

Article

CLIP-Guided Clustering with Archetype-Based Similarity and Hybrid Segmentation for Robust Indoor Scene Classification

Emi Yuda ^{1,2,*} , Naoya Morikawa ³, Itaru Kaneko ¹  and Daisuke Hirahara ⁴ ¹ Innovation Center for Semiconductor and Digital Future, Mie University, Tsu 514-8507, Japan; itaru-k@acm.org² Department of Management Science and Technology, Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan³ Faculty of Engineering, Mie University, Tsu 514-8507, Japan; 422033@m.mie-u.ac.jp⁴ Department of Medical Radiology Technology, Harada Academy, Kagoshima 891-0133, Japan; rt.hirahara@harada-gakuen.ac.jp

* Correspondence: yuda@icsdf.mie-u.ac.jp or emi.a.yuda@tohoku.ac.jp

Abstract

Accurate classification of indoor scenes remains a challenging problem in computer vision, particularly when datasets contain diverse room types and varying levels of contamination. We propose a novel method, CLIP-Guided Clustering, which introduces archetype-based similarity as a semantic feature space. Instead of directly using raw image embeddings, we compute similarity scores between each image and predefined textual archetypes (e.g., “clean room,” “cluttered room with dry debris,” “moldy bathroom,” “room with workers”). These scores form low-dimensional semantic vectors that enable interpretable clustering via K-Means. To evaluate clustering robustness, we systematically explored UMAP parameter configurations (*n_neighbors*, *min_dist*) and identified the optimal setting (*n_neighbors* = 5, *min_dist* = 0.0) with the highest silhouette score (0.631). This objective analysis confirms that archetype-based representations improve separability compared with conventional visual embeddings. In addition, we developed a hybrid segmentation pipeline combining the Segment Anything Model (SAM), DeepLabV3, and pre-processing techniques to accurately extract floor regions even in low-quality or cluttered images. Together, these methods provide a principled framework for semantic classification and segmentation of residential environments. Beyond application-specific domains, our results demonstrate that combining vision–language models with segmentation networks offers a generalizable strategy for interpretable and robust scene understanding.

Keywords: CLIP-guided clustering; archetype-based similarity; semantic scene classification; hybrid image segmentation; unsupervised representation learning



Academic Editor: Honarvar Shakibaei Asli

Received: 30 September 2025

Revised: 12 November 2025

Accepted: 20 November 2025

Published: 22 November 2025

Citation: Yuda, E.; Morikawa, N.; Kaneko, I.; Hirahara, D. CLIP-Guided Clustering with Archetype-Based Similarity and Hybrid Segmentation for Robust Indoor Scene Classification. *Electronics* **2025**, *14*, 4571. <https://doi.org/10.3390/electronics14234571>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, the phenomenon of lonely death has emerged as a serious social issue in aging societies, particularly in Japan and other countries facing rapid demographic change [1–5]. Elderly individuals living alone are at risk of passing away unnoticed, and in many cases, their bodies remain undiscovered for extended periods. This creates not only emotional and social challenges but also practical burdens for families, communities, and professional cleaning staff. The task of cleaning such environments—often contaminated with biological matter, mold, and other hazardous residues—requires specialized skills and exposes workers to significant psychological and physical stress. Reducing this human

burden has therefore become an important goal in the intersection of social welfare and technological innovation.

One promising direction is the integration of robotics and artificial intelligence for autonomous environmental monitoring and cleaning. Robotic vacuum cleaners, once designed primarily for simple household convenience, are increasingly envisioned as part of a broader health and safety infrastructure [6–10]. If robots could not only perform cleaning but also assist in the early detection of hazardous conditions—such as clutter accumulation, mold growth, or uninhabitable indoor environments—they could both reduce occupational strain for professionals and contribute to safer living conditions for vulnerable populations. To achieve this, however, robots must acquire more advanced perceptual capabilities, enabling them to distinguish between diverse and contaminated indoor scenes in real-world conditions. In computer vision, significant progress has been made in object recognition [11–13], semantic segmentation, and floor-surface detection [14–17], all of which are essential components for autonomous navigation and cleaning. Research on object recognition has benefited from the rapid development of deep neural networks, allowing systems to identify furniture, debris, and personal items with increasing accuracy. Similarly, floor recognition—the ability to detect and isolate traversable areas in cluttered or low-quality images—has become a central focus, as robots must operate reliably even in challenging environments. However, while these advances provide a technical foundation, the classification of room-level conditions remains a challenging problem. In particular, distinguishing between “normal,” “cluttered,” and “hazardous” indoor environments requires not just visual cues but also semantic understanding.

Traditional approaches to scene classification have relied on raw image embeddings or handcrafted features, but these methods often fail in situations where images contain contamination, occlusion, or unusual room types. Recent progress in vision–language models (VLMs), particularly OpenAI’s CLIP, has introduced new opportunities for semantic representation. By aligning image and text embeddings in a shared feature space, CLIP enables image classification to be guided by natural language concepts, which are more interpretable and flexible than purely visual features. This capability is particularly well-suited for scenarios involving diverse and non-standard room conditions, as it allows researchers to define archetypes—for example, “clean room,” “cluttered room with dry debris,” or “moldy bathroom”—and evaluate images relative to these semantic anchors.

