

FOCUS: A FREQUENCY-ORIENTED AND CLASS-UNDERREPRESENTED SEMANTIC SEGMENTATION FRAMEWORK FOR FOOD IMAGES

006 **Anonymous authors**

007 Paper under double-blind review

ABSTRACT

013 Generating high-quality semantic segmentation results from food images remains
 014 a challenging task, particularly in the presence of complex boundaries and class
 015 imbalance. Existing methods often struggle with blurred edges and underperform
 016 on long-tailed categories, limiting their generalizability in practical scenarios. To
 017 address these issues, we propose FOCUS, a novel semantic segmentation frame-
 018 work designed to enhance boundary precision and improve underrepresented class
 019 recognition. Specifically, we introduce a frequency-based strategy that selectively
 020 processes high-frequency components via differential convolution and integrates
 021 explicit edge supervision during training. This enables the model to better cap-
 022 ture fine-grained boundary details and improves edge discriminability. To mit-
 023 iate class imbalance, we introduce an enhanced gradient allocation mechanism
 024 that applies targeted matching supervision to underrepresented categories, thereby
 025 amplifying learning signals for low-shot classes and improving classification ac-
 026 curacy. Extensive experiments on benchmark datasets, FoodSeg103, UECFood-
 027 PixComplete, and Food50Seg, show that FOCUS consistently outperforms exist-
 028 ing approaches in both boundary quality and underrepresented class performance,
 029 validating its architectural effectiveness and robust generalization capability.

1 INTRODUCTION

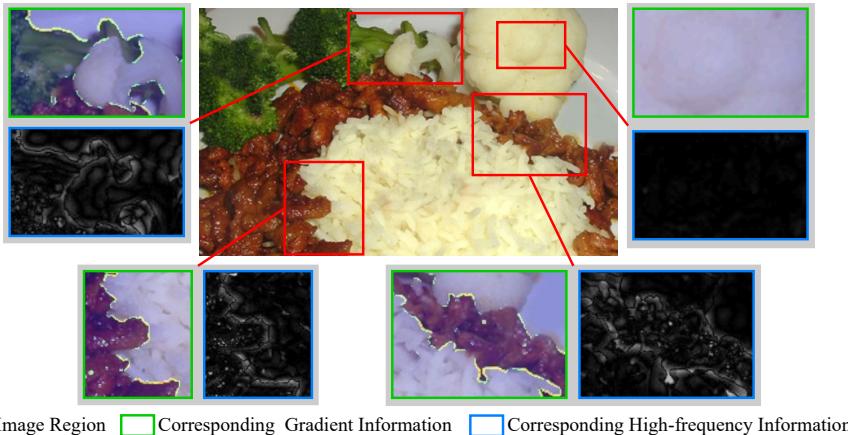


Figure 1: Gradient and frequency representation in food images. The green box shows gradient information, while the blue box highlights high-frequency components from the Fourier transform, with edges exhibiting abundant high-frequency content and high gradients.

Semantic segmentation of food images is a fundamental task in food computing, enabling various downstream applications such as food quality assessment Zhang et al. (2021b) and nutritional analysis He et al. (2020). Early approaches primarily relied on pixel-wise classification and performed

well when high-quality pixel-level annotations were available. However, in the food domain, such fine-grained annotations are often scarce, with datasets typically providing only coarse polygonal labels, which poses significant challenges for conventional methods. To overcome these limitations, recent research has increasingly adopted the mask classification paradigm Cheng et al. (2021), which directly predicts object-level masks. This strategy not only improves boundary localization but also enhances category recognition, thereby achieving higher-quality food image segmentation Min et al. (2019).

Despite recent advances, the inherent visual complexity and high intra-class variability of food items continue to present significant challenges for achieving robust segmentation. One major obstacle is the low quality of mask proposals near object boundaries. This issue arises primarily from the irregular and diverse shapes of food items Lin et al. (2017), as well as boundary interference caused by common food presentation styles such as overlapping and stacking. To address this, prior works have explored supervised attention mechanisms Chen et al. (2020) to enhance boundary awareness and refine mask predictions. However, these approaches typically rely on standard convolutional operations, which do not explicitly model boundary-specific structures. Because traditional convolutions operate over broad and unconstrained receptive fields, they often fail to effectively capture fine-grained edge details, resulting in suboptimal boundary localization. Therefore, improving mask proposal accuracy in boundary regions remains critical for advancing segmentation quality in food image analysis.

Another critical challenge observed in real-world scenarios is the pronounced class imbalance Cui et al. (2019), where certain categories are represented by only a limited number of training samples. These rare classes pose greater difficulties for model optimization, necessitating more effective strategies to fully exploit the available data during training. In food image generation, rare categories such as unique dishes or uncommon ingredients, are typically underrepresented, resulting in limited supervision and sparse gradient updates. Consequently, models struggle to learn stable and discriminative feature representations. As a result, the accuracy of mask classification for these classes is significantly compromised. Prior studies have attempted to mitigate class imbalance through approaches such as data augmentation Ghiasi et al. (2021) and loss re-weighting Wang et al. (2021). While these methods offer some improvement, they generally rely on explicit adjustments to the loss function or sample distribution and fail to provide richer and more diverse gradient signals. However, they rarely explore the optimization process from an algorithmic perspective. Consequently, they often fall short of fundamentally addressing the low mask classification accuracy associated with underrepresented classes.

