losses, hard-example mining, and high-resolution refinement strategies may further improve performance for extremely small defect regions. In particular, boundary-aware approaches better preserve fine spatial details. These approaches are promising directions for future work and may complement the findings of this study.

VI. Conclusion and Future Work

This study investigated semantic segmentation of extremely small defect regions in sliced apples under severe foreground–background class imbalance. Focusing on remaining skin and core fragments in UV images, a stepwise experimental framework was adopted to systematically evaluate the effects of loss functions, encoder architectures, and decoder designs on recall-oriented segmentation performance.

Experimental results demonstrated that region-based overlap loss functions, particularly Tversky and Focal Tversky losses, consistently outperformed pixel-wise reweighting approaches in terms of spatial consistency and F2-score. These findings indicate that explicitly optimizing spatial overlap is more effective than class-frequency reweighting alone when defect regions occupy only a few pixels.

In terms of network architecture, lightweight encoders, especially MobileNetV2, achieved superior and more stable performance compared to deeper models. This suggests that preserving fine-grained spatial features is more critical than increasing network depth for extremely small defect segmentation. Furthermore, UNet-based decoders with skip connections proved highly effective in reconstructing small and fragmented defect regions, highlighting the importance of direct feature transfer from encoder to decoder under extreme class imbalance.

Reproducibility analysis across multiple random seeds further confirmed that the combination of a lightweight encoder and a skip-connected decoder provides not only high accuracy but also stable optimization, which is essential for practical deployment in industrial food inspection systems. Qualitative results supported these findings by demonstrating reliable localization of most defect regions, with an acceptable trade-off between recall and minor false positives.

Overall, this work provides practical design guidelines for recall-oriented semantic segmentation of extremely small defects in sliced apples. The proposed framework and insights are directly applicable to food inspection tasks where minimizing missed detections is critical for quality assurance and safety. Future work will explore higher-resolution training, advanced post-processing strategies, and integration with real-time inspection pipelines to further improve robustness and deployment readiness.

References

[1] A. B. Oyenihi, Z. A. Belay, A. Mditshwa, and O. J. Caleb, “An apple a day keeps the doctor away: The potentials of apple bioactive constituents for chronic disease prevention,” Journal of Food Science, vol. 87, no. 6, pp. 2291–2309, 2022.

[2] Elsevier, “Apple — an overview,” ScienceDirect Topics, 2025.

[3] K. B. Patel, “A review: Machine vision and its applications,” International Journal of Computer Applications, vol. 70, no. 10, pp. 28–32, 2013.

[4] H. Zhao, “Advances and prospects in machine vision: A critical review based on CiteSpace,” IEEE Access, vol. 8, pp. 12345–12360, 2020.

[5] J. Rahmawan and O. Prima, “Quality inspection of processed apple based on ultraviolet imaging,” Computers and Electronics in Agriculture, vol. 162, pp. 89–97, 2019.

[6] A. Tempelaere et al., “Deep learning for apple fruit quality inspection using X-ray imaging,” in Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 4203–4212, 2023.

[7] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, “Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations,” in Proc. DLMIA, pp. 240–248, 2017.

[8] S. Lu, F. Gao, C. Piao, and Y. Ma, “Dynamic weighted cross entropy for semantic segmentation with extremely imbalanced data,” in Proc. International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), pp. 176–181, 2019.

[9] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, “Tversky loss function for image segmentation using 3D fully convolutional deep networks,” in Proc. MICCAI Workshops, pp. 379–387, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. CVPR, pp. 770–778, 2016.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. CVPR, pp. 4510–4520, 2018.

[12] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in Proc. ICML, pp. 6105–6114, 2019.

[13] O. Ronneberger, P. Fischer, and T. Brox, “UNet: Convolutional networks for biomedical image segmentation,” in Proc. MICCAI, pp. 234–241, 2015.

[14] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A nested UNet architecture for medical image segmentation,” IEEE Transactions on Medical Imaging, vol. 39, no. 6, pp. 1856–1867, 2020.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder–decoder with atrous separable convolution for semantic image segmentation,” in Proc. ECCV, pp. 801–818, 2018.

[16] H. Li, P. Xiong, J. An, and L. Wang, “Pyramid attention network for semantic segmentation,” in Proc. British Machine Vision Conference (BMVC), pp. 285–295, 2018.

[17] Z. Wang et al., “Revisiting Evaluation Metrics for Semantic Segmentation: Optimization and Evaluation of Fine-grained Intersection over Union,” in Proc. NeurIPS Datasets and Benchmarks Track, 2023.

[18] J. Tian, N. Mithun, Z. Seymour, H.-P. Chiu, and Z. Kira, “Striking the Right Balance: Recall Loss for Semantic Segmentation,” in Proc. International Conference on Learning Representations (ICLR), 2021.