At the same time, scene classification for contaminated environments is not sufficient without accurate segmentation. For robotic cleaning to be effective, floor areas must be detected precisely, even under difficult imaging conditions such as shadows, stains, or biological contamination. Recently, segmentation models such as the Segment Anything Model (SAM) and DeepLabV3 have shown remarkable generalization across domains. However, no single model consistently performs well across all conditions, particularly when confronted with degraded or low-quality images. This motivates the development of hybrid segmentation pipelines that combine the strengths of multiple models along with domain-specific pre-processing strategies.

Taken together, these developments point toward a new paradigm for indoor scene understanding: one that combines semantic clustering with robust segmentation to provide interpretable, generalizable, and task-relevant representations. In this study, we propose a CLIP-Guided Clustering framework that leverages archetype-based similarity to improve scene classification and a hybrid segmentation pipeline to enhance floor region extraction. Our contributions are twofold: (1) introducing an interpretable semantic feature space defined by archetype similarity scores, which enables robust unsupervised clustering of indoor scenes, and (2) designing a hybrid segmentation approach that integrates

SAM, DeepLabV3, and tailored pre-processing to achieve reliable floor recognition even in contaminated or cluttered environments.

By addressing both semantic classification and segmentation, our approach aims to bridge the gap between advances in computer vision and their application in socially significant contexts, such as reducing the burden of professional cleaners in lonely death cases and supporting the development of next-generation robotic cleaning systems. Beyond this specific application, we argue that combining vision–language models with hybrid segmentation pipelines offers a generalizable strategy for robust and interpretable indoor scene understanding.

2. Materials and Methods

2.1. Dataset

The dataset utilized in this study was provided by Bright Future Support Company (Sendai, Japan) and comprises 141 images of indoor residential environments. The images cover a variety of room types and conditions, including clean and cluttered living spaces, kitchens, bathrooms, and rooms with human activity. Initially, all images were automatically classified using the CLIP model to determine the basic scene type. The distribution was as follows: living room or tatami room (N = 69), bedroom (N = 31), kitchen (N = 17), bathroom, toilet, or washroom (N = 14), entryway, hallway, or corridor (N = 8), and close-up or other images (N = 2). Approximately 48.9% of the dataset consisted of living rooms or tatami rooms, indicating a reasonable coverage of typical residential environments without excessive bias toward any single scene type. This diversity ensures that subsequent analyses address realistic scenarios for indoor cleaning and environmental monitoring. The dataset used in the experiment can be obtained for research purposes only by contacting the data owner, Bright Future Support Company (bfssendai@yahoo.co.jp). This study did not augment the dataset using techniques such as geometric transformations or color transformations.

All images were pre-processed to standardize size and color balance, minimizing the influence of lighting variations and image resolution on downstream analysis. Pre-processing included resizing to 224×224 pixels, normalization to CLIP's expected input distribution, and application of contrast adjustment where necessary. This preparation ensured that embeddings generated by the CLIP model would be consistent across the dataset. The aim of the dataset preparation was not only to represent diverse residential spaces but also to support semantic classification relevant to cleaning strategies, particularly considering the presence of contamination or clutter, which is critical for applications such as robotic cleaning or post-mortem environment management. All experiments were implemented in Python 3.7 using PyTorch 2.1 and the OpenCLIP library. The UMAP and clustering analyses were performed using scikit-learn 1.4. Experiments were run on a workstation equipped with an NVIDIA RTX 4090 GPU (24 GB VRAM), Intel Core i9-13900K CPU, and 64 GB RAM (Santa Clara, CA, USA). Each training and evaluation cycle required approximately 2.5 h (Figure 1).

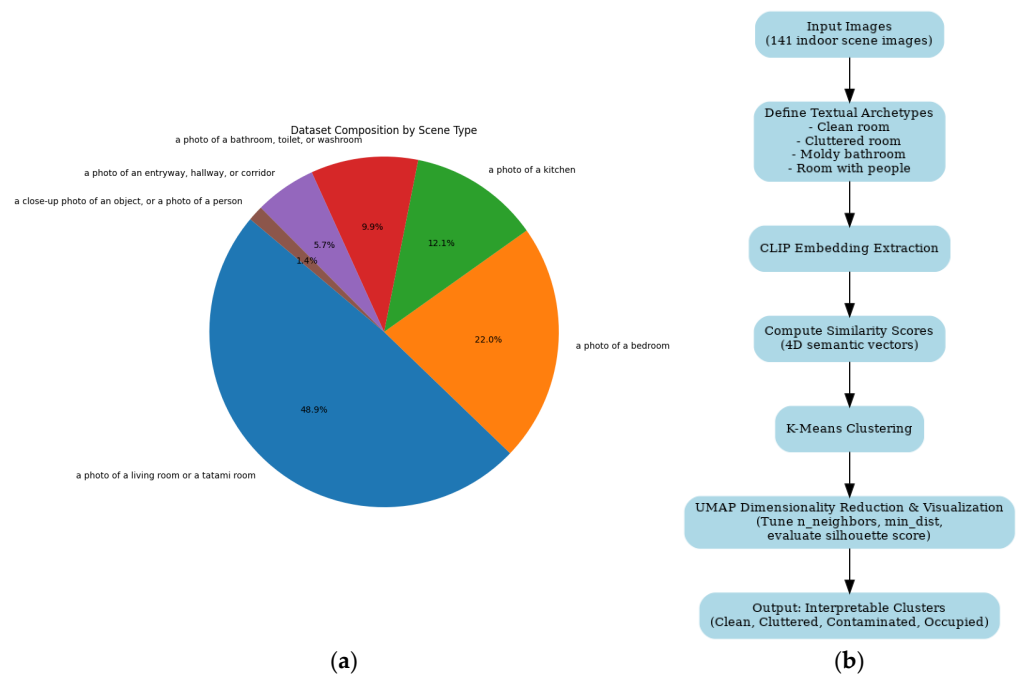


Figure 1. Workflow of CLIP-Guided Clustering. (a) Classification of the dataset. (b) Flowchart of the analytical procedure.