To address aforementioned challenges of boundary ambiguity and underrepresented class learning, we propose a novel framework, FOCUS, which enhances both mask proposal quality and mask classification, thereby improving overall segmentation accuracy. From a frequency-aware perspective (Figure 1), boundary regions exhibit sharp gradient transitions and abundant high-frequency components, reflecting the complexity of object edges. To effectively capture these signals, we introduce a Frequency-Aware Boundary Modeling mechanism that replaces standard convolution with differential convolution to explicitly incorporate edge priors. And this design employs a supervised attention mechanism to capture subtle boundary variations while balancing high- and low-frequency information, thereby generating more accurate masks. To further improve classification performance for underrepresented categories, we introduce an extra matching strategy that assigns multiple queries to each foreground object, thereby expanding their effective receptive fields without modifying the overall training paradigm (Figure 2). This strategy provides more diverse gradient signals, enabling stronger supervision and more stable learning for these underrepresented classes.

Our main contributions are as follows:

- We propose FOCUS, a novel mask classification-based semantic segmentation framework tailored for food images. FOCUS effectively addresses the challenges of food image segmentation and achieves state-of-the-art (SOTA) performance on multiple benchmark datasets, as validated by extensive ablation studies and qualitative results.
- We present a new edge representation paradigm that combines frequency-domain selective filtering with differential convolution. This design decouples fine-grained edge signals from background noise, enabling more accurate boundary modeling and addressing long-standing difficulties in capturing intricate object boundaries.

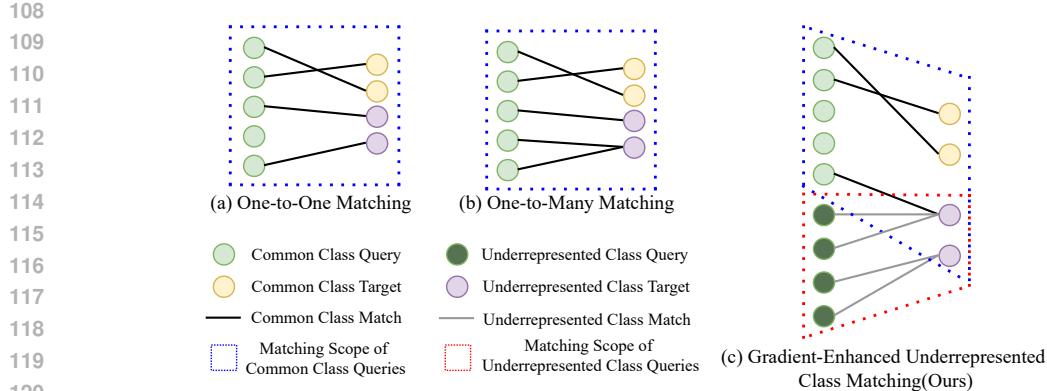


Figure 2: A comparison between existing matching strategies, including One-to-One Matching and One-to-Many Matching, and the Gradient-Enhanced Underrepresented Class Matching (Ours)

- We introduce a gradient-enhanced matching strategy that dynamically allocates extra queries to underrepresented and critical categories. This strategy improves classification accuracy and stabilizes training under class imbalance, thereby strengthening the overall effectiveness of the mask segmentation pipeline.

2 METHOD

2.1 OVERVIEW

Figure 3 illustrates the overall architecture of the proposed FOCUS framework, which consists of three main components: (1) the foundational backbone and pixel decoder for feature extraction, (2) the Frequency-Aware Boundary Modeling (FABM) that enriches boundary representations in critical feature maps, and (3) the Gradient-Enhanced Underrepresented Class Matching (GEUCM), which allocates extra queries to underrepresented classes to enhance classification accuracy.

2.2 FREQUENCY-AWARE BOUNDARY MODELING

Boundary regions typically exhibit sharp gradient transitions and rich high-frequency components, reflecting the inherent complexity of object edges. To effectively capture these signals, we design a Frequency-Aware Boundary Modeling (FABM) module, inspired by concepts from FADC Chen et al. (2024a) and DEA-Net Chen et al. (2024b). However, unlike the FADC and DEA-Net methods that perform fine-grained frequency decomposition, our approach simplifies frequency division into high and low frequencies, processing only the high-frequency components for further refinement. FABM primarily consists of a Frequency-Differentiated Block (FDB) and a convolutional layer.

Frequency Difference Block FDB primarily comprises two components: frequency-based feature selection and difference convolution. By isolating high-frequency signals and capturing pixel-level changes through difference-structured kernels, FDB effectively combines local detail sensitivity with global structural awareness for robust edge modeling.

