Kitchen Food Waste Image Segmentation and Classification for Compost Nutrients Estimation

Raiyan Rahman[1], Mohsena Chowdhury[1], Yueyang Tang[2], Huayi Gao[2], George Yin[2], and Guanghui Wang[1,*]

[1] Department of Computer Science, Toronto Metropolitan University, Toronto, ON, Canada
[2] VCycene Inc. Markham, ON, Canada
Email: {raiyan.rahman, mohsena.chowdhury}@torontomu.ca (R.R, M.C);
{cris, huayi,george}@virgohome.io(Y.T, H.G, G.Y);
wangcs@torontomu.ca (G.W)
*Corresponding author

***Abstract*—The escalating global concern over extensive food wastage necessitates innovative solutions to foster a net-zero lifestyle and reduce emissions. An effective home composter presents a convenient means of recycling kitchen scraps and daily food waste into nutrient-rich, high-quality compost. To capture the nutritional information of the produced compost, we have created and annotated a large high-resolution image dataset of kitchen food waste with segmentation masks of 19 nutrition-rich categories. Leveraging this dataset, we benchmarked four state-of-the-art semantic segmentation models on food waste segmentation, contributing to the assessment of compost quality of Nitrogen, Phosphorus, or Potassium. The experiments demonstrate promising results of using segmentation models to discern food waste produced in our daily lives. Based on the experiments, SegFormer, utilizing MIT-B5 backbone, yields the best performance with a mean Intersection over Union (mIoU) of 67.09. Class-based results are also provided to facilitate further analysis of different food waste classes.**

***Keywords*—semantic segmentation, deep learning, food waste, compost, nutrients**

I. Introduction

Food waste has significant implications for our lives, the environment, economic consequences, and the global community. Taking Canada as an example, it is reported that about 396 kilograms of food annually are wasted or lost per capita, making it one of the top food waste generators in the world [1]. In Canada, more than 32% of our methane (CH_4) production is contributed by food waste in landfills [2]. Turning food waste directly into organic compost is a great way to incentivize this transition into a net-zero sustainable lifestyle for each Canadian household while contributing to cutting down our carbon emissions and meeting the goals our government has set for the Paris Agreement, supporting national efforts to meet environmental targets and contribute to a more sustainable future.

To tackle this intricate issue, we have engineered a home composter named LILA, designed to seamlessly, efficiently, and inconspicuously transform kitchen-generated food waste into organic compost, as depicted in Fig. 1. LILA excels in consistently sorting waste, conditioning, and converting food waste into fully mature organic compost. This system offers a more effective and odor-free solution for recycling food waste, and the transformation of food waste to organic compost presents a promising solution, helping to mitigate waste generation and greenhouse gas emissions, and fostering a more sustainable and environmentally conscious lifestyle.

Fig. 1. The home composter would conveniently allow households to recycle their kitchen scraps and daily food waste into high-quality compost rich in nutrients. This paper aims to study the feasibility of using semantic segmentation techniques to identify food waste classes, enabling the capture of nutritional information from food waste.

The primary objective of this paper is to empower households to capture crucial nutritional information, specifically the NPK (Nitrogen, Phosphorus, and Potassium) values, associated with the compost generated by the composter. This information equips users with the knowledge to judiciously utilize the compost, fostering enhanced agricultural yields and the remediation of contaminated soil. Since various types of food waste, after the composting process, yield distinct NPK values, a key aspect is the automatic recognition of the type of food waste and its corresponding quantity. To achieve this, we propose to utilize computer vision techniques for the segmentation and recognition of different classes of food waste based on their images. This study endeavors to explore the effectiveness and feasibility of employing semantic segmentation techniques for this specific task.

The main contributions of this paper are summarized

Manuscript received June 19, 2024; revised July 5, 2024; accepted August 6, 2024; published March 21, 2025.

below.