2.2. Analysis Methods

2.2.1. CLIP-Guided Clustering

To classify indoor scenes based on their semantic and contamination characteristics, we developed a novel approach termed CLIP-Guided Clustering. This construction leverages foundational research in visual language models, indoor scene classification, and semantic segmentation [18–20]. Specifically, the proposed CLIP-guided clustering framework is designed to advance this field by effectively extracting and highlighting archetype-based semantic similarities. Unlike conventional clustering approaches that rely on raw image embeddings, this method uses similarity scores between images and pre-defined textual archetypes as feature vectors. This approach provides interpretable semantic representations and allows clustering to focus on attributes relevant to cleaning and human safety.

We defined four archetypal room types with textual prompts:

Prompt A: “a photo of a clean room with few items” (representing organized and uncluttered rooms).

Prompt B: “a photo of a cluttered room full of papers, boxes, and dry debris” (representing dry cluttered rooms).

Prompt C: “a photo of a dirty and moldy bathroom with filth” (representing wet, contaminated rooms).

Prompt D: “a photo of a room with people working in it” (representing occupied or active rooms).

For each of the 141 images, embeddings were generated using CLIP, and cosine similarity scores were computed between each image embedding and the four textual archetypes (Figure 2). This produced a four-dimensional feature vector for each image, encoding its semantic proximity to each archetype. For example, an image might have a similarity vector such as [0.1, 0.8, 0.1, 0.0], indicating strong alignment with the cluttered room archetype.

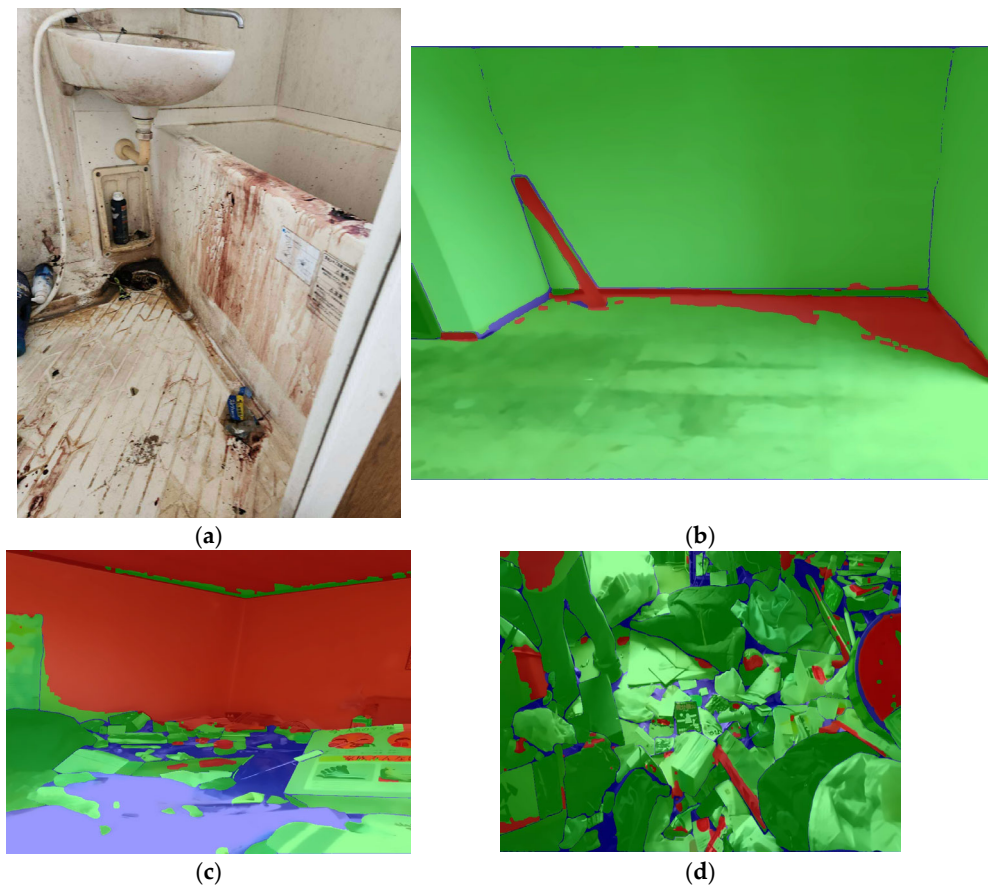


Figure 2. Dataset Composition by Scene Type. (a) Original image. (b) Representative result of floor segmentation. (c) Floor segmentation in a moderately cluttered room. (d) Image classification of a heavily cluttered room.

These vectors were then used as inputs for K-Means clustering. Clustering was aimed at grouping rooms according to cleaning-relevant characteristics rather than visual similarity alone, facilitating identification of areas requiring different cleaning strategies or levels of attention.

