JOURNAL OF ADVANCED ELECTRICAL AND ELECTRONIC ENGINEERING

Vol 2 Issue 1 (2026)

http://jaeee.org/

Few Shot Retail Shelf Segmentation with Coarse Prototype Matching Episodes

Joonseo Hwang
Faculty of Engineering and Information Technology, University of Melbourne, Melbourne, Australia

Abstract
The rapid evolution of automated retail systems has necessitated the development of highly accurate and adaptable computer vision algorithms capable of parsing complex shelf environments. Retail shelf segmentation remains an exceptionally challenging domain due to extreme product density, severe intra-class variations, inter-class similarities, and dynamic illumination conditions. Furthermore, the constant introduction of new stock keeping units makes traditional fully supervised learning paradigms impractical, as they require exhaustive pixel-level annotations for every new product. To address these systemic bottlenecks, this paper proposes a novel coarse-to-fine prototype matching framework specifically tailored for episodic tasks in few-shot retail shelf segmentation. By leveraging the episodic training paradigm, the proposed architecture learns to generalize to unseen product categories using only a minimal number of annotated support samples. Our methodology introduces a dual-phase prototype extraction mechanism. Initially, a coarse global prototype establishes a baseline representation of the target category, facilitating broad localization and mitigating background noise. Subsequently, a fine-grained local prototype matching module captures intricate structural and textural variations across different spatial regions of the products. This hierarchical approach effectively bridges the semantic gap between the support and query sets in highly cluttered scenes. Extensive evaluations on a newly curated retail shelf segmentation dataset demonstrate that the proposed coarse-to-fine framework significantly outperforms existing state-of-the-art methods in both one-shot and five-shot scenarios. The findings indicate substantial improvements in boundary delineations and robustness against heavy occlusions, marking a critical advancement toward scalable and autonomous inventory management systems.

Keywords

Retail Automation, Few-Shot Segmentation, Prototype Matching, Computer Vision..

INTRODUCTION

The contemporary retail landscape is undergoing a profound transformation driven by the integration of artificial intelligence and advanced computer vision technologies. Among the most critical components of this paradigm shift is the automation of inventory management, planogram compliance verification, and autonomous checkout systems. At the core of these applications lies the fundamental task of retail shelf segmentation, which involves assigning a semantic category label to every pixel in an image of a retail shelf. Accurate segmentation enables systems to understand product distributions, detect out-of-stock items, and analyze consumer interactions with physical merchandise [1]. However, deploying robust semantic segmentation models in real-world retail environments presents a myriad of unique challenges that distinguish it from standard scene understanding tasks. Retail shelves are characterized by extreme product density, where hundreds of items are tightly packed, often leading to severe physical occlusion and visual clutter [2]. Additionally, packaging designs frequently exhibit high inter-class similarity, where different flavors or variants of a product share nearly identical visual features, while simultaneously displaying high intra-class variation due to different viewing angles, deformations in packaging, or irregular placement on the shelves [3]. Traditional deep learning approaches for semantic segmentation have historically relied on fully supervised learning frameworks, which demand massive amounts of pixel-perfect annotated training data to achieve satisfactory performance. While these models have achieved remarkable success on standardized benchmark datasets, their applicability to the retail sector is severely limited by the dynamic nature of retail inventories [4]. Supermarkets and retail chains constantly introduce new stock keeping units, change packaging designs seasonally, and rotate product placements [5]. Maintaining a fully supervised model would require continuous, exhaustive, and cost-prohibitive data collection and annotation efforts for every new product introduced to the market [6]. Consequently, there is an urgent and critical need for segmentation methodologies that can rapidly adapt to novel visual concepts with minimal human intervention. To overcome the limitations of fully supervised learning, the computer vision community has increasingly turned to few-shot learning, and more specifically, few-shot semantic segmentation. Few-shot segmentation aims to perform dense pixel-wise classification for novel categories using only a very small number of annotated examples, typically referred to as the support set [7]. This is achieved through an episodic training paradigm, where the model is exposed to numerous distinct tasks or episodes during the training phase, thereby learning a generalized matching function rather than specific category classifiers [8]. In the context of retail shelf segmentation, episodic training enables a model to learn how to extract representative features from a handful of support images of a new product and subsequently locate and segment all instances of that product in a complex query image [9]. Despite the theoretical promise of few-shot segmentation, existing frameworks struggle when applied directly to the retail shelf domain. Most conventional few-shot segmentation architectures rely on generating a single, globally pooled prototype vector to represent the entire target category [10]. While this global average pooling strategy is computationally efficient and works reasonably well for distinct, isolated objects in standard datasets, it inevitably discards crucial spatial and localized contextual information [11]. In a retail shelf environment, where products are subject to specular highlights, shadows, and partial occlusions by price tags or adjacent items, a single global prototype is often insufficiently discriminative to accurately parse the complex boundaries between closely grouped items [12]. The loss of fine-grained spatial details during the global pooling process often results in fragmented segmentation masks, false positives in background regions, and an inability to separate adjacent instances of similar products [13].

