Intelligent Interior Design Search Engine Using Multi-Modal AI Techniques

Given Name Surname
dept. name of organization
(of Affiliation)
name of organization
(of Affiliation)
City, Country
email address or ORCID

Abstract— This paper presents a comprehensive approach to interior design using multimodal machine learning models. By combining visual and textual search capabilities, our proposed system, VisualizeVibe, enables users to visualize interior designs in their spaces based on various inputs. The system leverages models like YOLOv8, ConvNeXT, SBERT, and RoBERTa for feature extraction and similarity search, providing a robust and user-friendly tool for interior design enthusiasts and professionals.

I. INTRODUCTION

A. Overview

Interior design significantly impacts the aesthetic and functional aspects of living spaces. Traditional methods of interior design involve manual selection and matching of items, which can be time-consuming and subjective. With advancements in machine learning and artificial intelligence, there is a potential to automate and enhance this process. Our project, VisualizeVibe, aims to develop a multi-modal web application that combines visual and textual search capabilities to help users find and visualize interior designs that best match their preferences**.**

B. Problem Statement

Selecting the right combination of interior design items is a challenging task due to the vast variety of available options and the subjective nature of aesthetic preferences. Existing systems often fail to provide contextually relevant suggestions and do not leverage the full potential of multimodal data. There is a need for an intelligent system that can understand and combine visual and textual information to offer more accurate and personalized design recommendations**.**

C. Scope and Objectives

The scope of this project includes developing a web application that integrates advanced machine learning models for interior design. The primary objectives are:

- To utilize YOLOv8 for object detection in room scenes and interior items.
- To employ ConvNeXT for encoding images.
- To integrate SBERT and RoBERTa for textual feature extraction.
- To perform similarity calculations using cosine similarity for ranking and retrieving relevant items.
- To provide a user-friendly interface for seamless interaction and visualization.

D. Work Methodology

Our approach involves several key steps:

Data Collection and Preparation: Collecting and annotating a diverse dataset of interior design images and corresponding textual descriptions.

Model Selection and Training: Using pre-trained models (YOLOv8, ConvNeXT, SBERT, RoBERTa) and fine-tuning YOLOv8 on our dataset.

Feature Extraction: Extracting visual and textual features from the input data.

Similarity Calculation: Using cosine similarity to match and rank items based on their features.

System Integration: Developing a web application that combines these components into a cohesive user experience.

II. RELATED WORK

Multimodal Search Engine for Interior Design

The paper "What Looks Good with My Sofa: Multimodal Search Engine for Interior Design" by Tautkute et al. presents a pioneering approach to combining visual and textual search modalities in the domain of interior design. The authors developed a web-based search engine that integrates visual and textual queries to retrieve interior objects that share visual and aesthetic similarities with the user's input. This method addresses the limitations of traditional search engines that rely solely on either visual or textual input, thus enhancing the overall search experience.

A. Strengths

Integration of Visual and Textual Search Modalities: The search engine leverages state-of-the-art object detection algorithms, specifically YOLO 9000, combined with deep neural network-based visual search methods. This dual approach ensures high accuracy in detecting and retrieving relevant interior objects. By incorporating both visual and textual inputs, the search engine provides a more comprehensive and user-friendly search experience, catering to diverse user preferences and needs.

Effective Blending Method: The blending method proposed for merging visual and textual search results significantly improves the quality of the search outcomes. The reported 11% improvement in style similarity scores demonstrates the efficacy of this method in producing relevant and aesthetically pleasing results.

Accessible Web-Based Application: The practical implementation of the search engine as a web-based application underscores its real-world applicability and ease of use. This accessibility is a crucial factor in user adoption and satisfaction.

Robust Evaluation Metrics: The authors employed both Hit@k and a custom style similarity score to evaluate the search engine's performance. These metrics provide a thorough assessment of the accuracy and relevance of the search results, ensuring a reliable evaluation framework.

B. Weaknesses

Dataset Limitations: The dataset used in the study is restricted to specific products and styles available from IKEA. This narrow focus may not fully capture the diversity and variety of styles present in the broader interior design landscape. The lack of consideration for the contextual and stylistic similarity of retrieved objects limits the application of the search engine in practical interior design scenarios.

Scalability Issues: The computational demands for processing and blending large volumes of data pose scalability challenges, particularly when scaling up to larger datasets or more complex queries. This limitation could hinder the search engine's performance in more extensive applications. The scalability of the methods is limited and using such representations for long sequences (documents) tends to result in similar token distributions, leading to lower discriminative power and retrieval precision.

