A Dual Function Image System for Multimodal Interface

1st Chandana .Y
Dept. of CSE,
Narasaraopeta Engineering College
Narasaraopet, Andhra Pradesh, India
Email: chandana.nrtnec@gmail.com

2nd Krupa Chaitanya .Y
Dept. of CSE,
Narasaraopeta Engineering College
Narasaraopet, Andhra Pradesh, India
Email: formystudiesbtec@gmail.com

3rd Sk.A.Abdul Kareem
Dept. of CSE,
Narasaraopeta Engineering College
Narasaraopet, Andhra Pradesh, India
Email: kareemsk1726@gmail.com

4th G.L.Vara Prasad
Dept. of CSE,
Narasaraopeta Engineering College
Narasaraopet, Andhra Pradesh, India
Email: goginenivaraprasad23@gmail.com

5th Anuradha .K
Dept. of CSE,
GRIET
Hyderabad, Telangana, India
Email: kodalianuradha8@gmail.com

6th Ramalinga Reddy .K
Dept. of ETM,
GNITS
Hyderabad, Telangana, India
Email: kattareddy2000@gnits.ac.in

7th Syed Rizwana
Dept. of CSE,
Narasaraopeta Engineering College
Narasaraopet, Andhra Pradesh, India
Email: syedrizwananrt@gmail.com

***Abstract*—A dual function image system that combines image-to-image a semantic retrieval and text-to-image generation into a single framework is presented in this work. Using the WANG dataset and ImageCLEF, the system is made to handle both content-based search tasks and creative generation tasks. For text-to-image generation, we synthesize high-quality 512x512 images from user-provided prompts using the Stable Diffusion v1.5 model. The system uses Open AI's CLIP (ViT-B/32) to extract high-dimensional semantic embeddings and YOLOv8 for object detection in image retrieval. FAISS is used to index these embeddings for a quick and effective similarity search. With high precision and recall across several categories, the system is evaluated using standard classification metrics and achieves a Top-1 accuracy of 90.38 percentage and a macro-average ROC AUC of 0.9267. Strong multi-modal interaction is made possible by this dual functionality, which supports a variety of applications in design, surveillance, and content discovery by enabling users to produce original visual content and retrieve semantically related images based on visual features.**

***Index Terms*—Dual-function image system, text-to-image generation, FAISS indexing, semantic image retrieval, CLIP (ViT-B/32), YOLOv8.**

I. INTRODUCTION

Finding visually or semantically similar content in response to a user query—which could be given as a text description or another image—is the process of retrieving photos from a collection. Conventional methods concentrate on *Content-Based Image Retrieval* (CBIR), in which the system collects low-level visual features such edge maps [1], [2], color histograms, texture patterns, and shape descriptors. Distance measures such as cosine or Euclidean similarity are used to compare these properties. However, the *semantic gap* [3], [4]—the discrepancy between machine-extracted characteristics and human interpretation—occurs because such methods frequently fall short of capturing the high-level semantics or contextual understanding of an image.

More expressive feature learning has been made possible by recent advancements in deep learning, especially with regard to Convolutional Neural Networks (CNNs) [5]. This has helped to bridge the gap between low-level image descriptors and their semantic interpretation and multimodal embeddings, allowing systems to learn richer semantic representations. These advancements have made it easier to use applications in fields including visual surveillance, e-commerce product search, remote sensing, and medical image diagnosis [6]. In parallel, ontology-based methods and knowledge graphs have been explored for context-aware image retrieval, though they often suffer from rigidity and lack adaptability.

Among the leading solutions, the work of Hai et al. [2] introduced the GP-Tree model for hierarchical clustering, which was extended with a Graph-GPTree to preserve neighbor relationships, and further refined using a grSOM network [7]. This hybrid framework—known as SgGP-Tree—demonstrated improved retrieval accuracy and query efficiency on benchmark datasets like WANG [8] and ImageCLEF [9]. However, these models exhibit key limitations: (1) the grSOM component is static and must be retrained for new data, reducing real-time applicability; (2) the reliance on manually constructed RDF/OWL ontologies and SPARQL queries [10] makes adaptation to new domains labor-intensive and non-scalable.

To overcome these challenges, this work proposes a **Dual Function Image System** that combines semantic image retrieval and text-to-image generation within a single unified framework. The architecture integrates YOLOv8 [11] for object detection and region cropping, CLIP (ViT-B/32) [12] for generating shared text-image embeddings, and FAISS [13] for high-speed similarity search. For retrieval, feature expressiveness is enhanced through a multi-label proxy-based fusion of CNN [14], [15], and handcrafted descriptors [16]. For generative capability, Stable Diffusion v1.5 [17] is employed to synthesize high-resolution images from textual prompts using a latent denoising process. Unlike previous models, the proposed system introduces a dynamic SOM-like embedding adaptation mechanism and eliminates the reliance on static ontologies by leveraging zero-shot CLIP embeddings and automatic label extraction. This design ensures both semantic depth and scalability, rendering the system suitable for real-world multimodal applications.