Specifically, given an input feature map F_{in} , frequency-based feature selection first employs the Fast Fourier Transform (FFT) to map it from the spatial domain to the frequency domain. Then, based on a predefined frequency threshold α , two binary masks are constructed to represent the low-frequency and high-frequency components of the input feature, respectively. Finally, the frequency-domain feature is element-wise multiplied with these two masks and projected back into the spatial domain via the Inverse FFT, yielding the high- and low-frequency feature representations divided by the threshold.

$$F_{\text{low}} = \mathcal{F}^{-1}(M_{\text{low}} \mathcal{F}(F_{in})), \quad F_{\text{high}} = \mathcal{F}^{-1}(M_{\text{high}} \mathcal{F}(F_{in})), \quad (1)$$

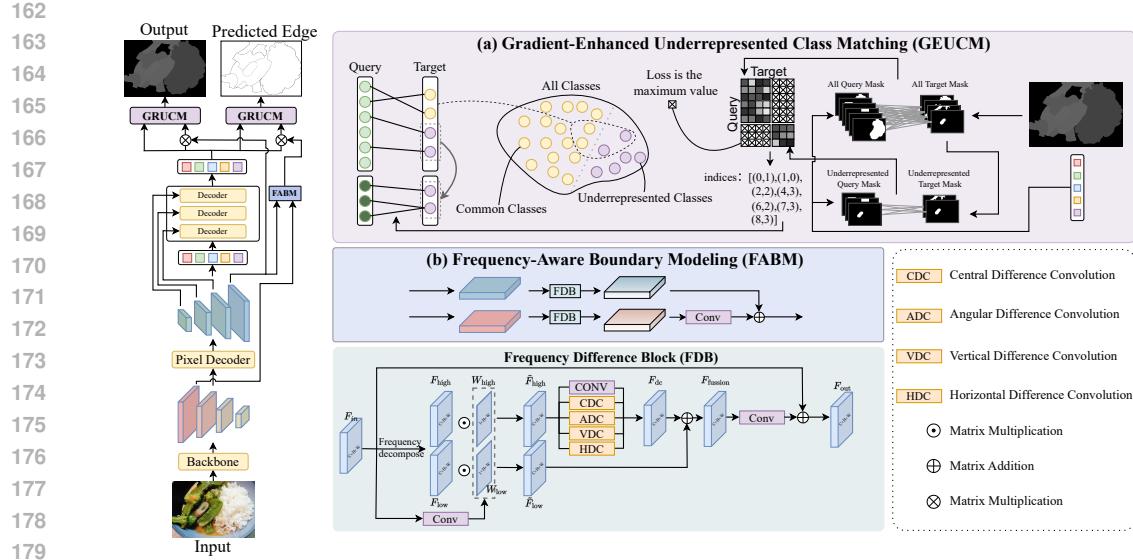


Figure 3: Architectural diagram of the proposed FOCUS framework. Given an input image, our model extracts multi-level features through a backbone and pixel decoder. These features, together with learnable queries, are fed into a Transformer decoder to produce the final segmentation output. To enhance the mask proposal quality and mask classification accuracy, we accordingly introduce two key modules: (a) FABM, which strengthens boundary representations in critical feature maps, and (b) GEUCM, which allocates additional queries to underrepresented classes to boost classification accuracy.

$$M_{\text{low}}(u, v) = \begin{cases} 1, & \max(|u|, |v|) < \alpha \\ 0, & \text{otherwise} \end{cases}, \quad M_{\text{high}}(u, v) = 1 - M_{\text{low}}(u, v) \quad (2)$$

Here, \mathcal{F} and \mathcal{F}^{-1} denote the FFT and inverse FFT, respectively. M_{low} and M_{high} are binary masks used to extract the corresponding low- and high-frequency components, and u and v indicate the frequency coordinates along the width and height dimensions. The resulting F_{low} and F_{high} represent the low-frequency and high-frequency spatial-domain features, respectively.

Next, a learnable convolutional layer is applied to selectively retain or suppress specific frequency components. By convolving the input feature F_{in} with this convolutional layer, two weighting maps, W_{low} and W_{high} , are generated to modulate the low-frequency and high-frequency features, respectively.

$$W_{\text{low}} = \text{Conv}_{\text{low}}(F_{in}), W_{\text{high}} = \text{Conv}_{\text{high}}(F_{in}). \quad (3)$$

$$\tilde{F}_{\text{low}} = W_{\text{low}} \odot F_{\text{low}}, \tilde{F}_{\text{high}} = W_{\text{high}} \odot F_{\text{high}}. \quad (4)$$

Here, $W_{\text{low}}, W_{\text{high}} \in \mathbb{R}^{1 \times H \times W}$ are represent spatial selection maps learned from the input feature F_{in} through Conv_{low} and $\text{Conv}_{\text{high}}$, respectively. The weighted low- and high-frequency features are denoted as \tilde{F}_{low} and \tilde{F}_{high} , respectively.

The weighted high-frequency feature \tilde{F}_{high} is further enhanced through a series of difference convolutions, which strengthen edge and fine-grained details, producing the feature F_{dc} .

$$F_{\text{dc}} = \text{DConv}(\tilde{F}_{\text{high}}), \quad (5)$$

The enhanced high-frequency feature is then fused with the weighted low-frequency feature \tilde{F}_{low} through element-wise addition to obtain the fused representation F_{fusion} .

$$F_{\text{fusion}} = F_{\text{dc}} + \tilde{F}_{\text{low}}. \quad (6)$$

Finally, F_{fusion} is further refined via a standard convolutional layer and combined with the original input feature F_{in} to produce the final output F_{out} .

