Efficient deep learning methods for food localization in canteen trays

Flavio Piccoli, Marco Buzzelli, Davide Marelli, Simone Bianco, Gianluigi Ciocca, Raimondo Schettini
University of Milano-Bicocca
{firstname}.{lastname}@unimib.it

2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI) | 979-8-3503-6213-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/RTSI61910.2024.10761232

***Abstract*—This paper presents a comparative study of various efficient state-of-the-art image segmentation models applied to the challenging task of food localization in trays in canteen environments. Using the UNIMIB2016 dataset, which comprises images of canteen trays with multiple food items, we evaluate the performance of ten deep learning-based methods in terms of their segmentation accuracy measured by the Jaccard Index, and computational efficiency measured by Multiply-Accumulate (MACs) operations and the number of parameters. Our results illustrate a trade-off between computational demand and accuracy, with DABNet achieving the highest accuracy but at a cost of lower efficiency compared to others, while models like ENet or EDANet offer a balanced solution suitable for real-time applications. The study not only benchmarks these models but also discusses the implications of different architectural choices, such as the use of dilated and depth-wise convolutions, which influence the models' performance. This work aims to guide the selection of appropriate segmentation models for dietary management systems in canteen settings, contributing to advancements in automated food service operations and dietary monitoring.**

***Index Terms*—Food localization, semantic segmentation, canteen automatization, industry 4.0**

I. Introduction

The rapid advancement in computer vision and image analysis has paved the way for innovative applications in various fields, notably in environments where automation and accuracy are crucial. One such application is food segmentation in canteen environments, which involves distinguishing between food items and non-food items within digital images [1], [2]. This task is foundational for numerous applications, including automated calorie estimation, waste management, and the enhancement of operations in the food industry service [3].

Canteens, whether in educational institutions, corporate offices, or public facilities, serve thousands of meals daily, necessitating efficient management and operation. Automating the process of identifying and analyzing what is on a plate can significantly contribute to more sustainable and health-conscious food management [4], [5]. By employing food segmentation techniques, it becomes feasible to monitor consumption patterns, manage inventory more effectively, and reduce food waste by adapting offerings according to real-time data on food preferences and consumption [6], [7].

Figure 1 shows an example of use of food localization. In a canteen scenario, customers fill the trays with the food and, at checkout, an intelligent system using computer vision and deep learning techniques (e.g. [5], [8]) automatically analyzes the trays providing information on the meal to be consumed. This information can then be exploited by the canteen to plan future food supply, and by the consumers to fill a food diary.

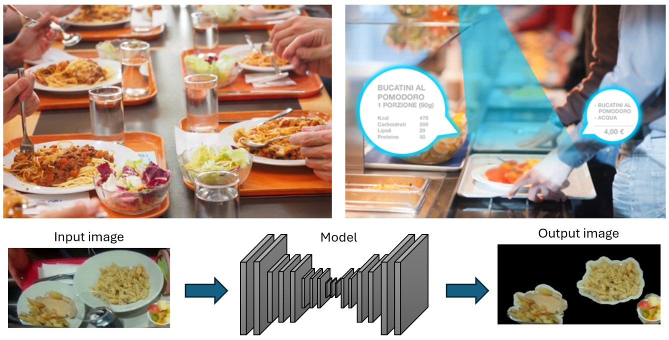

Fig. 1. Food localization in the context of a canteen scenario.

Moreover, the segmentation of food items from their surroundings and other non-food elements in canteen settings poses unique challenges. These include variations in food appearance due to cooking style, presentation, and overlapping items, as well as differing lighting conditions and background complexities [8]. Addressing these challenges requires robust algorithms that can generalize across various environments and conditions, making this an active area of research [9].

This paper aims to compare existing state-of-the-art machine learning techniques for food segmentation, evaluating their effectiveness on a dataset specifically curated to reflect the diverse and dynamic nature of canteen settings. Through this comparative analysis, we aim to identify which techniques are most effective and efficient, thereby contributing to the broader field of automated image processing in real-world applications and setting a benchmark for future innovations in the domain.