Generalization Challenges: The approach is heavily reliant on the specific dataset and object detection model used. Generalizing this method to other datasets or different types of interior design items may require substantial retraining and fine-tuning of the models. SVD assumes linear relationships between features, which limits its capability to capture complex non-linear relationships, affecting the overall retrieval performance.

Limitations of Textual Embedding Models: The CBOW model's approach of averaging word context and its unidirectional nature restrict its effectiveness in capturing the full context of words. The Skip-Grams model, although more effective, has higher memory requirements, impacting the search engine's performance.

User Experience Enhancements: While the web-based application is accessible, the user experience could be significantly improved with features such as real-time feedback, interactive customization, and personalized recommendations. These enhancements are crucial for user satisfaction and engagement.

C. Future Directions

Common Latent Space Mapping: Future research could explore approaches to common latent space mapping, allowing for the mapping of both textual and visual queries to a shared space. This advancement could enable more efficient and accurate similarity searches, enhancing the search engine's performance.

Dataset Expansion and Optimization: Expanding the dataset to include a wider variety of styles and products, along with optimizing the blending methods, could further improve the search engine's applicability and performance. A more diverse dataset would better represent the wide range of interior design styles and preferences.

Enhanced User Experience: Improving the user interface and adding features like real-time feedback, interactive customization, and personalized recommendations could significantly enhance the overall user experience. These enhancements are vital for increasing user engagement and satisfaction.

The paper by Tautkute et al. provides a robust foundation for the development of multimodal search engines in the interior design domain. While there are several strengths, including the integration of visual and textual search modalities, effective blending methods, and practical web-based implementation, there are also notable weaknesses such as dataset limitations, scalability issues, and the need for enhanced user experience. Our project, VisualizeVibe, aims to address these weaknesses by incorporating a more diverse dataset, exploring advanced common latent space mapping techniques, and enhancing user interaction features, thereby building on the foundational work presented in this paper.

By analyzing the strengths and weaknesses of the related work, we can better understand the landscape of current research and identify areas for improvement and innovation in our own project. This analysis informs our approach and guides the development of a more robust, scalable, and user-friendly multimodal search engine for interior design.

III. PROPOSED MODEL

A. Visual search

An object detection algorithm, specifically YOLOv8, is applied to the uploaded image. This algorithm identifies and locates objects of interest within the image, such as chairs, tables, or sofas.

Both a default YOLOv8 model and a custom-trained YOLOv8 model are employed to detect standard objects as well as additional classes that are relevant to interior design.

Once the objects are detected, their regions of interest (ROIs) are extracted as image patches.

The ConvNeXt model is used to extract high-dimensional features (embeddings) from these image patches as well as from the full image.

B. Textual search

Simultaneously, the textual query is processed using advanced NLP models, SBERT and RoBERTa, to generate high-quality sentence embeddings that capture the semantic meaning of the query.

C. Search engine

The features from the full image and the ROIs are combined using an attention mechanism. This mechanism assigns higher weights to features from regions with higher detection confidence, creating robust image embeddings.

The embeddings from SBERT and RoBERTa are ensembled to leverage the strengths of both models, resulting in a more comprehensive semantic representation of the text.

The combined image and text embeddings are used to calculate similarity scores with precomputed embeddings from the dataset.
Cosine similarity is employed to measure the similarity between the query embeddings and the dataset embeddings. The similarity scores from the visual and textual modalities are then combined using a late fusion technique.
The search results are ranked based on these combined similarity scores, ensuring that the most relevant items are presented to the user.

TABLE I. THIS IS THE HEADING FOR A TABLE

[a.] This is a table footnote.

You can cite your references in text by including the corresponding number, in square brackets [1]. If you need to cite a specific part of the source, you can include a page number [2, p. 13] or range [3, pp. 41–56].

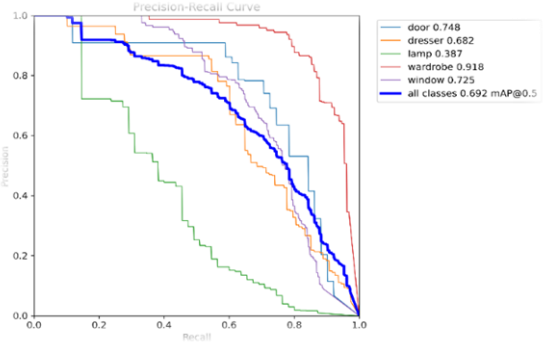
ACKNOWLEDGMENTS

IV. DATASET

In order to evaluate our proposed Style Search Engine, we collected our dataset images from Google and write it's description. Although several datasets for standard visual search methods exist, e.g. COCO dataset.
-it includes furniture items such as (chair,couch).
-room scenes such as (bedroom, living room).