- We have compiled and annotated a dataset from high-resolution images collected from the household kitchen scraps and food waste generated after meals. We further narrowed down the initially diverse food waste classes to 19 nutrition-rich categories, facilitating the estimation of final NPK values produced by the in-house composter.
- We conducted a performance evaluation of four state-of-the-art semantic segmentation models on the generated food waste dataset. This benchmarking process allowed us to capture nutritional information from food waste and assess the models' effectiveness in this context.
- The benchmark results underscore the effectiveness of semantic segmentation methods in discerning food waste within real household kitchens. This capability facilitates the straightforward capture of NPK values associated with food waste, streamlining the process of recycling it into compost through a home composter.

II. Related Work

Deep neural networks have exhibited remarkable success in diverse computer vision domains, adeptly handling tasks such as image classification [3], object detection [4], segmentation [5] [6], and recognition [7]. Prominent network models within this domain encompass Convolutional Neural Networks (CNNs) and Vision Transformers [8] [9], both extensively utilized in practical applications across various domains such as multimedia [10], depth estimation [11], agriculture [12], medical image analysis [13] [14] [15], and beyond.

In recent years, these techniques have found application in discerning food images for health and nutrition-related analysis tasks. Significantly, distinctions arise based on the associated task of the dataset and the level of annotation, encompassing categorizations such as dish, ingredient, or recipe. Two noteworthy public image segmentation datasets, namely UECFoodPix and UECFoodPixComplete, offer annotations at the dish level and collectively comprise 10,000 images [16]. These datasets prove to be well-suited for fundamental dish classification tasks [17], contributing to the progression of food-related research within the computer vision landscape.

Addressing the need for a comprehensive food image dataset, FoodSeg103 was introduced, featuring more than 104 ingredient food classes and a total of 7,118 images [18]. This dataset is distinctive in its annotation for semantic segmentation, providing detailed annotations at both the ingredient and dish levels. Notable alternatives include ETHFood101, Recipe1M, and Geo-Dish, primarily designed for dish classification and recipe generation.

Although these datasets serve as valuable benchmarks, it is essential to acknowledge that their emphasis is on entire meals depicted in images rather than on the aspect of food waste. Furthermore, images depicting food waste present a visual contrast, posing a challenge in repurposing existing datasets for this particular task. It is noteworthy that, to date, no publicly available dataset dedicated explicitly to food waste exists in the literature. This study is the first endeavor that is dedicated specifically to food waste segmentation.

III. Method and Experiment

In this section, we will present the dataset we collected to train the learning models, the neural network models we implemented and compared, as well as the experimental details we conducted.

A. Dataset

We gathered kitchen food waste from our clients and captured high-resolution images using smartphone cameras, resulting in a total of 3,128 images at a resolution of 4032×3024. Subsequently, we trained a team of engineers to manually annotate the masks for each food waste instance and assign a class name using the Labelbox platform. This effort led to the annotation of 93 distinct classes, generating 29,433 instance annotations from the captured images.

Aligned with the study's objective of identifying nutritionally rich food waste contributing to high-quality compost with NPK values, the initial set of 93 classes underwent refinement. Following guidelines from [19], the classes were narrowed down to 15 nutrition-rich categories and 4 nutrition-light but high-performing classes. After excluding images lacking annotations for the 19 selected classes, the instance annotations were consolidated for each image to create semantic segmentation masks. This process resulted in a dataset comprising 2,912 images. Table I listed the final 19 distinct food waste classes, together with the associated number of images and nutritional information for each class. The majority of these food waste classes exhibit nutritional richness or a balanced composition in terms of their NPK values, as stipulated by [20], with some classes being particularly rich in specific nitrogen (N), phosphorus (P), or potassium (K) values. Please note that we also include 4 nutrition-light classes due to their prevalence in our kitchen despite their lower nutritional content.

In comparison to other publicly available food datasets, ours is the first to primarily focus on the challenge of semantic segmentation within food waste classes post-meals. In our study, we adopt 10-fold cross-validation during the experiments. For this purpose, we shuffled the dataset randomly and split all images into 10 roughly equal groups, maintaining the class distribution if possible.