$$F_{\text{out}} = \text{Conv}(F_{\text{fusion}}) + F_{\text{in}}. \quad (7)$$

Difference Convolution Difference convolution employs differential kernels to highlight local pixel variations, making it effective for capturing gradients, edges, and structural details Yu et al. (2021); Su et al. (2021). Different kernel configurations (e.g., horizontal, vertical, angular, and central) emphasize gradient changes along specific directions Yu et al. (2020); Chen et al. (2024b). Unlike prior studies that mainly focus on designing new variants, our work integrates these operations with frequency-based feature selection in the FABM module, enabling explicit boundary enhancement for segmentation. The difference convolution can be formulated as follows:

$$Y(i, j) = \sum_{k=-1}^1 \sum_{l=-1}^1 I(i+k, j+l) \cdot W(k, l), \quad (8)$$

where $Y(i, j)$ is the pixel value of the output image after difference convolution, $I(i+k, j+l)$ is the pixel value of the input image at the position $(i+k, j+l)$, and $W(k, l)$ is the weight of the kernel at the corresponding position. In difference convolution, the kernel is designed with a differential structure. For example, in a horizontal 3×3 convolution, only the left and right columns contain learnable parameters, the middle column is fixed to zero, and the parameters in the same row are constrained to be negatives of each other. This design highlights pixel differences along the horizontal direction, thereby enhancing gradients and edge features.

Formally, the kernel can be expressed as:

$$W = \begin{bmatrix} -a_1 & 0 & a_1 \\ -a_2 & 0 & a_2 \\ -a_3 & 0 & a_3 \end{bmatrix},$$

where a_1, a_2, a_3 are learnable parameters.

FABM Loss Design To further enhance the model’s ability to capture object contours, we introduce an additional boundary prediction branch alongside the original mask classification loss. The loss is defined as $L_{\text{FABM}} = L_{\text{edge}} + \lambda_{\text{cls}} L_{\text{cls}}$, where all loss terms in L_{FABM} are computed using the predicted and ground-truth edge maps. The boundary supervision follows the same binary cross-entropy and Dice formulation used for mask prediction: $L_{\text{edge}} = \lambda_{\text{ce}} L_{\text{ce}} + \lambda_{\text{dice}} L_{\text{dice}}$. By explicitly supervising boundary information, this branch encourages the model to focus on precise contour details, leading to clearer boundary representations and improved overall segmentation performance.

2.3 GRADIENT-ENHANCED UNDERREPRESENTED CLASS MATCHING

The following section introduces the definition of underrepresented classes specific to our task, and then elaborates on the design of the Gradient-Enhanced Underrepresented Class Matching module.

Definition of Underrepresented Class In real-world food image datasets, category distributions are typically imbalanced, with certain classes appearing rarely or consistently yielding poor segmentation performance. We collectively refer to these as *underrepresented classes* to facilitate targeted optimization.

To formalize this, consider a dataset with n food categories. We begin by computing the occurrence frequency of each category, defined as the number of training images in which the category appears divided by the total number of training images. We then identify the k least frequent categories as a candidate underrepresented class set. From this set, we exclude j categories that demonstrate strong model performance on a validation set, those whose IoU or accuracy exceed a predefined threshold m . The remaining $k - j$ categories form Category Set A. Next, we evaluate segmentation performance on the remaining $n - k$ categories and select the j worst-performing ones based on the same metric, forming Category Set B. The final *underrepresented class* set is defined as the union of Category Sets A and B, capturing both rare and underperformed categories. The remaining categories are referred to as *common* categories.

Gradient-Enhanced Underrepresented Class Matching (GEUCM)

270 In the One-to-One matching scheme (Figure 2(a)), m queries generate m predicted masks, which are
 271 matched to n ground-truth masks using a cost matrix that computes pairwise losses. The Hungarian
 272 algorithm then solves the resulting $m \times n$ assignment problem. The remaining $(m - n)$ queries
 273 (typically close to 100 vs. few food objects) are labeled “no object,” introducing a large number
 274 of negative samples that dilute gradients for underrepresented classes. Although recent studies Li
 275 et al. (2024) have explored the One-to-Many matching (Figure 2(b)), assigning extra matches purely
 276 based on class frequency may degrade the performance of common classes, leading to suboptimal
 277 results in food segmentation.

278 To address this issue, we build on two key observations. First, rare classes contain few positive
 279 instances, and their gradients are easily overwhelmed by abundant negatives Li et al. (2024), in-
 280 creasing their match count strengthens their gradient signals and improves optimization balance.
 281 Second, queries naturally exhibit spatial biases Carion et al. (2020). For example, underrepresented
 282 classes in food images often occur in spatially predictable regions. Allocating dedicated queries to
 283 these classes promotes class-specific spatial priors, strengthening both their detectability and seg-
 284 mentability. Motivated by these insights, we propose Gradient-Enhanced Underrepresented Class
 285 Matching (GEUCM), a simple yet effective strategy that assigns extra matching queries for under-
 286 represented categories. This enhances their gradient signals without compromising the learning of
 287 common categories, improving gradient distribution during training and ultimately boost overall
 288 segmentation performance.

289 Our approach consists of two steps. First, we apply One-to-One Matching for m common queries,
 290 covering all ground-truth masks, comprising n common and u underrepresented masks. Second, we
 291 employ One-to-Many Matching (Figure 2(c)) for e extra queries targeting the u underrepresented
 292 classes, where each target is matched with k queries, with k defined as a hyperparameter. The
 293 matching strategies of these two steps operate independently, yielding a total of $m + u \times (k + 1)$
 294 matched pairs. The full procedure is outlined in Algorithm 1.

Algorithm 1 Gradient-Enhanced Underrepresented Class Matching

Input: A query set consisting of m common queries for all ground-truth masks and e extra queries
 297 for underrepresented ground-truth masks; a mask set containing n common ground-truth masks and
 298 u underrepresented ground-truth masks (for a single image inference).