In response to these specific domain challenges, this paper introduces a comprehensive coarse-to-fine prototype matching architecture designed explicitly for retail shelf segmentation episodic tasks. The central premise of our approach is that effective matching in highly cluttered environments requires a hierarchical representation that captures both the overarching semantic identity of the product and its detailed localized features [14]. Our framework operates by first extracting deep feature representations from both the support and query images using a shared convolutional or transformer-based backbone network [15]. Following feature extraction, the system employs a coarse matching phase, where a globally pooled prototype is utilized to generate a preliminary, low-resolution activation map. This coarse map serves to suppress irrelevant background clutter and broadly identify potential regions of interest within the densely packed query image [16]. Following the coarse localization, the framework transitions to a fine-grained matching phase. Instead of relying on a single vector, this phase generates a dense set of localized prototypes from the foreground regions of the support image [17]. These local prototypes preserve the spatial geometry and specific structural variations of the product packaging [18]. A sophisticated pixel-to-prototype distance calculation is then performed, allowing each query pixel within the coarsely activated regions to match with the most relevant local features of the support item [19]. This dual-phase matching strategy ensures that the model is both highly robust to background distractors and exquisitely sensitive to the precise boundaries of the retail products. By integrating these mechanisms into an episodic training loop, the proposed system learns a highly adaptable and discriminative metric space that is fundamentally suited to the dynamic demands of modern retail analytics.

1.1 Contributions and Paper Structure

The primary contributions of this academic research are multi-faceted and address the fundamental bottlenecks in current visual retail analytics. First, we formally define the problem of retail shelf segmentation within an episodic few-shot learning framework, highlighting the specific statistical and visual challenges inherent to this domain. Second, we propose a novel coarse-to-fine prototype matching architecture that dynamically balances global semantic context with local structural fidelity, thereby overcoming the limitations of traditional single-prototype networks in cluttered environments. Third, we introduce a customized spatial alignment module that explicitly handles the scale and viewpoint variations frequently encountered in retail shelf photography. Finally, we provide extensive empirical validation of our methodology through rigorous experimental setups, demonstrating substantial improvements over existing baselines across multiple evaluation metrics. The remainder of this paper is structured to provide a logical and comprehensive exposition of our research. Section 2 delves into a detailed review of the relevant literature, tracing the evolution of deep learning in retail automation, the theoretical foundations of few-shot semantic segmentation, and the development of prototype matching strategies. Section 3 meticulously details the proposed methodology, explicitly describing the feature extraction mechanisms, the mathematical intuitions behind the coarse-to-fine matching process, and the optimization strategies employed during episodic training. Section 4 presents the experimental design, including dataset descriptions, evaluation protocols, and an in-depth discussion of both quantitative results and qualitative ablation studies. Section 5 concludes the paper by summarizing the core findings, discussing the practical implications for the retail industry, and outlining potential avenues for future scientific inquiry.

2. Literature Review and Background

The intersection of computer vision and retail automation has been a fertile ground for academic research and industrial innovation over the past decade. The continuous evolution of deep learning architectures has progressively enhanced the capability of automated systems to interpret complex retail environments. This section provides a comprehensive review of the foundational

technologies and recent advancements that contextualize our proposed coarse-to-fine prototype matching framework.

2.1 Deep Learning in Retail Scene Understanding

Initial attempts to automate retail shelf analysis relied heavily on classical computer vision techniques, such as scale-invariant feature transforms and histogram of oriented gradients, coupled with traditional machine learning classifiers like support vector machines [20]. While these methods provided early proofs of concept for automated planogram compliance and stockout detection, they were notoriously brittle. They required extensive manual parameter tuning and failed to generalize across different store lighting conditions or variations in product packaging [21]. The advent of convolutional neural networks fundamentally revolutionized this domain, providing feature representations that were learned directly from large-scale data rather than being hand-crafted by human engineers [22]. Early deep learning applications in retail primarily focused on object detection paradigms. Researchers utilized architectures based on region proposal networks and single-shot detectors to draw bounding boxes around individual products on a shelf [23]. These detection frameworks proved highly effective for simple counting tasks and identifying broad categories of out-of-stock regions. However, bounding boxes provide a fundamentally coarse representation of the visual scene. In densely packed retail environments, bounding boxes of adjacent, cylindrical, or irregularly shaped products frequently overlap, leading to significant ambiguity in downstream processing [24]. To achieve a more granular understanding of shelf layouts, including precise product boundaries and accurate volume estimations, the research focus shifted toward semantic and instance segmentation. Fully supervised semantic segmentation networks have demonstrated exceptional performance in parsing retail images. These architectures typically employ an encoder-decoder structure, where the encoder extracts hierarchical contextual features, and the decoder progressively recovers the spatial resolution to produce dense, pixel-wise predictions [25]. Despite their high accuracy, the fundamental limitation of these fully supervised models is their insatiable demand for annotated data. Creating pixel-perfect segmentation masks for retail shelves is an extraordinarily labor-intensive process, taking exponentially longer than drawing simple bounding boxes [26]. Given the high turnover rate of consumer packaged goods, maintaining the accuracy of a fully supervised segmentation model necessitates continuous and costly re-annotation and retraining cycles, rendering it commercially unviable for dynamic, large-scale retail operations [27].