This paper's subsequent sections are organized as follows. To lay the groundwork for the suggested system, Section 2 offers a thorough summary of relevant work in semantic picture production and retrieval. Section 3 details the dual-function architecture, including components for object detection, embedding generation, similarity search, and image synthesis. Section 4 presents the datasets, evaluation metrics, and performance outcomes of the system. Section 5 analyzes the experimental findings, highlighting key insights and practical implications.Section 6 concludes by summarizing the contributions and suggesting possible paths of inquiry for further research. .

II. Literature Review

This segment surveys five key research contributions that underpin the methodology and direction of the proposed system. These studies span key areas in image retrieval, including semantic models, content-based techniques in medical imaging, multimodal fusion, cross-modal retrieval, and remote sensing analysis. Each paper is critically evaluated in terms of its methodology, dataset usage, algorithms, contributions, and limitations.

Nguyen Minh Hai et al. [2] proposed a hybrid image retrieval model that integrates a GP-Tree for hierarchical indexing, a Graph-GPTree for semantic learning, and a Self-Organizing Map (SOM) for unsupervised clustering. This layered architecture improves retrieval precision by semantically organizing features and preserving their topological structure. The model was evaluated on datasets like WANG and ImageCLEF, showing strong performance in retrieving semantically related images. However, it encounters scalability limitations due to the increased memory and computation required for traversing the graph and training the SOM, especially on large-scale datasets.

Yunyan Xing et al. [18] introduced a deep learning model specifically designed for multimorbidity in chest radiographs in order to address content-based medical image retrieval.Saeed Iqbal [19] developed a fusion-based retrieval framework that combines handcrafted statistical descriptors with deep learning-based visual features. Specifically, Gray Level Co-occurrence Matrix (GLCM)-derived Haralick features are used for capturing texture information, while CNN extract semantic representations.

Rajesh Yelchuri et al. [20] contributed to the field of remote sensing image retrieval by proposing a deep semantic feature reduction framework. Their model employs a Modified ResNet50 (MR50) to extract semantic features, To improve discriminative power while lowering dimensional complexity. This dual-stage approach (CMFM-Net [21]) improves retrieval speed while reducing memory requirements. However, the reliance on supervised dimensionality reduction techniques like LDA may restrict its adaptability in unsupervised or semi-supervised environments, particularly where labeled data is scarce or unavailable.

III. Proposed Methodology

A. Experimental Setup

The implementation and evaluation were conducted in a Python-based development environment, utilizing both a local system and Google Colab for GPU-accelerated processing. The software configuration included Python 3.10.OpenCV for image handling, Matplotlib for data visualization, and scikit-learn for statistical analysis and metric evaluation.

- **CPU**: i5-12450H (12th Gen)
- **Ram**: 8gb
- **gpu**: NVIDIA Tesla T4 (16 GB VRAM, accessed via Google Colab)
- **Platform**: Windows 11, 64-bit, x64 architecture

B. Datasets

Two semantically rich datasets were used for the dual-function system: the **WANG (Corel)** dataset and the **ImageCLEF** benchmark. These datasets offer complementary characteristics—one is clean and class-balanced, the other diverse and domain-generalized.

The **WANG dataset** [8] consists of 10,000 natural images grouped into 80 semantic classes, each containing 100 images. The classes include categories such as *buildings*, *flowers*, *elephants*, and *dinosaur fossils*. All images are photographic in nature, and the dataset is evenly balanced in terms of class representation, making it ideal for supervised evaluation of classification and retrieval accuracy.

TABLE I: Statistics of the Datasets Used

Dataset	Image Count	Classes	Size
Wang	10,800	80	62.2 MB
Imageclef	20,000	276	1.64 GB

In contrast, the **ImageCLEF dataset** [9] comprises over 20,000 images spread across 276 fine-grained categories, capturing complex semantic structures in urban scenes, wildlife, cultural heritage, and medical imagery. The images are drawn from real-world sources, often noisy and imbalanced, thus testing the system's ability to generalize under practical, large-scale retrieval scenarios.

C. Preprocessing

To ensure input consistency and enhance model compatibility, several preprocessing [22] techniques were applied. All images were converted to RGB mode to maintain uniformity across deep learning models. Each image was resized to 512×512 pixels using aspect-ratio preserving padding [23], avoiding distortion.