Parameter: k (number of extra matches per underrepresented ground-truth masks)

Output: $n + u \times (k + 1)$ matched query-mask pairs

```

301 1: Initialize:  $\mathcal{M}_1 \leftarrow \emptyset, \mathcal{M}_2 \leftarrow \emptyset$ 
302 2: Step 1: One-to-One Matching
303 3: Compute a matching loss matrix between the  $m$  common queries and all  $(n + u)$  ground-truth
304 4: Use the Hungarian algorithm to find the minimum-cost one-to-one assignment and store the
305 5: Step 2: Gradient-Enhanced One-to-Many Matching
306 6: for  $t = 1$  to  $k$  do
307 7:   Compute a matching loss matrix between the remaining  $e$  extra queries and  $u$  underrepre-
308 8:   sented ground-truth masks.
309 9:   Apply the Hungarian algorithm to select a match per underrepresented ground-truth masks
310 10:  and update  $\mathcal{M}_2$ ; remove the assigned queries from the extra query pool.
311 11: end for
312 12: return  $\mathcal{M}_1 \cup \mathcal{M}_2$ .
```

 315

3 EXPERIMENTS

 316

 317

3.1 DATASETS

 318

321 **FoodSeg103** Wu et al. (2021) is a large-scale dataset for food image semantic segmentation, con-
 322 taining 7118 high-quality images with pixel-level annotations across 103 common food categories.
 323 It presents challenges such as diverse categories, complex boundaries, and severe class imbalance,
 324 making it a strong benchmark for segmentation performance.

UECFoodPixComplete Okamoto & Yanai (2021) is a refined dataset for food classification and segmentation, comprising 10,000 images from 102 Japanese food categories. It improves upon the original UEC-FoodPix by manually correcting approximately 9,000 masks, addressing limitations of bounding box-based annotations and improving segmentation precision.

Food50Seg Aslan et al. (2020) includes 5,000 images across 50 food categories, with 100 pixel-level annotated images per class. It also provides 120,000 augmented images with various distortions, such as lighting variation, JPEG compression, and blur, enabling robustness evaluation.

3.2 EVALUATION METRICS

Following prior work, we evaluate model performance using mean Intersection over Union (mIoU), overall accuracy (aAcc), and mean accuracy (mAcc).

$$\text{mIoU} = \frac{1}{C} \sum_{i=1}^C \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i}, \quad \text{aAcc} = \frac{\sum_{i=1}^C \text{TP}_i}{\sum_{i=1}^C (\text{TP}_i + \text{FN}_i)}, \quad \text{mAcc} = \frac{1}{C} \sum_{i=1}^C \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}. \quad (9)$$

Here, C represents the total number of classes, and the TP_i denotes the number of pixels correctly classified as class i ; FP_i refers to the number of pixels incorrectly predicted as class i ; and FN_i represents the number of pixels whose true label is class i but are misclassified as other classes.

Method	FoodSeg103			UECFoodPixComplete		
	mIoU↑	aAcc↑	mAcc↑	mIoU↑	aAcc↑	mAcc↑
FPNKirillov et al. (2019)	27.28	75.23	36.7	53.34	83.93	67.21
DeeplabV3+Chen et al. (2018)	31.04	79.32	42.66	55.50	66.80	-
CCNetHuang et al. (2019)	28.6	78.90	47.80	64.62	87.96	77.50
UpernetXiao et al. (2018)	39.8	82.02	52.37	59.35	86.64	74.44
SETRZheng et al. (2021)	45.1	83.53	57.44	65.41	86.02	77.36
SegformerXie et al. (2021)	38.67	80.29	48.96	64.89	88.25	76.98
KNetZhang et al. (2021a)	40.18	82.01	52.01	64.88	87.63	76.94
FADCChen et al. (2024a)	50.77	85.10	63.14	-	-	-
VimZhu et al. (2024)	42.58	81.73	54.42	65.17	86.01	75.76
VMambaLiu et al. (2024)	44.38	82.94	55.99	64.18	86.37	75.63
GCNetYang et al. (2025)	46.25	83.94	56.21	65.33	87.46	76.73
SegMAN Fu et al. (2025)	50.14	85.22	62.11	68.24	88.14	80.39
STPPNWang et al. (2022)	40.3	82.13	53.98	-	-	-
Window AttentionDong et al. (2021)	31.4	77.62	40.3	-	-	-
ReLeM-CCNetWu et al. (2021)	47.10	85.50	59.50	-	-	-
FoodSAMLan et al. (2023)	46.42	84.10	58.27	66.14	88.47	78.01
CANetDong et al. (2024)	37.21	-	47.33	<u>68.90</u>	-	<u>80.60</u>
FDSNet Xiao et al. (2025)	47.34	-	60.04	-	-	-
Mask2FormerCheng et al. (2022)	51.37	86.40	64.39	68.39	89.34	78.68
Ours	52.84	86.89	65.72	71.24	90.24	81.09

Table 1: Performance comparison of different methods on FoodSeg103(%) and UECFoodPixComplete(%). Best results are in bold, second-best are underlined.