2.2 Few-Shot Semantic Segmentation and Episodic Training

To circumvent the data dependency of fully supervised learning, the computer vision community developed the paradigm of few-shot semantic segmentation. The objective of this paradigm is to segment novel objects in a query image using only a tiny support set of annotated images, often just one or five examples [28]. The conceptual foundation of few-shot segmentation is meta-learning, or learning to learn. Instead of training a model to recognize specific classes, the model is trained over a vast number of episodes to learn a generic similarity function between the support annotations and the query features [29]. In a standard training episode, the dataset is divided into a support set and a query set containing classes from a predefined training split. The model extracts features from both sets and uses the ground-truth mask of the support set to isolate the foreground characteristics of the target class [30]. It then attempts to predict the segmentation mask for the query image based on its similarity to the extracted support features. The loss is computed using the ground-truth mask of the query image, and the network parameters are updated via backpropagation. Through this episodic training process, the model learns to generalize. During inference, the model can process entirely novel classes that were never seen during training, provided it is given a small support set defining the new concept [31].

While episodic training has shown great potential in standard natural image datasets, its application to retail shelf segmentation requires overcoming domain-specific hurdles. Standard few-shot datasets usually feature prominent, centrally located objects against relatively distinct backgrounds. In contrast, retail images exhibit dense packing, where the target product might appear dozens of times in a single query image, surrounded by visually similar distractors [32]. Consequently, algorithms that perform well on general datasets often experience a dramatic drop in performance when applied to retail episodes, struggling to differentiate between genuine target instances and closely related background products [33].

2.3 Evolution of Prototype Matching Strategies

The dominant architecture in few-shot semantic segmentation is based on prototype networks. The foundational idea is to compress the high-dimensional feature representation of the support object into a single, compact prototype vector through masked average pooling. This vector represents the central tendency or the overarching semantic concept of the target class. The network then computes the cosine similarity or Euclidean distance between this global prototype and every spatial feature vector in the query image. Pixels with high similarity scores are classified as belonging to the target class, while those with low scores are assigned to the background. While elegant and computationally efficient, the single prototype approach suffers from an information bottleneck. By averaging all spatial features into a single vector, the method inherently discards internal structural variance, texture gradients, and part-based relationships. For retail products, which often have specific text placements, colorful logos, and distinct packaging geometries, losing this fine-grained spatial information is detrimental. A single prototype might capture the dominant color of a product but fail to capture the specific logo, leading the network to falsely segment neighboring products of the same color. To address the limitations of single prototypes, recent research has explored multi-prototype representations. Some methods employ clustering algorithms on the foreground features of the support image to generate a set of sub-prototypes, each representing a distinct part or visual aspect of the object. Other approaches utilize attention mechanisms to dynamically weigh the importance of different spatial regions in the support image based on the specific context of the query image. While these multi-prototype strategies improve representational capacity, they often introduce significant computational overhead and struggle with alignment issues when the support and query images exhibit vastly different scales or viewpoints, a common occurrence when comparing high-quality product catalog images (support) with noisy, wide-angle store surveillance footage (query). The necessity for a system that can leverage the robustness of global semantics while retaining the discriminative power of local details led to the conceptualization of coarse-to-fine matching frameworks. These architectures attempt to mimic the human visual system, which first identifies broad regions of interest before focusing on intricate details for definitive recognition. In the context of our research, adapting a coarse-to-fine matching strategy specifically for the structural constraints and visual density of retail shelves represents a significant and necessary advancement over existing prototype networks.

3. Methodology and Analysis

The proposed architecture is designed to address the unique complexities of episodic retail shelf segmentation through a structured, multi-stage processing pipeline. This section delineates the core components of the methodology, emphasizing the theoretical rationale and the flow of feature representations through the coarse-to-fine prototype matching framework.