B. Models

In recent years, deep neural networks have garnered significant success in semantic segmentation, greatly enhancing our ability to comprehensively understand images. Architectures like U-Net, PSPNet, and SegFormer have excelled in various semantic segmentation tasks across diverse domains, including health, agriculture, and autonomous vehicles [21] [22] [23]. In this paper, we investigate the efficacy of semantic segmentation in the specific context of identifying and localizing food waste. To establish a benchmark for our experiments, we employed the following four state-of-the-art segmentation models: PSPNet [24], SETR [25], Segmenter [26], and SegFormer [27].

PSPNet, introduced in [24], is a semantic segmentation model characterized by its incorporation of a pyramid pooling module. This module enables the capture of multiscale information, enhancing segmentation accuracy by fostering

TABLE I
TABLE DETAILING THE FOOD WASTE CLASSES PRESENT IN THE DATASET AND THEIR CORRESPONDING REPRESENTATION AS WELL AS NUTRITIONAL INFORMATION.

Class	# Images	Nutrition Information	Nitrogen, N (mg)	Phosphorus, P (mg)	Potassium, K (mg)
Banana Skin	485	Balanced	443.75	100	420
Egg Shell	581	Balanced	350	160	150
Lettuce Leaf	410	Nutrition-light	180	27	91
Hard Bread	261	Nutrition-rich	1970	212	250
Cooked Meat	201	Nutrition-rich	3260	280	476
Onion Skin	493	Balanced	431.25	300.36	161.20
Potato Skin	232	Nutrition-rich	3152	262.20	287.14
Apple Core	212	Nutrition-light	4.2	72	95
Orange	107	Nutrition-light	140	23	166
Waffle	43	Nutrition-rich, Nitrogen-rich	1510	254	217
Apple Peel	164	Nutrition-median, Potassium-rich	12.5	12	257.57
Corn Leaves	44	Nutrition-light	28	1.5	16.6
Cucumber	59	Nutrition-Median, Potassium-rich	13.06	24	147
Grape	98	Balanced	150	25	229
Orange Skin	629	Nutrition-median, Potassium-rich	93.75	21	212
Tea Bag	194	Nutrition-rich, Nitrogen-rich	4160	650	2000
Avocado Skin	196	Nutrition-rich, Nitrogen-rich	1100	141	459
Chicken Bone	161	Nutrition-rich, Phosphorus-rich	646.88	2040	40
Cooked Fish	58	Nutrition-rich, Nitrogen-rich	2610	205	372

context awareness. PSPNet has gained widespread popularity as a segmentation network, demonstrating state-of-the-art results across diverse computer vision applications. Its success extends to domains such as autonomous vehicles, medical image analysis, and other tasks requiring complex scene understanding.

The Segmentation Transformer (SETR) [25], is a trailblazer in leveraging transformer architecture for segmentation tasks. This model seamlessly integrates convolutional layers with transformer layers to extract features from image patches. The initial convolutional layers focus on capturing low-level features, while the transformer layers adeptly handle high-level semantic information and context. SETR adopts a hybrid architecture, demonstrating its efficacy through remarkable results in diverse segmentation benchmarks.

Segmenter [26] is a recent transformer model designed for semantic segmentation. Unlike SETR, this model is entirely transformer-based and adopts an encoder-decoder architecture. It maps a sequence of patch embeddings to pixel-level class annotations. The transformer encoder processes the sequence of patches, followed by decoding through either a point-wise linear mapping or a mask transformer. Notably, the Segmenter model excels in capturing long-range dependencies and contextual information within images, showcasing its effectiveness in comprehending complex scenes and accurately segmenting objects in images.

SegFormer [27] represents an extension of the transformer architecture tailored for computer vision applications. In SegFormer, a grid of image patches is treated as a sequence of tokens, which undergo processing by transformer layers. This model innovatively combines local self-attention with global self-attention. Local self-attention is applied within image patches to capture fine-grained details, while global self-attention captures long-range dependencies across patches. To address the high computational cost associated with self-attention in large images, SegFormer incorporates efficient attention mechanisms. These mechanisms enable the model to focus on crucial image regions, thereby reducing overall computational complexity. SegFormer has showcased competitive performance across various computer vision benchmarks, owing to its efficiency and accuracy, positioning it as a promising architecture for a range of segmentation tasks.