3.3 IMPLEMENTATION DETAILS

All experiments are implemented in PyTorch and conducted on two NVIDIA RTX 3090 GPUs with a total batch size of 4 (2 images per GPU). The model is trained using the AdamW optimizer with an initial learning rate of 0.0001, following a cosine annealing schedule that gradually decays the learning rate to zero. α , k , and u are set to 0.5, 2, and 20, respectively, while the threshold m is set to 0.3, 0.5, and 0.7 on the FoodSeg103, UECFoodPixComplete, and Food50Seg datasets, respectively. Unless otherwise specified, all hyperparameters are aligned with those in the Mask2Former baseline.

Method	Food50Seg		
	mIoU↑	aAcc↑	mAcc↑
DeeplabV3+Chen et al. (2018)	70.01	84.94	81.26
CCNetHuang et al. (2019)	74.25	82.56	81.36
UpernetXiao et al. (2018)	71.29	83.65	82.59
SETRZheng et al. (2021)	84.48	91.71	89.35
SegformerXie et al. (2021)	85.14	90.94	88.56
KNetZhang et al. (2021a)	83.74	92.27	90.18
VimZhu et al. (2024)	82.95	88.85	90.63
VMambaLiu et al. (2024)	84.18	91.11	90.92
GCNetYang et al. (2025)	86.38	90.34	91.20
SegMAN Fu et al. (2025)	89.57	95.48	95.37
Mask2FormerCheng et al. (2022)	92.77	96.45	96.17
Ours	93.42	96.78	97.02

Table 2: Performance comparison of different methods on Food50Seg(%).



Figure 4: Visualization of food image segmentation across different segmentation frameworks, including FOCUS. FOCUS demonstrates better performance in terms of segmentation edge precision and mask classification accuracy..

3.4 COMPARISON RESULTS

Quantitative Comparison We compare our proposed model against both recent SOTA general-purpose segmentation models and methods tailored to food image segmentation. As shown in Table 1 and Table 2, our method consistently outperforms all baselines, achieving new SOTA results across all datasets and metrics. While Mask2Former ranks as the second-best in most cases, our model surpasses it by a notable margin. Specifically, our model achieves mIoU scores of 71.24% on UECFoodPixComplete, 93.42% on Food50Seg, and 52.84% on FoodSeg103, representing a +1.47% improvement over Mask2Former on the most challenging dataset, FoodSeg103. This dataset presents a particularly challenging scenario due to its higher visual complexity and severe category imbalance, making these improvements even more significant.

Qualitative Comparison Figure 4 presents the qualitative comparison across different methods. Our approach demonstrates superior classification accuracy, particularly in challenging scenarios. In the first column, it accurately distinguishes overlapping components such as *cucumber* and *chicken*.

In the second column, our model correctly identifies *egg*, while GCNet and Mask2Former misclassify it as *pie* and *cake*, respectively. In the final column, our method successfully recognizes *Chinese butter* and precisely delineates the *spoon* within it, highlighting its strong capability in handling fine-grained targets and complex scenes.

3.5 ABLATION STUDY

In this section, we conduct ablation studies on FoodSeg103 to evaluate the effectiveness of the proposed components: FABM and GEUCM. As shown in Table 3a, both modules independently enhances model performance, and their combination yields the best results, showing clear synergy.

FABM	GEUCM	mIoU↑	aAcc↑	mAcc↑	Level1	Level2	mIoU↑	aAcc↑	mAcc↑
		51.37	86.40	64.39			51.37	86.40	64.39
✓		52.49	86.54	65.24	✓		51.85	86.02	64.24
	✓	52.24	86.55	64.40		✓	51.95	86.70	64.11
✓	✓	52.84	86.89	65.72	✓	✓	52.49	86.54	65.24

(a)

(b)

Table 3: (a) Ablation study of different modules on FoodSeg103(%). (b) Ablation study of FDB on FoodSeg103 (%). Level1 correspond to output features from the first stage of the backbone, while Level2 represents the final output feature from the Pixel Decoder. The first row of the table data (without checkmarks) represents the baseline.

α	mIoU↑	aAcc↑	mAcc↑
0.25	52.18	86.07	65.21
0.5	52.49	86.54	65.24
0.75	52.44	86.24	64.93

(a)

	mIoU↑	aAcc↑	mAcc↑
w/o L_{FABM}	51.94	86.38	64.72
w/ L_{FABM}	52.49	86.54	65.24

(b)

Table 4: (a) Ablation study of α on the FoodSeg103 (%). (b) Ablation study on the L_{FABM} the FoodSeg103 (%).

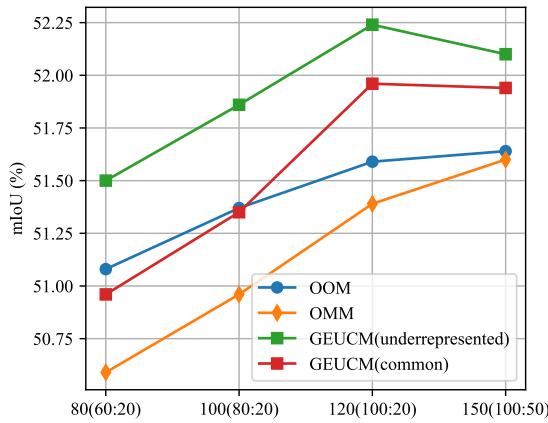


Figure 5: Comparison of different matching strategies in GEUCM on FoodSeg103 (%). The blue line denotes One-to-One Matching for all queries (OOM), the orange line denotes One-to-Many Matching for all queries (OMM), the green line represents GEUCM with extra queries for underrepresented classes, and the red line represents GEUCM with extra queries for common classes. The vertical axis indicates segmentation performance, while the horizontal axis shows the total number of queries. Numbers in parentheses specify the number of common and extra queries, respectively.