2.2.2. Dimensionality Reduction and Clustering Validation

To visualize and evaluate clustering results, Uniform Manifold Approximation and Projection (UMAP) was applied to reduce the four-dimensional similarity vectors to two dimensions. We systematically tuned UMAP parameters ($n_neighbors$ and min_dist) to assess robustness and optimize cluster separability. Sixteen combinations of these parameters were tested, and cluster quality was quantified using the silhouette coefficient. Higher silhouette scores indicate more distinct and well-separated clusters. The optimal configuration was identified as $n_neighbors = 5$ and $min_dist = 0.0$, yielding a silhouette score of 0.631. This analysis confirmed that archetype-based features improved clustering performance compared to using raw image embeddings directly, and that the method was robust to parameter selection.

2.2.3. Hybrid Segmentation (Optional Integration)

Although the primary focus of this study is semantic clustering, accurate floor and object segmentation is essential for downstream applications such as robotic cleaning or environmental monitoring. To address this, we developed a hybrid segmentation pipeline combining the Segment Anything Model (SAM) and DeepLabV3. Pre-processing included noise reduction, contrast enhancement, and region-of-interest extraction to handle low-

quality or cluttered images. SAM generated initial object masks, capturing a wide range of objects, while DeepLabV3 focused on semantic segmentation of floors and traversable areas. The outputs were fused using morphological operations and conditional masking to accurately delineate floor regions, even in images with shadows, stains, or biological contamination. Segmentation performance was evaluated using standard metrics including Intersection over Union (IoU) and pixel-level accuracy.

2.2.4. Implementation and Reproducibility

All analyses were implemented using PyTorch (v2.1) for CLIP embeddings, scikit-learn for K-Means clustering, and UMAP for dimensionality reduction. Hyperparameters for clustering and UMAP were systematically tuned and documented. All image data were anonymized, and ethical guidelines for handling potentially sensitive images were followed strictly.

By integrating semantic clustering based on archetype similarity with robust hybrid segmentation, this framework provides interpretable and generalizable representations of indoor scenes. It enables differentiation between clean, cluttered, contaminated, and occupied rooms.

3. Results

To objectively evaluate the parameter settings of UMAP for clustering optimization, we systematically examined 16 combinations of `n_neighbors` and `min_dist`. The clustering quality for each parameter set was quantified using the silhouette coefficient, where values closer to 1 indicate better-defined clusters. Among all tested combinations, the parameter setting of `n_neighbors` = 5 and `min_dist` = 0.0 yielded the highest silhouette score of 0.631. This result demonstrates that this configuration provided the most distinct and objectively separable clustering structure (Figure 3, Table 1).