TABLE II
RESULTS FROM SEMANTIC SEGMENTATION MODELS.

Method	Backbone	mIoU
PSPNet	ResNet50-D8	57.91 ± 2.59
SegFormer	MIT-B0	65.78 ± 3.22
SegFormer	**MIT-B5**	$\mathbf{67.08 \pm 3.25}$
SETR_naive	ViT-L	40.45 ± 3.34
Segmenter_mask	ViT-B_16	45.16 ± 3.71

C. Experiments

Pre-Processing: The initial images collected were of dimensions 4032×3024. The dataset was partitioned randomly into 10 approximately equal groups to facilitate 10-fold cross-validation. Subsequently, all images were resized to 1024×1024 and subjected to random cropping, retaining 75% of the image. Additionally, horizontal flipping was applied randomly with a 50% probability, and photometric distortion was introduced before the training process. To address the substantial class imbalance within the dataset, class weights were computed as outlined below.

$$W_{class\ i} = \frac{Total\ \ Pixels}{Class\ i\ \ Pixels} \tag{1}$$

Training Setup: The models were all implemented in PyTorch in Python using the MMSegmentation framework [28]. They were trained with 4 NVIDIA Tesla V100 GPUs with 48G memory in total. We used SETR_Naive with a ViT-L backbone, Segmenter_Mask with a ViT-B_16 backbone, and SegFormer with an MIT-B0 backbone.

Training Pipeline: For the PSPNet, Stochastic Gradient Descent (SGD) was used with a learning rate of 0.01, momentum of 0.9, and a weight decay of 0.0005. The model was trained with a batch size of 2 for 80,000 iterations. We adopt a polynomial learning rate decay schedule and employ SGD as the optimizer for the SETR and Segmenter models. We set the initial learning rate at 0.001. Momentum and weight decay are set to 0.9 and 0 respectively for all

TABLE III
CLASS-BASED RESULTS FROM THE SEMANTIC SEGMENTATION MODELS.

Class	Models			
	PSPNet	**SegFormer**	**SETR**	**Segmenter**
Banana Skin	70.51 ± 1.95	72.07 ± 0.41	48.31 ± 5.87	63.19 ± 3.68
Egg Shell	73.65 ± 2.05	74.99 ± 3.72	31.12 ± 2.45	58.65 ± 7.21
Lettuce Leaf	49.19 ± 9.58	57.89 ± 1.42	38.56 ± 4.93	39.33 ± 4.27
Hard Bread	69.51 ± 6.36	81.86 ± 3.98	55.56 ± 3.23	73.43 ± 2.38
Cooked Meat	56.61 ± 4.72	44.25 ± 7.93	38.46 ± 8.02	26.71 ± 5.36
Onion Skin	52.14 ± 5.55	57.32 ± 4.70	32.22 ± 3.17	39.16 ± 4.87
Potato Skin	33.63 ± 7.25	30.37 ± 8.34	30.89 ± 5.96	21.38 ± 10.02
Apple Core	58.93 ± 8.29	74.10 ± 4.70	32.83 ± 5.34	35.61 ± 11.14
Orange	51.01 ± 20.28	68.51 ± 4.48	63.34 ± 6.07	68.17 ± 1.35
Waffle	62.24 ± 24.38	88.31 ± 4.51	72.46 ± 20.35	22.34 ± 23.72
Apple Peel	40.82 ± 6.73	56.05 ± 10.91	20.07 ± 19.97	37.21 ± 8.11
Corn Leaves	68.17 ± 12.72	86.89 ± 2.61	44.58 ± 10.07	63.36 ± 8.05
Cucumber	65.64 ± 24.07	75.67 ± 5.31	63.58 ± 12.75	74.79 ± 2.47
Grape	64.41 ± 7.89	75.61 ± 7.67	65.72 ± 6.71	56.65 ± 5.24
Orange Skin	51.39 ± 4.38	56.23 ± 3.93	40.88 ± 3.84	50.96 ± 2.93
Tea Bag	56.19 ± 6.43	58.87 ± 6.62	23.07 ± 6.52	47.50 ± 7.84
Avocado Skin	49.16 ± 5.75	55.52 ± 2.85	28.27 ± 14.61	26.91 ± 9.13
Chicken Bone	56.13 ± 5.61	61.01 ± 3.21	42.51 ± 4.26	55.72 ± 5.94
Cooked Fish	40.83 ± 18.41	57.64 ± 2.03	49.21 ± 8.67	48.51 ± 1.67