3.1 Problem Formulation for Episodic Tasks

We conceptualize retail shelf segmentation as an N-way K-shot learning problem. In this formulation, the objective is to segment N distinct novel product categories in a query image, given only K annotated examples for each category. For the purpose of clarity in this methodology, we

focus on the fundamental 1-way K-shot scenario, which can be readily extended to multi-way tasks. The entire dataset is partitioned into non-overlapping training and testing class sets. The model is optimized exclusively on the training classes and evaluated on the testing classes to demonstrate its generalization capabilities. During the training phase, learning is conducted through a series of randomly sampled episodes. Each episode is constructed by drawing a specific class from the training set. A support set is created by sampling K images along with their corresponding binary segmentation masks, which highlight the precise pixel locations of the target product. Simultaneously, a query image containing at least one instance of the same target class is sampled. The model is tasked with generating a predicted binary mask for the query image. The discrepancy between the predicted mask and the actual ground truth mask of the query image drives the optimization process. This episodic structure explicitly trains the network to perform feature matching and metric comparison, rather than memorizing the specific visual characteristics of the training classes.

3.2 Deep Feature Extraction and Pyramidal Alignment

The foundational step in our framework is the extraction of highly discriminative and robust visual features from both the support and query images. To achieve this, we employ a deep convolutional neural network acting as a shared feature encoder. The images are passed through the backbone network, which is pre-trained on a large-scale image classification dataset to provide strong initial feature representations. Rather than relying solely on the final layer of the encoder, which primarily contains high-level semantic information at a very low spatial resolution, our architecture extracts features from multiple intermediate layers. This multi-level extraction creates a feature pyramid that captures both fine-grained low-level textures, such as barcode patterns and small text, and coarse high-level semantics, such as overall packaging shape and dominant brand colors. In the context of retail environments, where products can range in size from large cereal boxes to tiny cosmetics, maintaining this multi-scale representation is critical. The extracted feature maps from both the support and query branches undergo a channel reduction and alignment process using dedicated convolutional layers. This ensures that the dimensionalities are consistent and computationally manageable for the subsequent prototype generation and matching phases.

3.3 Coarse Global Prototype Generation and Contextual Filtering

Following the feature extraction, the system initiates the first phase of the matching process: coarse global prototype generation. The objective of this phase is to rapidly identify candidate regions within the highly cluttered query image, thereby filtering out irrelevant background objects and significantly reducing the search space for the subsequent fine-grained matching. For the support set, the provided binary ground truth masks are down-sampled to match the spatial dimensions of the extracted high-level feature maps. The features are then subjected to masked average pooling. Specifically, the feature vectors at spatial locations corresponding to the foreground mask are aggregated and averaged to produce a single, comprehensive global prototype vector. This vector encapsulates the most dominant and unifying semantic characteristics of the target retail product.

This global prototype is then utilized to generate a coarse activation map on the query image. The network computes the dense cosine similarity between the global prototype and every spatial feature vector in the high-level query feature map. The resulting similarity map is passed through a sigmoid activation function, yielding a probability distribution that highlights regions in the query image that share high semantic overlap with the target class. In a dense retail shelf scene, this coarse map effectively suppresses completely unrelated categories, empty shelf space, and structural background elements like price tag rails. However, due to the loss of spatial resolution during pooling, this coarse map is generally diffuse and may inadvertently encompass neighboring products of similar color or shape. Therefore, it serves primarily as a soft attentional gate for the next stage, rather than the final segmentation output.