In summary, the contributions of this paper are as follows:

- We conduct a thorough comparison of multiple state-of-the-art image segmentation models on a carefully annotated food tray dataset.
- We assess model performance not only based on segmentation accuracy using the Jaccard Index but also consider the computational efficiency, measured in terms of Multiply-Accumulate (MACs) operations and the number of parameters, highlighting the trade-offs between accuracy and computational demand.
- We provide insights into the impact of different architec-

979-8-3503-6213-8/24/$31.00 ©2024 IEEE

tural choices, such as the use of dilated convolutions and depth-wise separable convolutions, on the performance of the models.

- We discuss the practical implications of each model's performance, offering guidance for selecting appropriate models for real-time food segmentation in canteen environments, thus enhancing the operational efficiency and dietary management in these settings.

II. Related Works

Image segmentation has evolved dramatically with deep learning technologies, particularly when applied to real-time and mobile applications. The necessity for efficient computational strategies has led to the development of various techniques aimed at optimizing the performance of neural networks while minimizing resource usage. This section briefly introduces the foundational methods employed across different state-of-the-art architectures before delving into specific network designs.

A. Techniques for Efficient Image Segmentation

Dilated Convolutions: One prominent technique for enhancing the field of view in convolutional networks without increasing the number of parameters is the use of dilated convolutions [10]. By spacing out the kernel elements, dilated convolutions cover a larger input area, capturing more contextual information without the computational cost associated with larger kernels. This method is particularly useful in dense prediction tasks like semantic segmentation where capturing spatial hierarchies at different scales is crucial.

Depth-wise Separable Convolutions: Popularized by architectures like MobileNet [11], depth-wise separable convolutions [12] significantly reduce computational complexity and model size. This technique breaks down the convolution operation into two layers: a depth-wise convolution that applies a single filter per input channel, and a 1×1 point-wise convolution that combines the output channels. This approach reduces the computational cost and the number of parameters dramatically, enabling more efficient processing without a substantial drop in performance [13].

Kernel Decomposition: Another method to reduce the computational burden is kernel decomposition [14], where a standard convolutional kernel is decomposed into smaller, more manageable kernels. For example, a 3×3 kernel can be decomposed into a combination of 1×3 and 3×1 convolutions, reducing the parameter count and computational complexity. This technique often appears in networks aiming to balance accuracy and efficiency.

B. Types of Residual Blocks

Residual blocks, particularly those used in ResNet architectures [15], have been pivotal in enabling the training of very deep networks by using skip connections to mitigate the vanishing gradient problem. Different variations of residual blocks have been adapted to further enhance the efficiency of semantic segmentation networks:

Bottleneck Blocks: These blocks reduce the dimensionality at the first layer, process activations through a smaller dimensional space, and then restore the dimensions at the last layer, which saves a significant amount of computation especially when the input and output dimensions are large.

Bottleneck-1D Blocks: Adapted for more efficiency, these blocks modify the bottleneck structure by using 1D convolutions, which can reduce the computational cost while still maintaining effective channel interactions and spatial hierarchy.

Non-Bottleneck Blocks: These blocks do not reduce the channel dimensions but may incorporate mechanisms like dilated convolutions and separable convolutions to enhance feature extraction without excessive computation.

Non-Bottleneck-1D Blocks: Combining the features of non-bottleneck design with 1D convolutions, these blocks aim to provide an efficient pathway for maintaining spatial relationships and channel-wise feature processing with reduced computational demands.

Employing these techniques and block designs, modern architectures strive to achieve a delicate balance between computational efficiency and segmentation accuracy.

C. Decoding and Upsampling Techniques in Image Segmentation

Decoding Strategies: In pursuit of efficiency, several semantic segmentation architectures opt to modify or entirely skip the decoding process. Typically, this results in faster inference times at the expense of output resolution, which can be crucial for applications requiring fine-grained detail. For example, some methods, like DABNet [16] and EDANet [17], choose a minimalist approach where they generate coarse segmentation maps directly from low-dimensional representations without traditional decoding. This strategy is beneficial in scenarios where speed is more critical than pixel-perfect accuracy.