The dataset is divided into 14 categories based on the room class and items: Living Room, Bedroom, Chair, Clock, Bed, Couch, Dining Table, Potted Plant, Door, Dresser, Lamp, Wardrobe, Window, table.

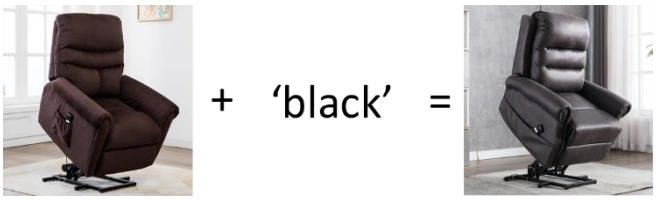
Below are the visualizations of the training results:



This template was adapted from those provided by the IEEE on their own website.

V. RESULTS

The top-ranked search results are visualized, providing an intuitive interface for users to see the most relevant images and items based on their query.



- **Training Details (YOLOv8):**

The training was conducted for a total of 50 epochs with a batch size of 8 and an image size of 640 for 5 classes (door, dresser, lamp, wardrobe, window). The optimizer was set to auto, with AdamW chosen and a learning rate of 0.001111. Mixed precision training was enabled using AMP, and a patience parameter of 100 was set to manage early stopping. The model was pretrained, and various data augmentation techniques were employed, including Blur, MedianBlur, ToGray, CLAHE, Mosaic, and others. Additional YOLO-specific parameters included settings such as box=7.5, cls=0.5, dfl=1.5, pose=12.0, among others. The model architecture consisted of 225 layers, with 11,137,535 parameters and 28.7 GFLOPs. The training time for one epoch was approximately 230 seconds using a Tesla T4 graphics card. The AdamW optimizer was used with an initial learning rate determined automatically (lr=0.001111). TensorBoard was enabled for monitoring. The model employed various data augmentation techniques such as Blur, MedianBlur, ToGray, CLAHE, and Mosaic. The final model configuration included details like specific layers, parameters, and module arguments.

Below are the performance metrics captured during the selected epochs:

- **Visual Search:**

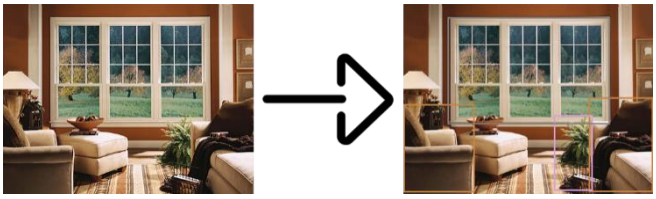
The visual search system demonstrated robust performance in detecting and retrieving relevant objects and descriptions. The combination of advanced object detection models, attention-based feature extraction, and effective similarity calculation techniques contributed to the overall success of the system. The integration of multimodal embeddings and late fusion further enhanced the retrieval accuracy, providing a comprehensive solution for visual search tasks.

Object Detection and Feature Extraction:

- **Object Detection:**

The ConvNeXTv1 model was used to extract high-dimensional features from the entire image. In addition, the regions of interest (ROIs) detected by YOLOv8 were expanded and passed through ConvNeXTv1 to extract detailed features. We chose ConvNeXTv1

because of its efficiency in feature extraction compared to other feature extraction models.



Feature Extraction:

The default YOLOv8 model was utilized to detect objects from the COCO dataset, focusing on key classes such as 'bed', 'chair', 'couch', 'dining table', 'clock', and 'potted plant'. Additionally, a custom-trained YOLOv8 model was specifically designed to detect additional classes like 'door', 'dresser', 'lamp', 'wardrobe', and 'window'. The models were evaluated based on their ability to detect these objects in various images, with high-confidence detections visualized with bounding boxes and annotated with the class name and confidence score.