Ablation of FDB in FABM We evaluated the effectiveness of the Frequency Difference Block (FDB) in FABM by inserting it at different feature levels and analyzing its impact on segmentation performance. As shown in Table 3b, incorporating FDB at individual feature levels yields clear gains in both mIoU and aACC. Integrating FDB across all levels achieves the highest mIoU, the primary metric for segmentation, highlighting the benefit of full-level deployment.

Choice of α in FABM We perform an ablation study on the threshold α used for binary high–low frequency separation in FABM. As shown in Table 4a, the performance is not highly sensitive to the choice of α . The results indicate that boundary refinement cues in food images are largely captured

Method	mIoU	aAcc	mAcc
SegformerXie et al. (2021)	50.10	82.36	61.77
KNetZhang et al. (2021a)	52.21	82.41	67.34
VMambaLiu et al. (2024)	51.64	81.23	66.85
SegMANFu et al. (2025)	53.21	83.77	67.84
Mask2formerCheng et al. (2022)	54.96	84.19	68.55
Ours	55.39	84.31	69.02

Table 5: Performance comparison of different methods on ADE20K(%).

by high-frequency components, thus, further subdividing the frequency spectrum would introduce unnecessary complexity without meaningful gains. Therefore, we set α to 0.5 in all experiments.

Ablation of FABM Loss To validate the contribution of the FABM Loss, we conducted an ablation study that isolates this term. As shown in Table 4b, removing the L_{FABM} leads to a clear drop in in multiple metrics, confirming the effectiveness of the edge-aware constraint.

Choice of Matching Strategy in GEUCM As shown in Figure 5, our results demonstrate that, compared to other matching strategies, the proposed GEUCM(underrepresented) consistently delivers the best performance across various settings by introducing extra matching mechanisms exclusively for underrepresented classes. This highlights the effectiveness of our matching strategy.

Additionally, we observe that increasing the number of queries can somewhat enhance segmentation. However, this improvement is not guaranteed, when the proportion of extra matches becomes too large (i.e., an excessive number of underrepresented classes), performance may decline. This is likely because common classes are already optimized, offering limited room for further gains, and excessive attention may shift learning capacity away from the underrepresented classes.

3.6 COMPLEXITY ANALYSIS

We performed a comprehensive assessment of the computational efficiency of the proposed components. For FABM, both the FFT-based frequency decomposition and the multi-branch convolutional design are computationally lightweight, adding only approximately 4M parameters and 13 GFLOPs. For GEUCM, the additional cost primarily arises from the extra queries. With 20 extra queries (i.e., $u = 20$), GEUCM contributes fewer than 1M additional parameters and about 4 GFLOPs. In total, our method introduces roughly 5M additional parameters and 17 GFLOPs relative to the baseline, demonstrating that the performance gains are achieved with minimal computational overhead.

3.7 GENERALIZATION ANALYSIS

To assess our method’s generalization, we perform additional experiments on ADE20K Zhou et al. (2017). We compare our method with the several SOTA methods. As shown in Table 5, our approach achieves mIoU, aAcc, and mAcc scores of 55.4, 84.3, and 69.0, respectively. Relative to the second-best method, Mask2Former, this represents absolute improvements of 0.4%, 0.1%, and 0.5%, demonstrating that our method generalizes well beyond the food segmentation domain.

4 CONCLUSION

In this study, we present FOCUS, a semantic segmentation framework tailored for food images that enhances both the mask proposal quality and mask classification accuracy. Specifically, we introduce a frequency-oriented edge enhancement module that sharpens boundary details during mask proposal generation, and a Gradient-Enhanced Underrepresented Class Matching strategy that assigns extra queries to underrepresented classes, thereby improving classification accuracy. Extensive evaluations on three benchmark food-image datasets demonstrate the effectiveness of our approach in advancing segmentation performance.

540 REFERENCES

541

542 Sinem Aslan, Gianluigi Ciocca, Davide Mazzini, and Raimondo Schettini. Benchmarking algo-
543 rithms for food localization and semantic segmentation. *Int. J. Mach. Learn. Cybern.*, 11(12):
544 2827–2847, Jun. 2020.

545 Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and
546 Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on*
547 *computer vision*, pp. 213–229. Springer, 2020.

548 Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-
549 decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of*
550 *the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

551 Linwei Chen, Lin Gu, Dezhi Zheng, and Ying Fu. Frequency-adaptive dilated convolution for
552 semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*
553 *Pattern Recognition*, pp. 3414–3425, 2024a.

554 Xier Chen, Yanchao Lian, Licheng Jiao, Haoran Wang, YanJie Gao, and Shi Lingling. Supervised
555 edge attention network for accurate image instance segmentation. In *Computer Vision–ECCV*
556 *2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII*
557 *16*, pp. 617–631. Springer, 2020.

558 Zixuan Chen, Zewei He, and Zhe-Ming Lu. Dea-net: Single image dehazing based on detail-
559 enhanced convolution and content-guided attention. *IEEE transactions on image processing*, 33:
560 1002–1015, 2024b.

561 Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. Per-pixel classification is not all you
562 need for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision*
563 *and pattern recognition*, 2021.