Full convolutions: Also known as deconvolutions, transposed convolutions or fractionally strided convolutions, full convolutions are a common method for upsampling and are often used to reverse the spatial downsampling effect of conventional convolutions [18]. This technique involves padding zeros in the input data, which allows the convolution to upsample the feature map instead of reducing its size. Networks like ENet [19] use transposed convolutions to gradually restore the feature map dimensions to that of the input image, facilitating fine detail in the output segmentation.

Max Unpooling: the max-pooling indices recorded during the downsampling phase are used to perform non-linear upsampling in the decoder. This method directly uses the spatial positions from the pooling process to guide the placement of values in the upsampling phase, thus helping to better reconstruct the structure of the input image without additional parameters.

Two-Branch Systems: FastSCNN [20] and ContextNet [21] illustrate architectures employing a two-branch system where one branch processes the input at full resolution to preserve spatial details, and another branch processes a downsampled version to efficiently extract high-level semantic information. The features from both branches are then merged, balancing detail and context in the final segmentation output. This method effectively captures both fine and coarse features, which are crucial for accurate segmentation.

Bilinear Upsampling: Some networks opt for bilinear upsampling due to its simplicity and efficiency, as it does not involve learning any additional parameters. This technique uses linear interpolation to estimate the values at upscaled positions, providing a smooth and computationally light way to increase the resolution of feature maps. LEDNet [22] and several other models utilize bilinear upsampling to efficiently produce higher-resolution outputs from encoded features.

Attention Mechanisms in Upsampling: Incorporating attention mechanisms during the upsampling process is a newer trend aimed at enhancing the model's focus on relevant features during reconstruction. Networks like FPENet [23] use an attention-based module to selectively emphasize important features and suppress less useful ones, improving the clarity and accuracy of the segmented output. This approach is particularly effective in complex scenes where differentiation between objects and background can be challenging.

Each of these techniques provides a different balance of accuracy, computational efficiency, and resource requirements. They reflect the diverse strategies employed by researchers to tackle the challenges inherent in real-time semantic segmentation, particularly in resource-constrained environments. The choice of technique often depends on the specific requirements and constraints of the application at hand, such as the need for speed over resolution or vice versa.

III. Deep learning models under investigation

This section provides a detailed overview of the ten state-of-the-art deep learning-based image segmentation models that were evaluated in this study for the task of food localization.

DABNet [16] addresses the trade-off between accuracy and inference speed essential for autonomous systems by introducing a Depth-wise Asymmetric Bottleneck module. This module utilizes depth-wise asymmetric and dilated convolutions to enhance computational efficiency. DABNet demonstrates a balance between speed and precision, making it a viable option for real-time semantic segmentation.

ESNet [24] and **EDANet** [17] both focus on enhancing the efficiency of convolutional operations. ESNet uses factorized convolutions and a symmetric network design to reduce computational demand while maintaining high accuracy. In contrast, EDANet leverages an asymmetric convolution structure and dense connectivity to optimize both computational cost and model size for high-speed inference.

ENet [19], designed for tasks requiring low latency, is another prominent architecture that significantly reduces the computational burden compared to traditional methods. It achieves this by focusing on early downsampling, allowing it to be much faster and less resource-intensive while maintaining reasonable accuracy.

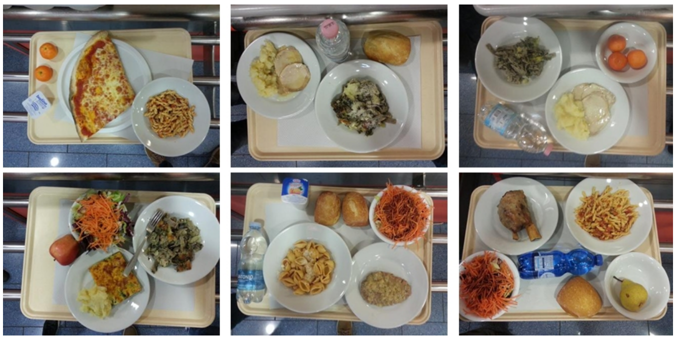

Fig. 2. Examples of tray images in the UNIMIB2016 dataset.