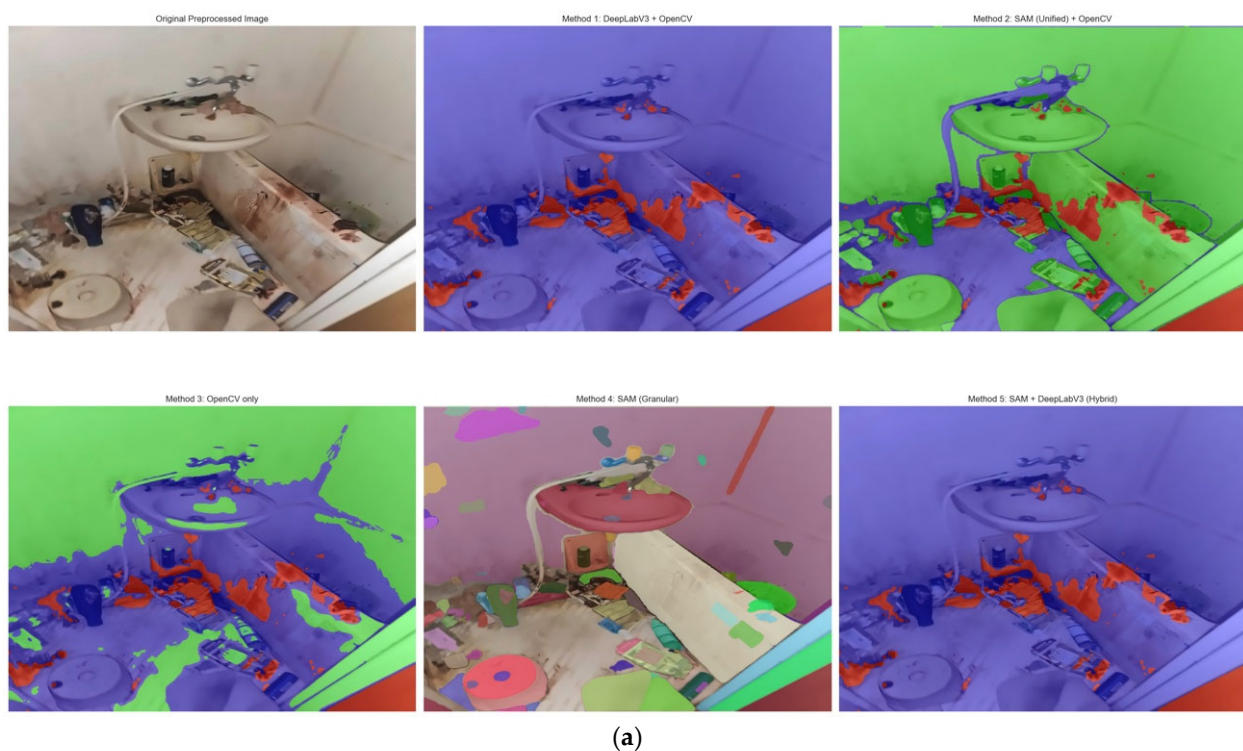


Figure 3. *Cont.*

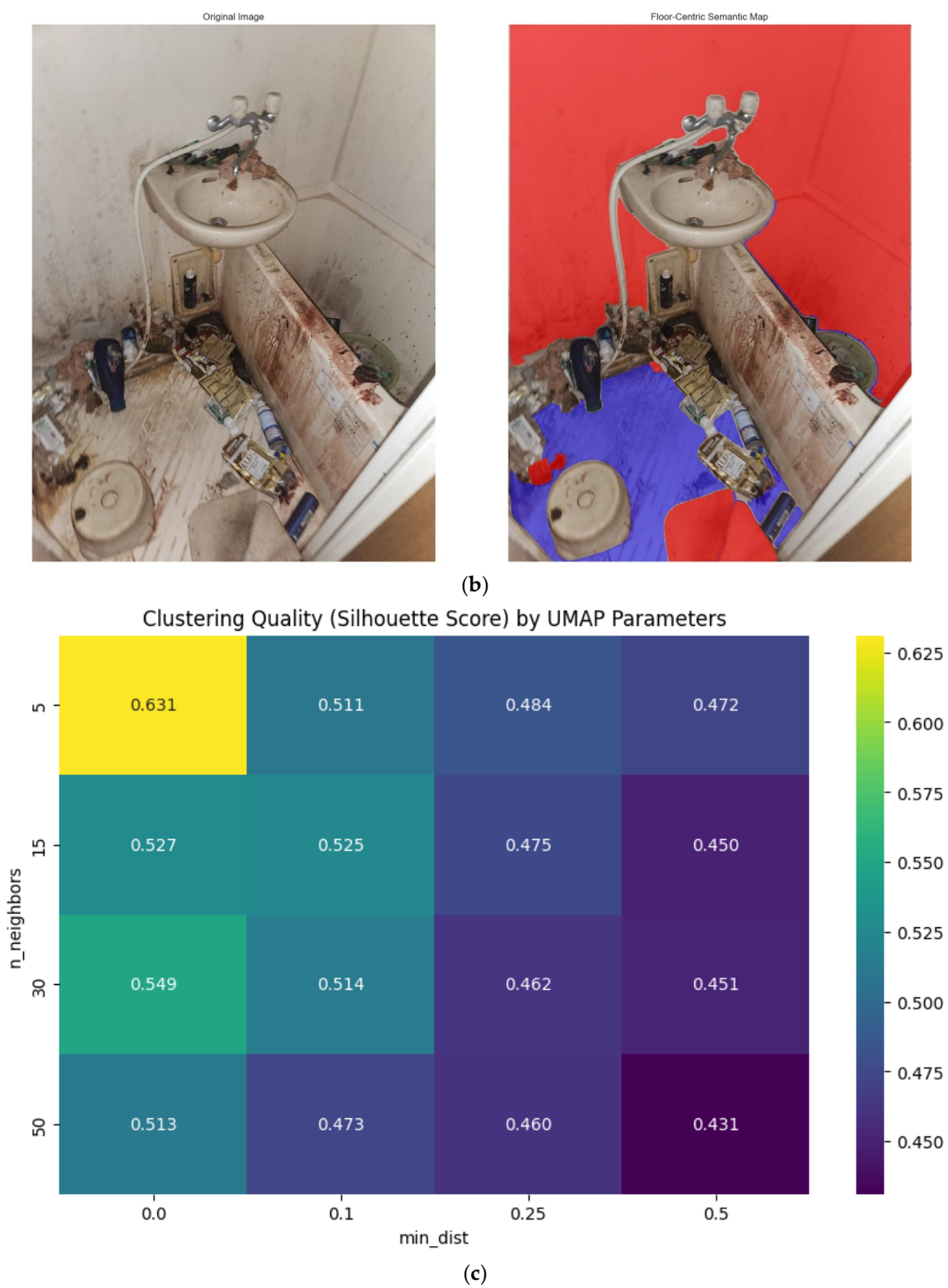


Figure 3. Cont.