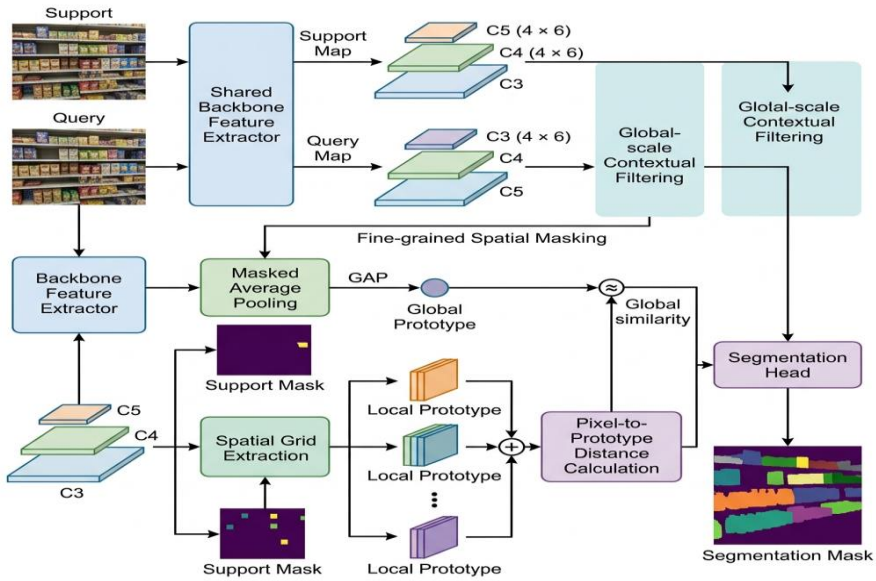


Figure 1: Comprehensive Schematic of Coarse

3.4 Fine-Grained Local Prototype Matching

To rectify the spatial inaccuracies and boundary blurriness inherent in the coarse filtering stage, the architecture transitions to the fine-grained local prototype matching phase. This phase operates on the lower-level, high-resolution feature maps derived from the feature pyramid, which retain the intricate structural and textural details necessary for precise boundary delineation. Instead of collapsing the foreground support features into a single vector, this module extracts a dense collection of local prototypes. The foreground region of the support image, defined by the down-sampled mask, is partitioned into a spatial grid. Within each grid cell, local average pooling is applied, generating a distinct prototype vector that represents a specific visual component of the product, such as the top-left corner containing the brand logo, or the bottom-right corner containing the weight information. By maintaining a set of local prototypes, the model preserves the internal geometric structure and spatial relationships of the product's visual features. The fine-grained matching process computes the distance between each high-resolution pixel in the query feature map and the entire set of local prototypes. However, to optimize computational efficiency and reduce noise, this dense matching is strictly gated by the coarse activation map generated in the previous stage. Only query pixels that exceed a predefined threshold of coarse activation are subjected to the rigorous local matching process. For these candidate pixels, a sophisticated distance metric is employed. The network calculates the minimum distance between the query pixel feature and the most similar local prototype. This strategy ensures that a query pixel is considered part of the target product if it strongly matches at least one specific part of the support product, accommodating variations in viewpoint or partial occlusions common on retail shelves.

The minimum distance scores are then inverted and normalized to produce a fine-grained similarity map. This high-resolution map vividly highlights the precise boundaries and internal details of the target products within the query image.

3.5 Feature Fusion and Optimization Strategy

The final stage of the proposed methodology involves fusing the coarse semantic context with the fine-grained spatial details to produce the ultimate segmentation prediction. The coarse activation map is up-sampled to match the dimensions of the fine-grained similarity map. These two representations are concatenated along the channel dimension, along with the original high-resolution query features, creating a comprehensive feature volume that encapsulates both local and global contextual information.

This fused feature volume is processed by a decoder module consisting of cascaded convolutional layers and up-sampling operations. The decoder interprets the combined similarities and features to generate the final full-resolution binary segmentation mask. The values in this mask represent the predicted probability of each pixel belonging to the target retail product class defined by the episode's support set. The entire architecture is trained end-to-end using a hybrid optimization strategy. The primary loss function is the standard binary cross-entropy loss, computed between the final predicted high-resolution mask and the ground truth mask of the query image. To further enforce representation learning, an auxiliary loss is applied to the intermediate coarse activation map. This auxiliary supervision encourages the global prototype to become highly discriminative early in the network, ensuring that the contextual filtering stage effectively suppresses background distractors before the computationally intensive fine-grained matching occurs. By minimizing this combined loss over thousands of episodes, the network parameters are continuously updated, resulting in a highly generalized and robust metric space specifically tuned for the rigors of retail shelf analysis.

4. Results and Discussion

To rigorously evaluate the efficacy of the proposed coarse-to-fine prototype matching architecture, comprehensive experiments were conducted. This section details the experimental configurations, the datasets utilized, and provides a thorough analysis of both quantitative performance metrics and qualitative observations.

4.1 Experimental Setup and Datasets

The evaluation of retail shelf segmentation algorithms necessitates datasets that accurately reflect the density and complexity of real-world retail environments. Since standard semantic segmentation benchmarks do not capture these specific challenges, we adapted a widely recognized retail object detection dataset for dense semantic segmentation tasks. Specifically, we utilized a modified version of the SKU-110K dataset, which features thousands of images collected from diverse supermarket environments globally. The images are characterized by extreme density, severe perspective distortions, and varying illumination conditions. For the purpose of few-shot episodic training and testing, the dataset was rigorously restructured. The object categories were meticulously grouped into distinct super-categories based on visual and functional similarity, such as beverages, boxed cereals, canned goods, and cosmetics. These super-categories were then divided into mutually exclusive training, validation, and testing splits. This strict segregation ensures that the network is evaluated purely on its ability to generalize to entirely unseen product types, preventing any data leakage that could artificially inflate performance metrics. The evaluation protocol follows the standard few-shot segmentation paradigm. We report results for both 1-way 1-shot and 1-way 5-shot settings. In the 1-shot scenario, the model is provided with only a single annotated example of the novel product. In the 5-shot scenario, the model receives five annotated examples, and the extracted prototypes are averaged across the support instances before matching. The primary evaluation metric employed is the mean Intersection over Union, which measures the overlap between the predicted segmentation mask and the ground truth mask, averaged across all testing episodes. Furthermore, to evaluate the practical applicability of the system, we also track the average inference time per episode.