Epoch	Classification Loss	Precision (P)	Recall (R)	mAP50	mAP50-95
46	0.6079	0.646	0.641	0.684	0.468
47	0.5832	0.686	0.586	0.684	0.469
48	0.552	0.708	0.615	0.688	0.481
49	0.5281	0.74	0.587	0.694	0.476
50	0.5301	0.72	0.606	0.692	0.481

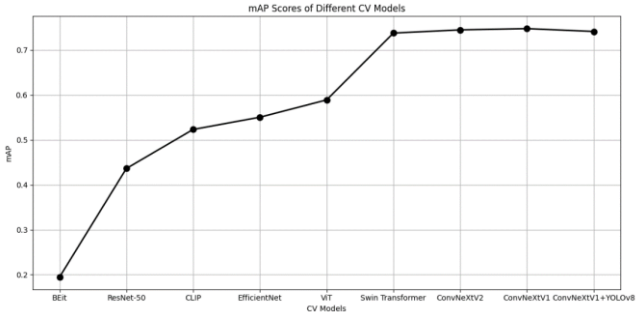


Figure 1: Line plot showing mean Average Precision (mAP) scores of different computer vision models. The plot compares various models, such as BEiT, ResNet-50, CLIP, EfficientNet, ViT, Swin Transformer, ConvNeXtV2, ConvNeXtV1, and ConvNeXtV1 combined with YOLOv8.

- **Textual Search:**
  - The textual search system demonstrated strong performance in retrieving relevant text descriptions based on semantic similarity. The combination of advanced NLP models (SBERT and RoBERTa) and their ensemble embeddings significantly improved the accuracy and relevance of the search results. The use of late fusion further enhanced the retrieval performance, making the system robust and effective for textual search tasks. The visualizations provided clear evidence of the semantic clustering capabilities of the embeddings, supporting the overall effectiveness of the textual search system.

Text Embedding and NLP Models:

- **Text Descriptions:**

The text descriptions were encoded using two advanced NLP models, SBERT (Sentence-BERT) and RoBERTa. SBERT is specifically designed for creating semantically meaningful sentence embeddings, while RoBERTa is an optimized version of BERT, which is robust in handling diverse text inputs. By using these models, high-dimensional embeddings for each text description were generated. To take advantage of the strengths of both SBERT and RoBERTa, their imputations were combined using a weighted averaging technique. This combined approach aims to capture a more comprehensive semantic representation of text descriptions.

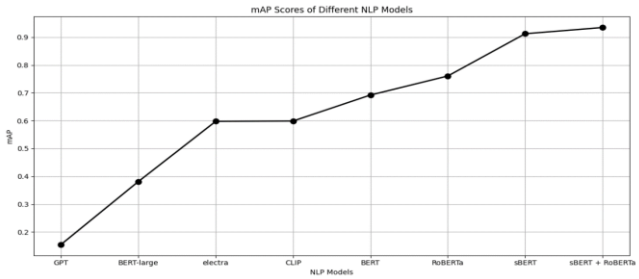


Figure 2: Line plot showing mAP scores of different NLP models. Models compared include GPT, BERT-large, Electra, CLIP, BERT, RoBERTa, SBERT, and a combination of SBERT and RoBERTa.

- **Fusion:**

The fusion techniques employed in the system significantly improved the robustness and accuracy of the feature representations and embeddings. The attention-based feature combination effectively integrated global and local information, while the ensemble embeddings leveraged the strengths of both SBERT and RoBERTa models. Late fusion further enhanced retrieval performance by combining similarity scores from different modalities, resulting in a comprehensive and effective multimodal search system. The visualizations provided clear evidence of the improved semantic clustering and class separability achieved through fusion, supporting the overall effectiveness of the system.

Late Fusion for Similarity Calculation:

- **Similarity Scores:**

Late fusion involved combining similarity scores from different modalities (image and text). This was achieved by averaging the similarity scores from the image embeddings and the text embeddings. The resulting combined similarity score provided a more accurate measure of similarity, improving the overall retrieval performance.

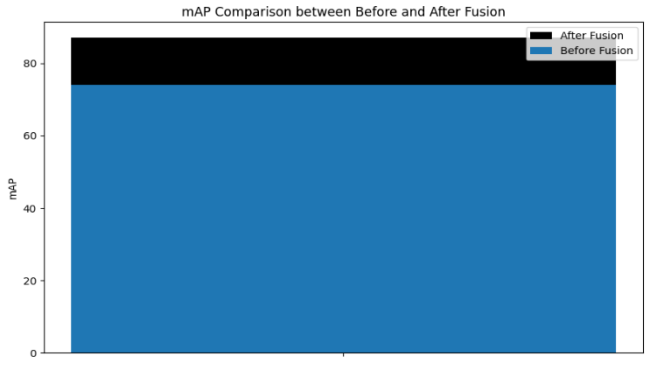


Figure 3: compares the mean Average Precision (mAP) before and after a fusion process(CV).