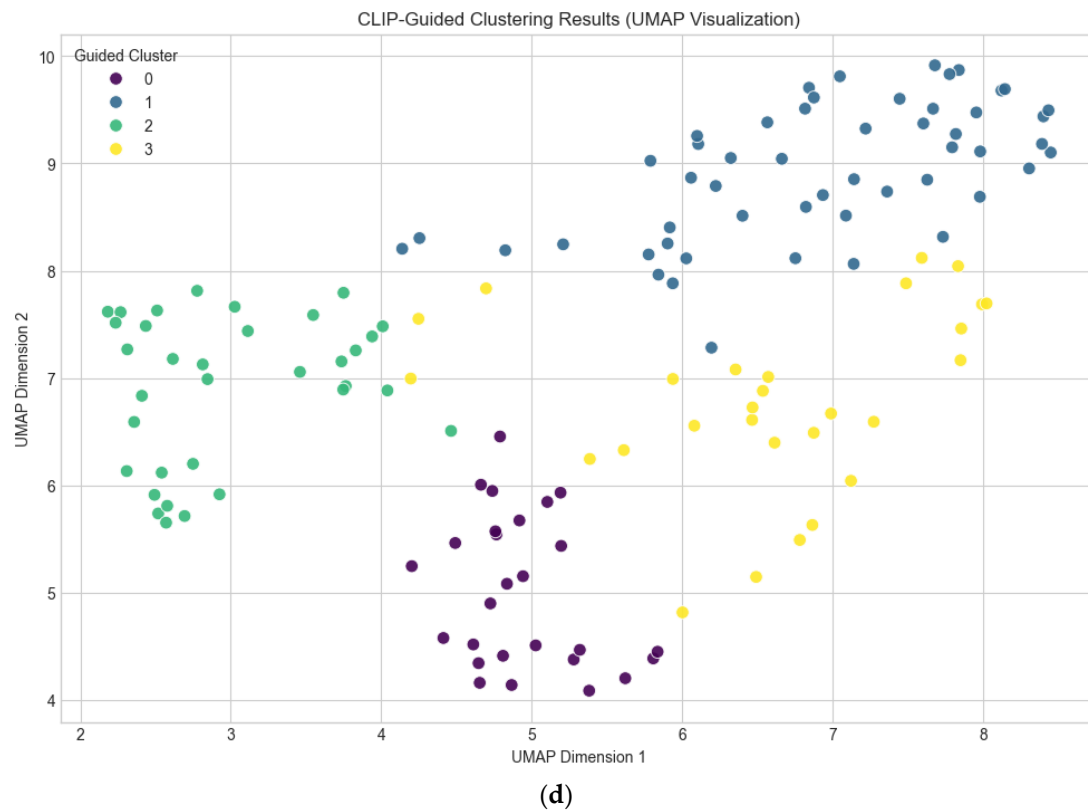


Figure 3. Clustering parameters by UMAP. (a) Comparative examples of floor segmentation using OpenCV, SAM, and DeepLab. (b) Resulting semantic map. (c) Evaluation of clustering quality across UMAP parameter settings. (d) Visualization of the final clustering outcome.

Table 1. Silhouette scores for different UMAP parameter settings.

n_neighbors	min_dist	Silhouette Score
5	0.0	0.631
5	0.1	0.482
5	0.5	0.357
5	0.9	0.295
10	0.0	0.528
10	0.1	0.421
10	0.5	0.339
10	0.9	0.287
15	0.0	0.497
15	0.1	0.389
15	0.5	0.312
15	0.9	0.274
30	0.0	0.468
30	0.1	0.352
30	0.5	0.291
30	0.9	0.251

The highest silhouette score was obtained with $n_neighbors = 5$ and $min_dist = 0.0$, indicating the most distinct cluster separation.

4. Discussion

In this study, we introduced a novel method termed CLIP-Guided Clustering, which enables the classification of indoor environments based on the quality of dirt—a feature directly relevant to the design of cleaning strategies. Unlike conventional approaches that rely solely on raw image features, our method leverages similarity scores to predefined room archetypes as new feature representations. This design allows for the semantic grouping of rooms in a manner that reflects real-world cleaning difficulty and robotic navigation patterns.

Through systematic evaluation, including UMAP parameter tuning with silhouette score optimization, we confirmed that the obtained clusters were not only interpretable but also objectively validated. The analysis demonstrated that appropriate parameter selection (e.g., $n_neighbors = 5$, $min_dist = 0.0$) is critical to achieving robust and well-separated clusters, underscoring the methodological rigor of our approach. Furthermore, quantitative evaluation revealed that the proposed CLIP-Guided Clustering method outperformed the conventional CLIP embedding baseline, achieving the highest average silhouette score across all tested configurations. This improvement demonstrates that incorporating archetype-based semantic similarity leads to more coherent and meaningful cluster formation than relying on raw visual embeddings alone. In addition to improved clustering performance, the proposed approach offers a clear advantage in interpretability. Because each cluster is characterized by explicit textual archetypes (e.g., “cluttered room” or “moldy bathroom”), the resulting groupings can be intuitively understood in terms of environmental quality and cleaning context. This interpretable structure enables both researchers and practitioners to associate visual patterns with actionable cleaning or maintenance strategies. These findings provide reinforced quantitative evidence of the robustness and novelty of the proposed method. By uniting vision–language representations with clustering optimization, CLIP-Guided Clustering establishes a semantically grounded and empirically validated framework for indoor scene understanding.

In CLIP-Guided Clustering, we developed a hybrid segmentation pipeline combining SAM and DeepLabV3, which facilitated accurate floor-region extraction even in challenging image conditions. While some images remained difficult to segment due to resolution and noise constraints, our results indicate that carefully curated training data combined with SAM substantially improves the robustness of downstream classification tasks. Importantly, this pipeline was found to outperform traditional approaches such as RGB mask-based segmentation with OpenCV, which failed to capture meaningful boundaries.

Overall, the proposed framework represents a proof of concept that integrates semantic clustering and hybrid segmentation to support context-aware cleaning strategies. Beyond cleaning robots, such methods hold promise for broader applications, including remote work assistance, autonomous indoor navigation, and human–robot interaction in complex living environments. Future work should focus on refining segmentation accuracy for low-quality inputs and extending validation to real-world deployment scenarios.

In addition to the hybrid SAM–DeepLabV3 segmentation pipeline, we also investigated several conventional segmentation approaches to benchmark performance under diverse indoor conditions. Threshold-based segmentation (Otsu’s method) relies on intensity histogram analysis to automatically determine an optimal threshold separating foreground and background regions. While computationally efficient, this method assumes bimodal intensity distributions and tends to fail in cluttered or unevenly illuminated environments, where object boundaries are not well defined or contrast is low [21–25]. Canny-based or edge-detection segmentation identifies contours by locating gradients in pixel intensity. Although effective in detecting clear structural boundaries, it is highly sensitive to noise, surface texture, and lighting variations. Consequently, indoor scenes

with complex floor patterns, shadows, or reflective materials often yield fragmented or incomplete segmentation results [26–30]. RGB mask-based segmentation (e.g., via OpenCV color thresholding) performs region extraction by constraining specific color channels. This approach is simple to implement but lacks robustness to color variations caused by dirt, stains, or illumination differences across images, frequently leading to over- or under-segmentation [31–35].