564 Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-
565 attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF*
566 *conference on computer vision and pattern recognition*, 2022.

567 Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based
568 on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision*
569 *and pattern recognition*, pp. 9268–9277, 2019.

570 Xiaoxiao Dong, Wei Wang, Haisheng Li, and Qiang Cai. Windows attention based pyramid network
571 for food segmentation. In *2021 IEEE 7th International Conference on Cloud Computing and*
572 *Intelligent Systems (CCIS)*, pp. 213–217. IEEE, 2021.

573 Xiaoxiao Dong, Haisheng Li, Xiaochuan Wang, Wei Wang, and Junping Du. Canet: cross attention
574 network for food image segmentation. *Multimedia Tools Appl.*, 83(21):60987–61006, Dec. 2024.

575 Yunxiang Fu, Meng Lou, and Yizhou Yu. Segman: Omni-scale context modeling with state space
576 models and local attention for semantic segmentation. In *IEEE/CVF Conference on Computer*
577 *Vision and Pattern Recognition*, 2025.

578 Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and
579 Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation.
580 In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp.
581 2918–2928, 2021.

582 Jiangpeng He, Zeman Shao, Janine Wright, Deborah Kerr, Carol Boushey, and Fengqing Zhu. Multi-
583 task image-based dietary assessment for food recognition and portion size estimation. In *2020*
584 *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 49–54. IEEE,
585 2020.

586 Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet:
587 Criss-cross attention for semantic segmentation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*
588 *(ICCV)*, pp. 603–612, Oct. 2019.

594 Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmen-
 595 tation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 9404–9413, Jun.
 596 2019.

597

598 Xing Lan, Jiayi Lyu, Hanyu Jiang, Kun Dong, Zehai Niu, Yi Zhang, and Jian Xue. Foodsam: Any
 599 food segmentation. *IEEE Transactions on Multimedia*, 2023.

600 Shan Li, Lu Yang, Pu Cao, Liulei Li, and Huadong Ma. Frequency-based matcher for long-tailed
 601 semantic segmentation. *IEEE Transactions on Multimedia*, 2024.

602

603 Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense
 604 object detection. In *Proceedings of the IEEE international conference on computer vision*, pp.
 605 2980–2988, 2017.

606

607 Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin
 608 Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information
 609 processing systems*, 37:103031–103063, 2024.

610

611 Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. A survey on food computing.
 612 *Acm Computing Surveys (CSUR)*, 52(5):1–36, 2019.

613

614 Kaimu Okamoto and Keiji Yanai. Uec-foodpix complete: A large-scale food image segmentation
 615 dataset. In *Lect. Notes Comput. Sci.*, pp. 647–659, Jan. 2021.

616

617 Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu.
 618 Pixel difference networks for efficient edge detection. In *Proceedings of the IEEE/CVF interna-
 619 tional conference on computer vision*, pp. 5117–5127, 2021.

620

621 Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen,
 622 Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation.
 623 In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp.
 624 9695–9704, 2021.

625

626 Qiankun Wang, Xiaoxiao Dong, Ruimin Wang, and Hao Sun. Swin transformer based pyramid
 627 pooling network for food segmentation. In *2022 IEEE 2nd International Conference on Software
 628 Engineering and Artificial Intelligence (SEAI)*, pp. 64–68. IEEE, 2022.

629

630 Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven CH Hoi, and Qianru Sun. A large-scale
 631 benchmark for food image segmentation. In *Proceedings of the 29th ACM international confer-
 632 ence on multimedia*, pp. 506–515, 2021.

633

634 Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for
 635 scene understanding. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 418–434, Sep. 2018.

636

637 Zhiyong Xiao, Yang Li, and Zhaohong Deng. Food image segmentation based on deep and shallow
 638 dual-branch network. *Multimedia Systems*, 31(1):1–14, 2025.

639

640 Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Seg-
 641 former: Simple and efficient design for semantic segmentation with transformers. *Proc. NeurIPS*,
 642 34:12077–12090, Jun. 2021.

643

644 Guoyu Yang, Yuan Wang, Daming Shi, and Yanzhong Wang. Golden cudgel network for real-time
 645 semantic segmentation, 2025. URL <https://arxiv.org/abs/2503.03325>.

646

647 Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and
 648 Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In
 649 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5295–
 650 5305, 2020.

651

652 Zitong Yu, Yunxiao Qin, Hengshuang Zhao, Xiaobai Li, and Guoying Zhao. Dual-cross central
 653 difference network for face anti-spoofing. *arXiv preprint arXiv:2105.01290*, 2021.

654

655 Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image
 656 segmentation. *Proc. NeurIPS*, 34:10326–10338, Jun. 2021a.

648 Xiaolei Zhang, Jie Yang, Tao Lin, and Yibin Ying. Food and agro-product quality evaluation based
649 on spectroscopy and deep learning: A review. *Trends in Food Science & Technology*, 112:431–
650 441, 2021b.

651 Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei
652 Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from
653 a sequence-to-sequence perspective with transformers. In *Proc. IEEE/CVF Conf. Comput. Vis.
654 Pattern Recognit. (CVPR)*, pp. 6881–6890, Jun. 2021.

655 Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene
656 parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and
657 pattern recognition*, pp. 633–641, 2017.

658 Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vi-
659 sion mamba: efficient visual representation learning with bidirectional state space model. In
660 *Proceedings of the 41st International Conference on Machine Learning*, pp. 62429–62442, 2024.

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701