- **Fusion Techniques:**
  Various fusion techniques were compared, including Hybrid, Deep Cross-Modal, Hierarchical, Early, Intermediate, Slow, Tensor, and Late fusion. Late fusion outperformed the other methods, suggesting that integrating modalities at a later stage preserves more information from each modality, leading to better performance.

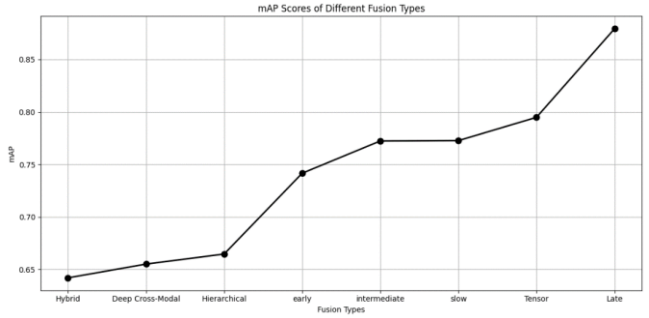


Figure 4: Line plot showing mAP scores of different fusion types.This graph explores various fusion techniques, such as Hybrid, Deep Cross-Modal, Hierarchical, Early, Intermediate, Slow, Tensor, and Late Fusion.

- **t-SNE Visualization of Fused Embeddings::**
  The t-SNE plot was used to visualize the fused embeddings in a 2D space. Each cluster in the t-SNE plot represents a different class, indicating how well the fusion process maintained class separability and semantic grouping.

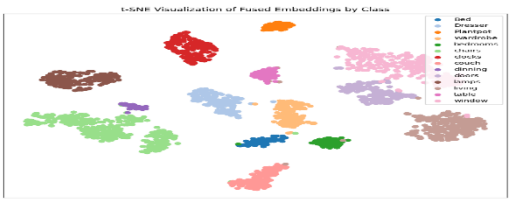


Figure 5:Each color represents a different class, indicating how well the embeddings cluster similar items together.

VI. REFRENCE

[1] Tautkute, Ivona, et al. "What looks good with my sofa: Multimodal search engine for interior design." 2017 Federated Conference on Computer Science and Information Systems (FedCSIS). IEEE, 2017. .

[2] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

[3] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert networks." arXiv preprint arXiv:1908.10084 (2019).

[4] Reis, Dillon, et al. "Real-time flying object detection with YOLOv8." arXiv preprint arXiv:2305.09972 (2023).

[5] Liu, Zhuang, et al. "A convnet for the 2020s." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[6] Tautkute, Ivona, et al. "Deepstyle: Multimodal search engine for fashion and interior design." IEEE Access 7 (2019): 84613-84628.

[7] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[8] Bao, H., Dong, L., Piao, S. and Wei, F., 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

[9] Koonce, B. and Koonce, B., 2021. ResNet 50. Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization, pp.63-72.

[10] Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J. and Zavolan, M., 2021. CLIP and complementary methods. *Nature Reviews Methods Primers*, *1*(1), pp.1-23.

[11] Koonce, B. and Koonce, B., 2021. EfficientNet. Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization, pp.109-123.

[12] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J. and Yan, S., 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 558-567).

[13] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022).

[14] Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S. and Xie, S., 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16133-16142).

[15] Nath, S., Marie, A., Ellershaw, S., Korot, E. and Keane, P.A., 2022. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. *British Journal of Ophthalmology*, *106*(7), pp.889-892.

[16] Wang, S., Guo, Y., Wang, Y., Sun, H. and Huang, J., 2019, September. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics* (pp. 429-436).

[17] Clark, K., Luong, M.T., Le, Q.V. and Manning, C.D., 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

[18] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[19] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H. and Ng, A.Y., 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 689-696).

[20] Boulahia, S.Y., Amamra, A., Madi, M.R. and Daikh, S., 2021. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, *32*(6), p.121.

[21] Zadeh, A., Chen, M., Poria, S., Cambria, E. and Morency, L.P., 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

[22] Bethe, H.A., 1979. The fusion hybrid. *Physics Today*, *32*(5), pp.44-51.

[23] Majumder, N., Hazarika, D., Gelbukh, A., Cambria, E. and Poria, S., 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-based systems*, *161*, pp.124-133.

[24] Hosseinpour, H., Samadzadegan, F. and Javan, F.D., 2022. CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images. *ISPRS journal of photogrammetry and remote sensing*, *184*, pp.96-115.