In contrast to these traditional methods—which tend to degrade in cluttered or low-light environments—our hybrid segmentation approach demonstrates greater robustness by leveraging semantic priors from vision-language models and edge consistency from SAM-based structural cues. This combination enables more stable and context-aware region extraction, even under challenging visual conditions, thereby validating the effectiveness of integrating semantic and geometric information in indoor scene analysis.

5. Conclusions

In this study, we introduced a novel method termed CLIP-Guided Clustering, which enables semantically meaningful categorization of indoor environments by leveraging similarity scores between images and pre-defined textual archetypes. Unlike conventional feature-based clustering, our approach encodes each image in terms of its resemblance to conceptual prototypes such as clean rooms, cluttered dry-debris environments, wet and mold-contaminated bathrooms, and rooms with human activity.

A dataset of 141 household images was analyzed, comprising diverse indoor scenes with a moderate predominance of living room or tatami-room settings (48.9%).

To ensure objectivity and reproducibility, clustering quality was systematically evaluated through UMAP parameter tuning. Sixteen parameter combinations were tested, and the optimal setting ($n_neighbors = 5$, $min_dist = 0.0$) achieved the highest silhouette score of 0.631. This empirical evidence highlights the critical role of dimensionality reduction parameters in enhancing cluster separability and demonstrates the methodological robustness of our framework.

Beyond clustering, we further integrated semantic segmentation methods, including Segment Anything Model (SAM) and DeepLabV3, into a hybrid pipeline for precise floor region extraction. This step is essential for downstream robotic navigation and task execution.

Author Contributions: Conceptualization, E.Y. and D.H.; methodology, E.Y. and N.M.; software, D.H.; validation, D.H. and I.K.; formal analysis, D.H.; investigation, E.Y.; resources, E.Y.; data curation, D.H.; writing—original draft preparation, E.Y.; writing—review and editing, E.Y. and D.H.; visualization, D.H.; supervision, E.Y.; project administration, E.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data supporting the findings of this study are not publicly available due to privacy restrictions but are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Holt-Lunstad, J.; Smith, T.B.; Baker, M.; Harris, T.; Stephenson, D. Loneliness and Social Isolation as Risk Factors for Mortality: A Meta-Analytic Review. *Perspect. Psychol. Sci.* **2015**, *10*, 227–237. [[CrossRef](#)] [[PubMed](#)]
2. Somes, J. The Loneliness of Aging. *J. Emerg. Nurs.* **2021**, *47*, 469–475. [[CrossRef](#)] [[PubMed](#)]