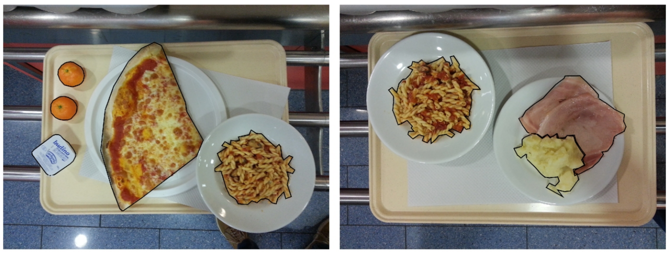

Fig. 3. Examples food annotations in the UNIMIB2016 dataset.

CGNet [26] proposes a Context Guided Network utilizing a novel block that efficiently processes local and global context information, which is crucial for accurate semantic segmentation. This model is particularly designed for mobile devices, offering a substantial reduction in parameter count while improving segmentation accuracy.

FastSCNN [20] and **ContextNet** [21] both introduce innovative methods to combine low-level feature extraction with high-level semantic information efficiently. FastSCNN employs a dual-branch approach that merges features at different resolutions, optimizing both accuracy and computational speed. Similarly, ContextNet utilizes a pyramid representation to balance detail and context, achieving effective segmentation at higher frame rates.

FSSNet [25], **FPENet** [23], and **LEDNet** [22] each propose solutions to optimize the trade-offs between model complexity, inference speed, and accuracy. FSSNet uses a factorized architecture with dilated convolutions to enhance the field of view without excessive parameter increase. FPENet and LEDNet focus on encoding multi-scale contextual features and employing novel network components like attention mechanisms to improve both performance and efficiency.

IV. Experimental setup

A. Dataset

As the evaluation dataset we consider the UNIMIB2016 dataset [27], that is widely used in several research works. The dataset is publicly available at http://www.ivl.disco.unimib.it/activities/food-recognition/ and is designed to support the development and evaluation of food recognition algorithms, particularly for dietary monitoring applications in canteen

TABLE I
MAIN CHARACTERISTICS OF STATE-OF-THE-ART MODELS FOR EFFICIENT IMAGE SEGMENTATION.

Model	**Architecture**			**Layers**		**Upsampling technique**	**Convolution**		
	Encoder	Decoder	dual path	activation func.	Residual Layer		dilated	depth-wise	factoriz.
ESNet [24]	ResNet [15]	Custom		ReLU	PFCU	Transposed Conv.	✓		✓
ENet [19]	Custom	Custom		PReLU	BottleNeck	Transposed Conv.	✓		✓
FSSNet [25]	ResNet [15]	Custom		PReLU	BottleNeck-1D	Bilinear + Transposed	✓		✓
FPENet [23]	Custom	Custom		ReLU	BottleNeck-1D	MEU (Bilinear + attention)	✓	✓	
LEDNet [22]	ResNet [15]	Custom		ReLU	BottleNeck-1D	Bilinear	✓		✓
DABNet [16]	Custom	No decoder (/3)		ReLU	Non-BottleNeck-1D		✓	✓	✓
EDANet [17]	Custom	No decoder (/8)		ReLU	BottleNeck-1D		✓		✓
CGNet [26]	Custom	No decoder (/8)		PReLU	Non-BottleNeck		✓	✓	
FastSCNN [20]	MobileNet-V2 [11]	No decoder (/1)	✓	ReLU	BottleNeck	Merge with full res. branch		✓	
ContextNet [21]	MobileNet-V2 [11]	No decoder (/1)	✓	ReLU6	BottleNeck	Merge with full res. branch	✓	✓	

TABLE II
COMPARISON OF SEGMENTATION ACCURACY (JACCARD INDEX) AND COMPUTATIONAL REQUIREMENTS (MAC) FOR EFFICIENT SEGMENTATION MODELS IN FOOD LOCALIZATION ON THE UNIMIB2016 DATASET.