the experiments on both the models. For the SegFormer network, we trained the models using AdamW optimizer and a batch size of 2 for 160K iterations. The learning rate and weight decay were set to an initial value of 0.00006 and 0.01 respectively and then used a "poly" LR schedule with factor 1.0 by default. To calculate the loss during training, we utilized Cross-Entropy Loss. All other hyperparameters were kept consistent with their default implementation.

Model Evaluation: To evaluate our models, we used the intersection over union (IoU) metric, a foundational evaluation metric used in detection and segmentation.

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{2}$$

IV. RESULT AND DISCUSSION

In our experiments, we implemented and compared four state-of-the-art semantic segmentation models on the generated food waste dataset. We performed 10-fold cross-validation to obtain statistically meaningful results. The performance of the four models with different backbones is shown in Table II. The results show the mean and standard deviation of the mIoU for each model. It is evident that SegFormer with an MIT-B5 backbone achieved the best result with an mIoU of 67.08 ± 3.25. To further delineate the segmentation results based on each specific food waste class, we also provide the class-based segmentation results in Table III. Out of the 19 food waste classes, egg shells and banana skins were consistently the top performing classes while potato skin, cooked fish, and apple peel were the worst performing. Waffle, cucumber, and orange classes also had the highest standard deviation, likely due to their low training samples.

Fig. 2 presents a visualization of the qualitative results of the four segmentation models. The figure illustrates the original images and the disparities between semantic segmentation predictions and ground truths involving eggshells, onion skin, and hard bread. All models demonstrate commendable performance. We can see that SegFormer and PSPNet consistently generate superior prediction masks characterized by precise boundaries, effectively avoiding confusion with visually similar objects in the images. In contrast, SETR and Segmenter exhibit challenges in distinguishing between backgrounds that share visual similarities with other classes, resulting in less distinct segmentation

In the initial dataset, classes with high nutritional values in NPK were retained, given their significant contribution to the nutritional content of the resulting compost. However, this selective inclusion led to a class imbalance in the food waste dataset, impacting the model's performance, as evident in the class-wise results presented in Table III. The close relationship between image count and class performance is apparent, as reflected by the high standard deviation for classes with a relatively lower number of images in the dataset. The models tend to overfit the data due to the lower training samples, resulting in reduced generalizability. Based on our observations and the class-specific results, an optimal number of images for consistent and accurate results is estimated to fall within the range of 400 to 600 images per class.

The qualitative results indicate that classes with low standard deviation exhibit more consistent, accurate, and higher-quality mask predictions. This observation is particularly evident in classes such as eggshells, onion skin, hard bread, and banana peels. Conversely, classes like oranges and orange peels, as well as apple cores and apple peels, display a substantial overlap in visual characteristics, presenting a greater challenge in distinguishing one class from the other. As a result, these classes exhibit higher difficulty levels, leading to increased variability and less precise segmentation outcomes.

V. CONCLUSION

Semantic segmentation is essential for predicting NPK values from images. In this study, we utilized high-resolution images of kitchen scraps and food waste, narrowing down the initially diverse waste classes to 19 nutrition-rich classes for experimentation. Based on the dataset, we have benchmarked four state-of-the-art semantic segmentation models by incorporating various data augmentation techniques and leveraging pre-trained weights.