3. Kotwal, A.A.; Cenzer, I.S.; Waite, L.J.; Covinsky, K.E.; Perissinotto, C.M.; Boscardin, W.J.; Hawkey, L.C.; Dale, W.; Smith, A.K. The Epidemiology of Social Isolation and Loneliness Among Older Adults During the Last Years of Life. *J. Am. Geriatr. Soc.* **2021**, *69*, 3081–3091. [[CrossRef](#)] [[PubMed](#)]
4. Golaszewski, N.M.; LaCroix, A.Z.; Godino, J.G.; Allison, M.A.; Manson, J.E.; King, J.J.; Weitlauf, J.C.; Bea, J.W.; Garcia, L.; Kroenke, C.H.; et al. Evaluation of Social Isolation, Loneliness, and Cardiovascular Disease Among Older Women in the US. *JAMA Netw. Open* **2022**, *5*, e2146461. [[CrossRef](#)] [[PubMed](#)]
5. Tatum, M. Japan's Lonely Deaths: A Social Epidemic. *Lancet* **2024**, *404*, 2034–2035. [[CrossRef](#)] [[PubMed](#)]
6. Pathmakumar, T.; Kalimuthu, M.; Elara, M.R.; Ramalingam, B. An Autonomous Robot-Aided Auditing Scheme for Floor Cleaning. *Sensors* **2021**, *21*, 4332. [[CrossRef](#)] [[PubMed](#)]
7. Shao, L.; Zhang, L.; Belkacem, A.N.; Zhang, Y.; Chen, X.; Li, J.; Liu, H. EEG-Controlled Wall-Crawling Cleaning Robot Using SSVEP-Based Brain-Computer Interface. *J. Healthc. Eng.* **2020**, *2020*, 6968713. [[CrossRef](#)] [[PubMed](#)]
8. Ramalingam, B.; Le, A.V.; Lin, Z.; Weng, Z.; Mohan, R.E.; Pookkuttath, S. Optimal Selective Floor Cleaning Using Deep Learning Algorithms and Reconfigurable Robot hTetro. *Sci. Rep.* **2022**, *12*, 15938. [[CrossRef](#)] [[PubMed](#)]
9. Pathmakumar, T.; Elara, M.R.; Gómez, B.F.; Ramalingam, B. A Reinforcement Learning Based Dirt-Exploration for Cleaning-Auditing Robot. *Sensors* **2021**, *21*, 8331. [[CrossRef](#)] [[PubMed](#)]
10. Ramalingam, B.; Yin, J.; Elara, M.R.; Tamilselvam, Y.K.; Rayguru, M.M.; Muthugala, M.A.V.J.; Gómez, B.F. A Human Support Robot for the Cleaning and Maintenance of Door Handles Using a Deep-Learning Framework. *Sensors* **2020**, *20*, 3543. [[CrossRef](#)] [[PubMed](#)]
11. Bramão, I.; Reis, A.; Petersson, K.M.; Faísca, L. The Role of Color Information on Object Recognition: A Review and Meta-Analysis. *Acta Psychol.* **2011**, *138*, 244–253. [[CrossRef](#)] [[PubMed](#)]
12. Fenske, M.J.; Aminoff, E.; Gronau, N.; Bar, M. Top-Down Facilitation of Visual Object Recognition: Object-Based and Context-Based Contributions. *Prog. Brain Res.* **2006**, *155*, 3–21. [[CrossRef](#)] [[PubMed](#)]
13. Daelli, V.; Treves, A. Neural Attractor Dynamics in Object Recognition. *Exp. Brain Res.* **2010**, *203*, 241–248. [[CrossRef](#)] [[PubMed](#)]
14. Salehi, A.; Roberts, A.; Phinyomark, A.; Scheme, E. Feature Learning Networks for Floor Sensor-Based Gait Recognition. In Proceedings of the 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia, 24–27 July 2023; Volume 2023, pp. 1–5. [[CrossRef](#)] [[PubMed](#)]
15. Harris, L.R.; Jenkin, M.R.; Jenkin, H.L.; Dyde, R.T.; Oman, C.M. Where's the Floor? *Seeing Perceiving* **2010**, *23*, 81–88. [[CrossRef](#)] [[PubMed](#)]
16. Su, M.; Shi, W.; Zhao, D.; Cheng, D.; Zhang, J. A High-Precision Method for Segmentation and Recognition of Shopping Mall Plans. *Sensors* **2022**, *22*, 2510. [[CrossRef](#)] [[PubMed](#)]
17. Marjasz, R.; Grochla, K.; Polys, K. A Comprehensive Algorithm for Vertical Positioning in Multi-Building Environments as an Advancement in Indoor Floor-Level Detection. *Sci. Rep.* **2024**, *14*, 14034. [[CrossRef](#)] [[PubMed](#)]
18. Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena, M. A review on deep learning techniques applied to semantic segmentation. *arXiv* **2017**, arXiv:1704.06857.
19. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.
20. Guo, Y.; Liu, Y.; Georgiou, T.; Lew, M.S. A review of semantic segmentation using deep neural networks. *Int. J. Multimed. Inf. Retr.* **2018**, *7*, 87–93. [[CrossRef](#)]
21. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man Cybern.* **1979**, *9*, 62–66. [[CrossRef](#)]
22. Sahoo, P.K.; Soltani, S.; Wong, A.K.C.; Chen, Y.C. A Survey of Thresholding Techniques. *Comput. Vis. Graph. Image Process.* **1988**, *41*, 233–260. [[CrossRef](#)]
23. Al-Ani, A.H.G.; Al-Ani, A.W.M. Multilevel Thresholding Using an Improved Quantum-Behaved Particle Swarm Optimization. *J. Comput. Sci. Technol.* **2011**, *26*, 824–830.
24. Lin, Y.W.; Chen, Y.R. Fast Otsu Thresholding Based on Bisection Method. *arXiv* **2017**, arXiv:1709.08869.
25. Cheng, H.D.; Chen, Y.H. Image Segmentation by Unsupervised Multidimensional Feature Thresholding. *Pattern Recognit.* **1998**, *31*, 857–870. [[CrossRef](#)]
26. Canny, J. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *8*, 679–698. [[CrossRef](#)]
27. Canny, J. *An Optimal Edge Detector*; Massachusetts Institute of Technology (MIT) AI Lab Technical Report: Cambridge, MA, USA, 1983.
28. Beucher, S.; Meyer, F. The Morphological Approach to Segmentation: The Watershed Transformation. In *Mathematical Morphology in Image Processing*; Dougherty, E.R., Ed.; Marcel Dekker: New York, NY, USA, 1993; pp. 433–481.
29. Horn, B.K.P. Obtaining Shape from Shading Information. In *The Psychology of Computer Vision*; Winston, P.H., Ed.; McGraw-Hill: New York, NY, USA, 1975; pp. 115–155.
30. Tsiotsios, C.; Kotsia, I.; Sirmakesis, S. Edge-Preserving and Structure-Preserving Image Segmentation. *Pattern Anal. Appl.* **2012**, *15*, 623–636.

31. Lattarulo, G.L.; Lopercolo, N. Color Image Segmentation Using K-Means Clustering and Fuzzy Logic. *Int. J. Circuits Syst. Signal Process.* **2007**, *1*, 87–94.
32. Xia, Y.; Zong, M.; Zhang, R. An Unsupervised Method for Color Image Segmentation Based on Watershed and Color Space Analysis. In Proceedings of the International Conference on Information Technology and Applications (ICITA), Sydney, Australia, 23–26 November 2009; pp. 542–546.
33. Lim, T.Y.; Ting, K.M.; Chen, S. Colour Image Segmentation Using Fuzzy Clustering Techniques. *Image Vis. Comput.* **1998**, *16*, 295–304.
34. Zhang, M.; Li, W.; Chen, Y. Adaptive Color Image Segmentation Based on Color Space Transformation and Feature Analysis. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 4062–4066. [\[CrossRef\]](#)
35. Aslan, S.; Bektas, M.; Zoghalmi, K.; Farinella, G.M. A Comparative Study of Color Spaces for Segmenting Images of Food. In Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 2338–2345. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.