Model	**Jaccard Index ↑**	**MAC (M) ↓**
DABNet [16]	93.50%	1314
ESNet [24]	93.44%	3382
EDANet [17]	93.41%	1115
ENet [19]	92.81%	616
CGNet [26]	92.64%	884
FastSCNN [20]	91.96%	221
ContextNet [21]	91.90%	222
FSSNet [25]	91.72%	564
FPENet [23]	89.99%	192
LEDNet [22]	83.99%	1451

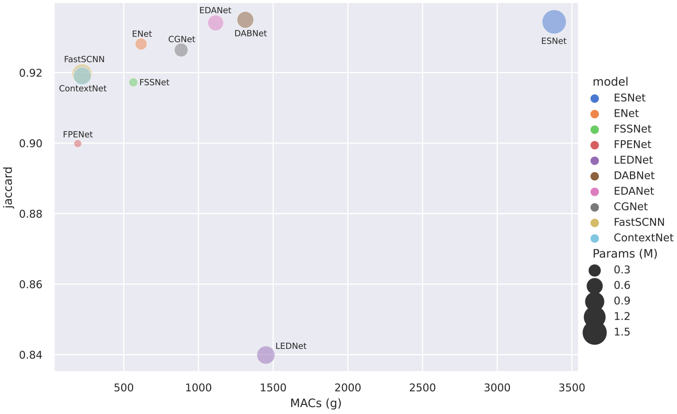

Fig. 4. Graph illustrating the performance of the efficient segmentation models under investigation, applied to food localization on the UNIMIB2016 dataset. The plot shows a general linear relationship between MACs (Multiply-Accumulate Operations) and segmentation performance. Notably, LEDNet deviates from this trend with lower-than-expected performance, while ESNet exhibits higher MACs with minimal performance gains. Additionally, the size of each circle represents the number of parameters in the model, indicating the model complexity.

settings. It comprises 1,027 images of canteen trays, each presenting multiple food items in various arrangements, for a total of 3,616 food instances across 73 distinct food classes. Figure 2 shows some example of tray images in the dataset. Each food item within the dataset has been meticulously segmented using polygonal boundaries to create high-quality ground truth annotations (see Figure 3).

To adapt this dataset for the specific task of food and non-food detection, binary masks have been created that differentiate food items from non-food background elements on each tray. This modification facilitates the training of segmentation models that can accurately identify and isolate food regions from the surrounding environment.

B. Training

The training process for our food segmentation models employs the Adam optimizer. We set the learning rate to 0.0001, and we optimize the models using binary cross-entropy as the loss function.

C. Metrics

To evaluate the performance of our segmentation models, we use the binary Jaccard index, also known as the Intersection over Union (IoU) metric. Formally, it is:

$$\text{Jaccard Index (IoU)} = \frac{\sum_{i=1}^{N}(p_i \wedge g_i)}{\sum_{i=1}^{N}(p_i \vee g_i)} \qquad (1)$$

where p_i represents the predicted set of food pixels and g_i is the ground truth of the i-th sample.

D. Hardware

The experiments were conducted on a machine equipped with an Intel Core i7 processor and an NVIDIA TITAN Xp graphics card, which features 12 GB of dedicated memory.

V. RESULTS

The results of this evaluation are summarized in Table II, which lists the Jaccard Index scores for each model tested on the UNIMIB2016 dataset. Additionally, Figure 5 visually represents these models in a bidimensional space where the x-axis shows the Multiply-Accumulate (MACs) operations and the y-axis the Jaccard Index performance. The size of each dot is proportional to the number of parameters in the model, providing a clear visualization of the trade-offs between computational complexity and segmentation accuracy.

DABNet emerged as the top-performing model with a Jaccard Index of 93.50%, despite having a lower MACs value compared to both ESNet and LEDNet. This highlights DABNet's efficiency in achieving high segmentation accuracy with comparatively fewer computational resources.