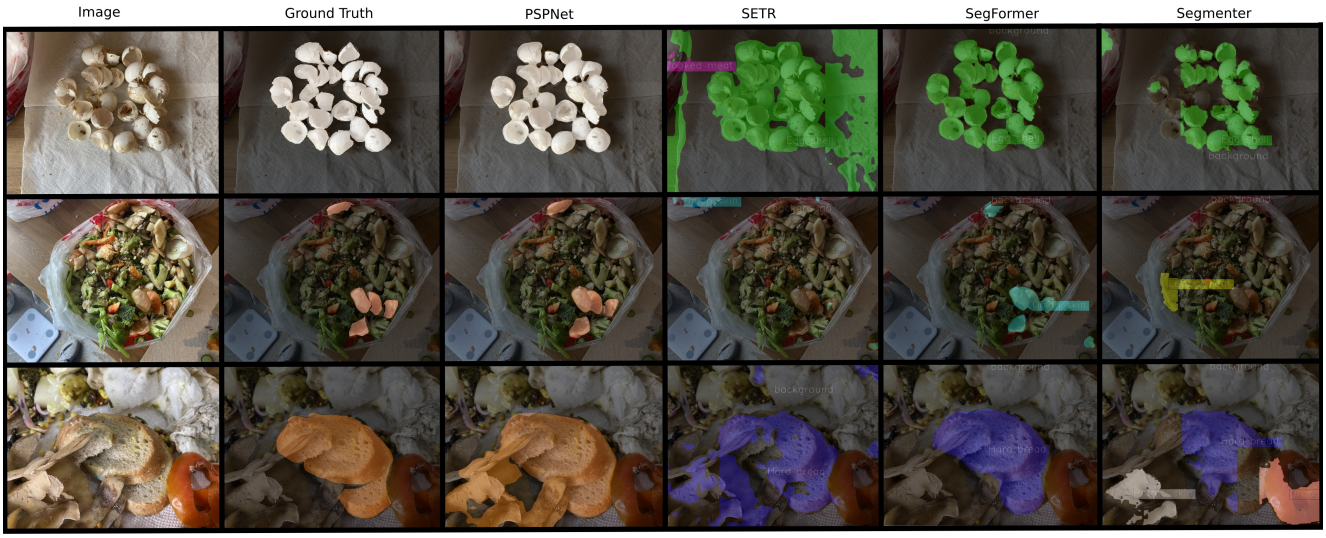

Fig. 2. Visual results comparing the ground truths to their corresponding segmentation predictions are presented for cases involving egg shells, onion skins, and hard bread across various segmentation models.

Our results highlight that transformer-based models, particularly SegFormer, exhibit superior accuracy. Additionally, from qualitative assessments, PSPNet emerges as a strong choice, generating high-quality masks for classes it can effectively learn. PSPNet also demonstrates proficiency in distinguishing the background class from foreground food waste. The promising outcomes from our experiments provide a clear path for future work, wherein we aim to extrapolate NPK values for each food waste class based on their detected masks. By expanding our dataset to include more images for adequate representation of each class, we plan to explore further enhancements in semantic segmentation to achieve higher-quality masks for more accurate calculation of NPK values across various food waste classes.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

The first two authors contributed equally to this work. Y.T, H.G, G.Y collected and analyzed the data. R.R prepared the dataset and implemented and experimented with one CNN-based model. M.C augmented the dataset and implemented and experimented with three transformer-based models. R.R, M.C and G.W conducted the paper writing and final version conceptualization, supervision, review, and editing by G.W. All authors approved the final version.

Funding

This work was partly supported by the Natural Sciences and Mitacs Accelerate Program under grant number ALLRP 576912-22.

References

[1] H. Byrnes, and T. Frohlich, "Canada produces the most waste in the world. The US ranks third," *USA TODAY*, July 12, 2019.

[2] B. Weber, "Canada among heaviest food wasters on the planet, report says," *Global News*, April 3, 2018.

[3] X. Chen, Y. Qin, W. Xu, *et al.*, "Improving vision transformers on small datasets by increasing input information density in frequency domain," in *CVPR Workshops*, 2022.

[4] K. Li, M. Fathan, K. Patel *et al.*, "Colonoscopy polyp detection and. classification: Dataset creation and comparative evaluations," *Plos ONE*, 2021.