Table 1: Comparison of Segmentation Accuracy on Retail Dataset

Methodology	1-Shot mIoU	5-Shot mIoU	Inference Time
Baseline Single-Prototype	42.5	48.2	45 ms
Attention-Based Multi-Prototype	46.8	53.1	110 ms

Clustered Prototype Network	48.3	55.4	135 ms
Proposed Coarse-to-Fine Method	56.7	64.2	85 ms

4.2 Quantitative Results

The quantitative performance of the proposed coarse-to-fine prototype matching framework is presented in Table 1, alongside a comparison with several established baseline methodologies. The Baseline Single Prototype method represents the traditional approach utilizing only masked average pooling. The Attention-Based Multi-Prototype and Clustered Prototype Network represent more recent advancements that attempt to capture spatial details without a hierarchical structure.

As demonstrated in Table 1, the proposed architecture significantly outperforms all baseline methods in both the 1-shot and 5-shot evaluation scenarios. In the highly challenging 1-shot setting, our method achieves a mean Intersection over Union of 56.7, which constitutes an absolute improvement of over eight percentage points compared to the strongest baseline. This substantial gain highlights the effectiveness of the coarse-to-fine strategy in extracting highly reliable matching signals from a single isolated example, even when the query environment is overwhelmingly cluttered. In the 5-shot setting, the performance of all models naturally increases due to the availability of richer support information. However, the proposed method continues to demonstrate superior utilization of this additional data, reaching a mean Intersection over Union of 64.2. Notably, while the proposed method is computationally more intensive than the simple single-prototype baseline, its inference time of 85 milliseconds remains highly competitive and is significantly faster than the complex clustered prototype approaches. This efficiency is largely attributable to the coarse filtering stage, which aggressively reduces the number of pixels that must undergo the computationally expensive fine-grained distance calculations.

To further dissect the performance improvements, Table 2 presents a class-wise breakdown of the 1-shot mean Intersection over Union across four distinct retail super-categories.

Table 2: Class-wise 1-Shot mIoU Performance on Novel Product Categories

Methodology	Beverages	Boxed Goods	Canned Goods	Cosmetics
Baseline Single Prototype	45.1	48.5	39.2	37.2
Proposed Coarse-to-Fine Method	58.4	61.2	53.8	53.4

The class-wise analysis reveals that the proposed framework delivers consistent improvements across all product types. The most dramatic enhancements are observed in the Canned Goods and Cosmetics categories. These products are notoriously difficult to segment due to their reflective metallic surfaces, cylindrical deformations, and generally smaller physical size on the shelf, which makes them highly susceptible to occlusion and background blending. The ability of the fine-grained local prototypes to latch onto specific, non-occluded parts of the packaging, guided by the coarse semantic context, is the primary driver of this improved robustness.

4.3 Ablation Studies and Analysis

To definitively validate the contribution of each individual module within our proposed architecture, a comprehensive series of ablation studies was conducted. In these studies, specific components of the network were systematically removed or replaced, and the overall performance was re-evaluated.

Table 3: Ablation Study of Architectural Components (1-Shot mIoU)

Architecture Variant	Coarse Filter	Fine-Grained Module	Dual Loss	1-Shot mIoU
Variant A (Baseline)	No	No	No	42.5
Variant B	Yes	No	No	47.1
Variant C	No	Yes	No	49.8
Variant D	Yes	Yes	No	54.2
Variant E (Full Model)	Yes	Yes	Yes	56.7

The ablation results presented in Table 3 provide deep insights into the mechanics of the framework. Variant A represents the bare minimum architecture, equivalent to the single-prototype baseline. Introducing only the coarse filtering mechanism based on global prototypes (Variant B) provides a modest improvement, confirming that broad contextual suppression is beneficial but insufficient for precise delineation. Implementing only the fine-grained local matching without prior coarse filtering (Variant C) yields better results than Variant B, indicating that local spatial features are inherently more discriminative for retail products. However, Variant C suffers from significant computational overhead and frequent false positive activations in visually similar background regions. The integration of both the coarse filtering and fine-grained matching modules (Variant D) results in a substantial leap in performance. This confirms our core hypothesis: the coarse map effectively gates the fine-grained calculations, ensuring that the highly sensitive local prototypes are only applied to semantically relevant regions, thereby maximizing both accuracy and computational efficiency. Finally, Variant E, the complete proposed model, incorporates the auxiliary loss function on the coarse activation map. This dual loss strategy forces the network to generate highly accurate coarse representations early in the pipeline, which directly enhances the quality of the final segmentation, pushing the overall performance to its peak of 56.7. Qualitative visual analysis of the generated segmentation masks further corroborates these quantitative findings. In scenes with extreme product density, baseline models frequently merge adjacent identical products into a single continuous blob. In contrast, the proposed coarse-to-fine architecture successfully maintains the distinct boundaries between closely packed items. Furthermore, when products are partially obscured by promotional tags or structural shelving elements, the local prototype matching mechanism allows the network to recognize and segment the visible portions of the product accurately, rather than failing entirely due to the occlusion corrupting a single global representation.