The analysis reveals that models utilizing bottleneck and non-bottleneck residual layers, such as ENet and ESNet, perform differently. ENet, which uses bottleneck layers, and DABNet, employing non-bottleneck-1D blocks, demonstrate that less restrictive feature processing can sometimes yield better performance. Moreover, the inclusion of dilated convolutions in models like SQNet and ERFNet enhances their ability to capture detailed spatial hierarchies effectively, which is reflected in their high performance.

Conversely, models designed with a strong focus on efficiency, such as LEDNet and FastSCNN, show some of the lowest performance metrics. This underscores a notable trade-off between efficiency and segmentation accuracy, highlighting the challenge of balancing these aspects in practical applications.

Furthermore, ContextNet demonstrates efficient performance with its pyramid representation for capturing context at various scales, offering a good balance between speed and accuracy. FPNet, which uses a feature pyramid to encode multi-scale contextual features, shows promise but may require further optimization to improve its efficiency and effectiveness in segmenting detailed food items.

In conclusion, while advanced architectural features like depth-wise and dilated convolutions offer significant computational advantages, they must be carefully balanced to maintain high performance in segmentation tasks. For real-time applications in environments like canteens, where both efficiency and accuracy are critical, the choice of model architecture might lean towards those like EDANet or ENet and CGNet which provide a good balance. Future work could explore combining these efficient designs with lightweight attention mechanisms or advanced upsampling techniques to further enhance both speed and performance.

VI. Conclusion

This study assessed ten state-of-the-art image segmentation models on the UNIMIB2016 dataset designed for food recognition in canteen environments, identifying a balance between efficiency and accuracy as crucial for real-time applications.

DABNet achieves the best accuracy score in terms of Jaccard Index. It is less efficient when compared to other models like ESNet and EDANet that obtained comparative accuracy with fewer parameters. These models can be valid alternatives if efficiency is of paramount importance, such as in scenarios requiring the use of compact edge devices. Our analysis and comparison will help researchers and practitioners to select models offering the best trade-off in terms of food localization accuracy and computation efficiency.

Future research should explore several directions to address the identified limitations and enhance the generalizability and applicability of the findings. Firstly, we plan to perform the same assessment on additional datasets that include a wider variety of food types, tray arrangements, and lighting conditions to thoroughly evaluate the generalization capability of the compared methods. Future research should also explore hybrid architectures that combine efficient and self-expanding convolution techniques, and integrate attention mechanisms to enhance model focus on salient features. Expanding the dataset to include diverse food types and environments, alongside testing models in real-world settings, will help improve the robustness and applicability of segmentation technologies.

Acknowledgement

Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for tender No. 341 of 15 March 2022 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU. Project code PE00000003, Concession Decree No. 1550 of 11 October 2022 adopted by the Italian Ministry of University and Research, CUP D93C22000890001, Project title "ON Foods - Research and innovation network on food and nutrition Sustainability, Safety and Security – Working ON Foods".

This work was partially supported by the MUR under the grant "Dipartimenti di Eccellenza 2023-2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

References

[1] S. O'Halloran, G. Eksteen, M. Gebremariam, and L. Alston, "Measurement methods used to assess the school food environment: a systematic review," *International journal of environmental research and public health*, vol. 17, no. 5, p. 1623, 2020.

[2] S. Aslan, G. Ciocca, D. Mazzini, and R. Schettini, "Benchmarking algorithms for food localization and semantic segmentation," *Int. Journal of Machine Learning and Cybernetics*, vol. 11, pp. 2827–2847, 2020.

[3] N. Martinez-Perez, L. E. Torheim, N. Castro-Díaz, and M. Arroyo-Izaga, "On-campus food environment, purchase behaviours, preferences and opinions in a norwegian university community," *Public Health Nutrition*, vol. 25, no. 6, pp. 1619–1630, 2022.

[4] C. Wanjek, *Food at work: Workplace solutions for malnutrition, obesity and chronic diseases*. International Labour Organization, 2005.