[5] F. Wahid, G. Raju, S. Joseph, D. Swain, O. Das *et al.*, "A novel fuzzy-based thresholding approach for blood vessel segmentation from fundus image," *Journal of Advances in Information Technology*, vol. 14, no. 2, pp. 185–192. 2023

[6] L. He, J. Lu, G. Wang, G. *et al.*, "SOSD-Net: Joint semantic object segmentation and depth estimation from monocular images," *Neurocomputing*, vol. 440, pp. 251–263, 2021.

[7] Y. Yang, T. Zhang, G. Li *et al.*, "An unsupervised domain adaptation model based on dual-module adversarial training," *Neurocomputing*, vol. 475, pp. 102–111, 2022.

[8] X. Chen, Q. Hu, Q. *et al.*, "Accumulated trivial attention matters in vision transformers on small datasets," in *Proc. WACV*, 2023.

[9] W. Ma, X. Tu, B. Luo *et al.*, "Semantic clustering based deduction learning for image recognition and classification," *Pattern Recognition*, vol. 124, 2022.

[10] W., Xu, C. Long, R. Wang *et al.*, "DRB-GAN: A dynamic res-block generative adversarial network for artistic style transfer," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021

[11] W. Chao, X. Wang, Y. Wang *et al.*, "Learning sub-Pixel disparity distribution for light field depth estimation," *IEEE Transactions on Computational Imaging*, vol. 10, 2023.

[12] T. Zhang, K. Li, X. Chen *et al.*, ñAphid cluster recognition and detection in the wild using deep learning models," *Scientific Reports*, 2023.

[13] B. Wilson, D. Tucci, D. Moses *et al.*, "Harnessing the power of artificial intelligence in Otolaryngology and the communication sciences," *Journal of the Association for Research in Otolaryngology (JARO)*, vol. 23, pp. 1–31, 2022.

[14] T. Gayathri and K. S. Kumar, "A deep learning based effective model for brain tumor segmentation and classification using MRI images," *Journal of Advances in Information Technology*, vol. 14, no. 6, pp. 1280–1288, 2023.

[15] X. Xiao, Q. Hu, G. Wang "Edge-aware multi-task network for integrating quantification segmentation and uncertainty prediction of liver tumor on multi-modality non-contrast MRI," in *Proc. of MICCAI*, 2023.

[16] T. Ege and K. Yanai, "A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice," in *Proc. of MADiMa*, 2019.

[17] K. Okamoto and K. Yanai, "UEC-FoodPIX complete: A large-scale food image segmentation dataset," in *Proc. of ICPR Workshops*, 2021.

[18] X. Wu *et al.*, "A large-scale benchmark for food image segmentation," in *Proc. of ACM MM*, 2021.

[19] REFRESH. (2019). FoodWasteExplorer. [Onine]. Available: URL: http://foodwasteexplorer.eu

[20] FoodData Central, U.S. Department of Agriculture, Agricultural Research Service.

[21] A. M. Bur *et al.*, "Interpretable computer vision to detect and classify structural laryngeal lesions in digital flexible laryngoscopic im-ages," *Otolaryngology-Head and Neck Surgery*, vol. 169, no. 6, pp. 1564–1572, 2023. doi: 10.1002/ohn.411

[22] R. Rahman *et al.*, "On the real-time semantic segmentation of aphid clusters in the wild," in *Proc. of CVPR Workshops*, 2023.

[23] S. Gowravaram *et al.*, "Prescribed grass fire mapping and rate of spread measurement using NIR images from a small fixed-wing UAS," in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2023.

[24] H. Zhao, J. Shi *et al.*, "Pyramid Scene Parsing Network," arXiv preprint, arXiv:1612.01105, 2017.

[25] S. Zheng *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. of CVPR*, 2021.

[26] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in P*roc. of CVPR*, 2021.

[27] E. Xie, et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," arXiv preprint, arXiv:2105.15203, 2021.

[28] MMSegmentation Contributors, "MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark," 2020.