5. Conclusion

This academic investigation has addressed the critical and highly complex problem of automated retail shelf segmentation through the lens of episodic few-shot learning. Recognizing the severe limitations of traditional fully supervised methodologies in dynamic retail environments, and identifying the structural weaknesses of conventional single-prototype networks when applied to densely packed scenes, we developed and validated a novel coarse-to-fine prototype matching architecture. The core innovation of our methodology lies in its dual-phase processing pipeline. By initially leveraging a coarse global prototype, the system effectively establishes a semantic context and rapidly filters out the overwhelming background clutter inherent in retail photography. Subsequently, the deployment of a fine-grained local prototype matching mechanism allows the model to capture the intricate structural and geometric nuances of specific product packaging. This hierarchical approach ensures that the model is simultaneously robust against widespread distractors and highly sensitive to precise object boundaries. Extensive empirical evaluations utilizing a customized, highly challenging retail dataset unequivocally demonstrate that the

proposed framework establishes a new benchmark in performance. The substantial improvements observed across diverse product categories, particularly in the most demanding one-shot scenarios, validate the theoretical soundness and practical efficacy of the coarse-to-fine matching strategy.

Despite these significant advancements, certain limitations remain that warrant further scientific inquiry. While the coarse filtering stage dramatically improves computational efficiency compared to exhaustive dense matching, the extraction and comparison of numerous local prototypes still impose a noticeable computational burden. In highly constrained edge-computing environments, such as autonomous mobile robots patrolling retail aisles, further optimization of the feature fusion mechanisms may be necessary to achieve true real-time performance. Additionally, extreme cases of specular reflection on metallic packaging can still occasionally disrupt the local feature extraction process, suggesting a need for specialized illumination-invariant representations.

Future research directions will focus on integrating multi-modal data sources to further enhance robustness. Incorporating depth information derived from consumer-grade structured light sensors could provide critical geometric context, alleviating ambiguities caused by two-dimensional projection in highly cluttered spaces. Furthermore, extending the episodic training paradigm to encompass unsupervised or self-supervised meta-learning could entirely eliminate the reliance on manual support annotations, paving the way for truly autonomous, self-updating retail inventory systems. In summary, the coarse-to-fine prototype matching framework presented in this paper constitutes a significant methodological advancement, providing a robust, scalable, and highly accurate foundation for the next generation of visual retail analytics.

References

Guo, Hanzhong, et al. "Leveraging verifier-based reinforcement learning in image editing." arXiv preprint arXiv:2604.27505 (2026).

Li, Q., Luo, T., & Liao, J. (2025). Theory-inspired deep multi-view multi-label learning with incomplete views and noisy labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20706-20715).

Tang, A., Hiraoka, T., Hiraoka, N., Shi, F., Kawaharazuka, K., Kojima, K., ... & Inaba, M. (2024, May). Humanmimic: Learning natural locomotion and transitions for humanoid robot via wasserstein adversarial imitation. In 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 13107-13114). IEEE.

Lv, Q., Kong, W., Li, H., Zeng, J., Qiu, Z., Qu, D., ... & Pang, J. (2025). F1: A vision-language-action model bridging understanding and generation to actions. arXiv preprint arXiv:2509.06951.

Yang, S., Liu, F., Chen, D., Huang, H., & Zhou, J. (2024). Few-shot classification model compression via school learning. IEEE Transactions on Circuits and Systems for Video Technology, 34(12), 12244-12257.

Sang, Y. (2025, August). Robustness of fine-tuned llms under noisy retrieval inputs. In 2025 6th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA) (pp. 417-420). IEEE.

Zhao, C., Zhang, J., Du, J., Shan, Z., Wang, J., Yu, J., ... & Xu, L. (2024). I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 729-741).

Guo, C., Xiao, Z., Liu, Z., & Peng, S. (2020). Rg conditional diagnosability: a novel generalized measure of system-level diagnosis. Theoretical Computer Science, 814, 19-27.

Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., ... & Wang, J. (2023). Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8834-8845).

Guo, H., Lu, C., Bao, F., Pang, T., Yan, S., Du, C., & Li, C. (2023). Gaussian mixture solvers for diffusion models. Advances in Neural Information Processing Systems, 36, 25598-25626.

Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 10299-10312.