[5] G. Ciocca, P. Napoletano, and R. Schettini, "Food recognition and leftover estimation for daily diet monitoring," in *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops*, ser. Lecture Notes in Computer Science, vol. 9281. Springer, 2015, pp. 334–341.

[6] S. Aslan, G. Ciocca, and R. Schettini, "Semantic food segmentation for automatic dietary monitoring," in *2018 IEEE 8th Int. Conference on Consumer Electronics - Berlin (ICCE-Berlin)*, 2018, pp. 1–6.

[7] L. García-Herrero, C. Costello, F. De Menna, L. Schreiber, and M. Vittuari, "Eating away at sustainability. food consumption and waste patterns in a us school canteen," *Journal of Cleaner Production*, vol. 279, p. 123571, 2021.

[8] M. Buzzelli, G. Ciocca, P. Napoletano, and R. Schettini, "Analyzing and recognizing food in constrained and unconstrained environments," in *Proceedings of the 3rd Workshop on AIxFood*, 2021, pp. 1–5.

[9] S. Bianco, M. Buzzelli, G. Chiriaco, P. Napoletano, and F. Piccoli, "Food recognition with visual transformers," in *2023 IEEE 13th International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*. IEEE, 2023, pp. 82–87.

[10] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[11] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[12] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[13] D. Haase and M. Amthor, "Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved mobilenets," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 600–14 609.

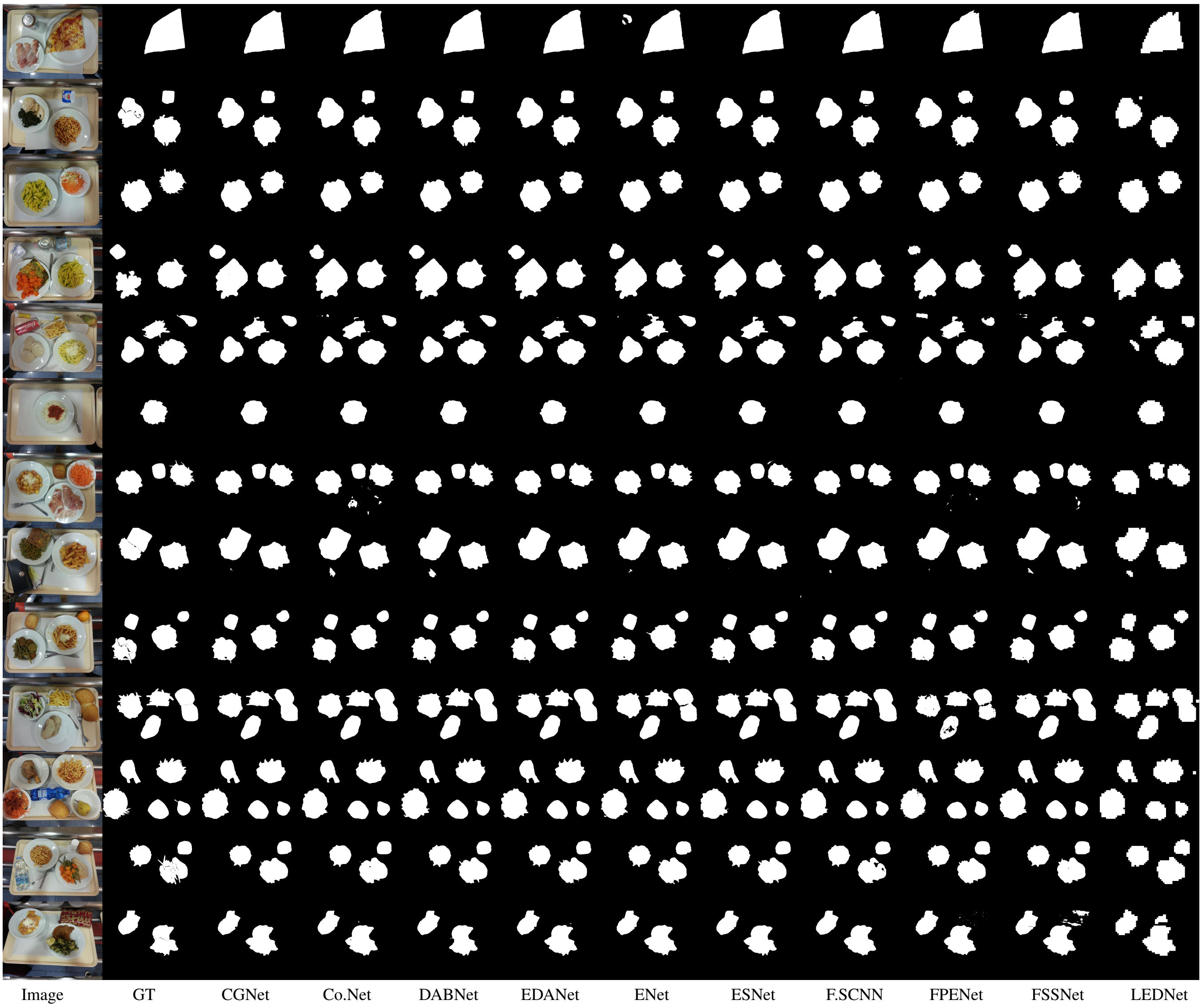

Fig. 5. Visual results of several efficient segmentation models applied to the task of food localization on the UNIMIB2016 dataset.

[14] M. Wang, B. Liu, and H. Foroosh, “Factorized convolutional neural networks,” in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 545–553.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] G. Li, I. Yun, J. Kim, and J. Kim, “Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation,” *arXiv preprint arXiv:1907.11357*, 2019.

[17] S.-Y. Lo, H.-M. Hang, S.-W. Chan, and J.-J. Lin, “Efficient dense modules of asymmetric convolution for real-time semantic segmentation,” in *Proceedings of the 1st ACM International Conference on Multimedia in Asia*, 2019, pp. 1–6.

[18] V. Dumoulin and F. Visin, “A guide to convolution arithmetic for deep learning,” *arXiv preprint arXiv:1603.07285*, 2016.

[19] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, “Enet: A deep neural network architecture for real-time semantic segmentation,” *arXiv preprint arXiv:1606.02147*, 2016.

[20] R. P. Poudel, S. Liwicki, and R. Cipolla, “Fast-scnn: Fast semantic segmentation network,” *arXiv preprint arXiv:1902.04502*, 2019.

[21] R. P. Poudel, U. Bonde, S. Liwicki, and C. Zach, “Contextnet: Exploring context and detail for semantic segmentation in real-time,” *arXiv preprint arXiv:1805.04554*, 2018.

[22] Y. Wang, Q. Zhou, J. Liu, J. Xiong, G. Gao, X. Wu, and L. J. Latecki, “Lednet: A lightweight encoder-decoder network for real-time semantic segmentation,” in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 1860–1864.

[23] M. Liu and H. Yin, “Feature pyramid encoding network for real-time semantic segmentation,” *arXiv preprint arXiv:1909.08599*, 2019.

[24] Y. Wang, Q. Zhou, J. Xiong, X. Wu, and X. Jin, “Esnet: An efficient symmetric network for real-time semantic segmentation,” in *Pattern Recognition and Computer Vision: Second Chinese Conference, PRCV 2019, Xi’an, China, November 8–11, 2019, Proceedings, Part II 2*. Springer, 2019, pp. 41–52.

[25] X. Zhang, Z. Chen, Q. J. Wu, L. Cai, D. Lu, and X. Li, “Fast semantic segmentation for scene perception,” *IEEE Transactions on Industrial Informatics*, vol. 15, no. 2, pp. 1183–1192, 2018.

[26] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, “Cgnet: A light-weight context guided network for semantic segmentation,” *IEEE Transactions on Image Processing*, vol. 30, pp. 1169–1179, 2021.

[27] G. Ciocca, P. Napoletano, and R. Schettini, “Food recognition: a new dataset, experiments, and results,” *IEEE journal of biomedical and health informatics*, vol. 21, no. 3, pp. 588–598, 2016.