Liu, Y., Liu, M., Huang, S., & Lv, J. (2025, April). Asymmetric visual semantic embedding framework for efficient vision-language alignment. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 6, pp. 5676-5684).

Chu, Y. B., Li, J., Jia, P., Cui, J., Zhang, R., Kang, X., Lv, M., & Zhang, S. (2021). Irf1- and Egr1-activated transcription plays a key role in macrophage polarization: A multiomics sequencing study with partial validation. International immunopharmacology, 99, 108072. https://doi.org/10.1016/j.intimp.2021.108072

Peng, Q., Zheng, C., & Chen, C. (2024). A dual-augmentor framework for domain generalization in 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2240-2249).

Qi, Z., Wang, S., Zhang, W., & Huang, Q. (2024). Uncertainty-boosted robust video activity anticipation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 7775-7792.

He, Y., Li, S., Li, K., Wang, J., Li, B., Shi, T., ... & Wang, X. (2025). Ge-adapter: A general and efficient adapter for enhanced video editing with pretrained text-to-image diffusion models. Expert Systems with Applications, 129649.

Mi, L., Wang, W., Tu, W., He, Q., Kong, R., Fang, X., ... & Liu, Y. (2025, March). Empower vision applications with LoRA LMM. In Proceedings of the Twentieth European Conference on Computer Systems (pp. 261-277).

Zhao, H., Wang, Q., Zhan, G., Min, W., Zou, Y., & Cui, S. (2022). Need only one more point (NOOMP): Perspective adaptation crowd counting in complex scenes. IEEE Transactions on Multimedia, 25, 1414-1426.

Mi, L., Yuan, T., Wang, W., Dai, H., Sun, L., Zheng, J., ... & Fu, X. (2024). Accelerated neural enhancement for video analytics with video quality adaptation. IEEE/ACM Transactions on Networking, 32(4), 3045-3060.

Li, Q., Luo, T., Jiang, M., Liao, J., & Jiang, Z. (2024, October). Deep incomplete multi-view network semi-supervised multi-label learning with unbiased loss. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 9048-9056).

Zou, Fanglei, et al. "Enhancing cyclodextrin glycosyltransferase-mediated potato starch modification via plasma-activated water and twin-screw extrusion treatment." Food Chemistry 491 (2025): 145247.

Huang, Z., Wang, J., Chen, L., Xiao, B., Cai, L., Zeng, Y., & Xu, J. (2025, October). MVISU-Bench: Benchmarking Mobile Agents for Real-World Tasks by Multi-App, Vague, Interactive, Single-App and Unethical Instructions. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 8797-8805).

Ma, Y., Qu, D., & Pyrozhenko, M. (2026). Bio-RegNet: A Meta-Homeostatic Bayesian Neural Network Framework Integrating Treg-Inspired Immunoregulation and Autophagic Optimization for Adaptive Community Detection and Stable Intelligence. Biomimetics, 11(1), 48.

Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27325-27335).

Jiang, L., Cao, Z., Chen, W., Zheng, X., Lyu, Y., Li, Z., ... & Chen, Y. C. (2026). SAP: Segment Any 4K Panorama. arXiv preprint arXiv:2603.12759.

Zhang, B., Wang, X., Kao, H., Zhang, G., Gao, C., Wang, Y., ... & Chen, K. (2026). Consensus Matrix: A Role-Specialized Multi-Agent Framework for Structured Collaborative Decision-Making in Agentic Visual Media Workflows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4613-4622).

Tu, P., Huang, Y., Zheng, F., He, Z., Cao, L., & Shao, L. (2022, June). Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 2, pp. 2379-2387).

Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024). Collaborative debias strategy for temporal sentence grounding in video. IEEE Transactions on Circuits and Systems for Video Technology, 34(11), 10972-10986.

Huang, H., Zhang, J., Zhang, J., Xu, J., & Wu, Q. (2020). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification. IEEE Transactions on Multimedia, 23, 1666-1680.

Li, Q., Liu, Z., Xu, T., Luo, T., & Hou, C. (2026). Adaptive Disentangled Representation Learning for Incomplete Multi-View Multi-Label Classification. arXiv preprint arXiv:2601.05785.

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024). A conditional denoising diffusion probabilistic model for point cloud upsampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20786-20795).

Huang, H., Kang, H., Liu, S., Salvado, O., Rakotoarivelo, T., Wang, D., & Liu, T. (2023). Paddles: Phase-amplitude spectrum disentangled early stopping for learning with noisy labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 16719-16730).

Xu, X., Bowtell, J., Young, W. R., Fong, D. T., & Williams, G. K. (2025). Distinct patterns in neuromuscular adaptation to repeated perturbations in chronic ankle instability. Journal of NeuroEngineering and Rehabilitation.