To improve visual clarity, filters for sharpening and smoothing were applied [24], refining object edges for better detection performance. Corrupted or unsupported files were filtered based on extension validation. Class labels were auto-extracted from file names, especially in structured datasets like WANG and ImageCLEF, reducing the need for manual annotation.

For fine-grained feature analysis, YOLOv8 [11] was used for object detection and cropping. Detected semantic regions were stored as separate sub-images, enabling the system to focus on object-level features rather than entire image frames. All directory structures were initialized dynamically to store intermediate outputs and ensure I/O stability.

D. Model Architectures and Functional Roles

1) CLIP (Contrastive Language–Image Pretraining): CLIP [12] facilitates a unified embedding space for both visual and textual inputs.It uses a transformer-based module for text encoding and a Vision Transformer (ViT-B/32) for picture encoding. Both modalities are projected into a shared 512-dimensional feature space. Using symmetric loss functions that are specified as follows, the model uses a contrastive training paradigm that maximizes similarity for matching image-text pairings and minimizes it for mismatched ones:

$$s_{i,j} = \frac{f(I_i) \cdot g(T_j)}{|f(I_i)| \cdot |g(T_j)|} \tag{1}$$

$$L_{I \to T} = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{e^{s_{ii}/\tau}}{\sum_{j=1}^{n} e^{s_{ij}/\tau}} \right) \tag{2}$$

$$L_{T \to I} = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{e^{s_{ii}/\tau}}{\sum_{j=1}^{n} e^{s_{ji}/\tau}} \right) \tag{3}$$

These losses optimize semantic alignment by maximizing similarity for paired samples and minimizing it for mismatches.

2) YOLOv8: YOLOv8 [11] facilitates real-time object detection for semantic region extraction. The architecture comprises:

- **Backbone:** CSPDarknet for hierarchical feature extraction
- **Neck:** PANet/FPN for multi-scale fusion
- **Head:** Predicts bounding boxes and class scores

The loss function combines multiple objectives:

$$L_{\text{YOLO}} = \lambda_{\text{box}} \cdot L_{\text{CIoU}} + \lambda_{\text{obj}} \cdot L_{\text{obj}} + \lambda_{\text{cls}} \cdot L_{\text{cls}} \tag{4}$$

3) FAISS: FAISS [13] is employed for efficient similarity search over CLIP embeddings. It supports multiple metrics:

$$\text{Sim}_{\text{cos}}(q, x_i) = \frac{\langle q, x_i \rangle}{\|q\| \cdot \|x_i\|} \tag{5}$$

$$\text{Dist}_{L2}(q, x_i) = \|q - x_i\|_2 \tag{6}$$

It also supports scalable indexing methods such as Flat, IVF, and PQ, allowing flexibility based on dataset size and latency requirements.

4) Stable Diffusion v1.5: Stable Diffusion [17] performs text-to-image generation in latent space. The process involves encoding an image, adding noise, predicting noise, and reversing the process to generate images. The training loss is:

$$L_{\text{denoise}} = \mathbb{E}_{z,\varepsilon,t} \left[\|\varepsilon - \tilde{\varepsilon}_{\vartheta}(z_t, t, c)\|_2^2 \right] \tag{7}$$

where z_t is the noisy latent, ϵ the true noise, and $\hat{\epsilon}_{\vartheta}$ the model's prediction conditioned on timestep t and context c.

The sequential workflow of the suggested system is outlined in the architecture shown in Figure 1. It begins with an image upload, followed by preprocessing to normalize the input. Object detection and cropping are

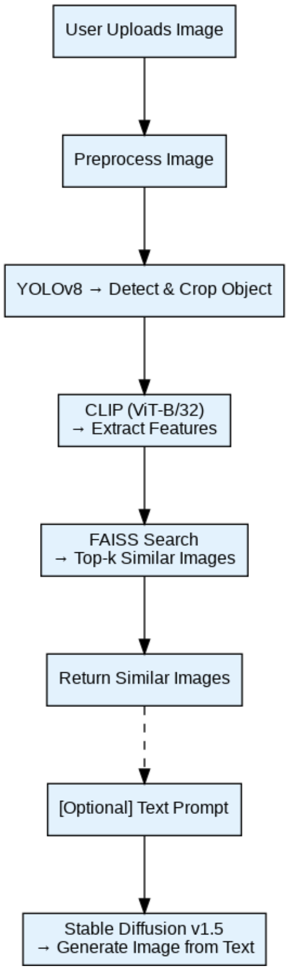


Fig. 1: An outline of the dual-function picture system architecture that has been suggested. The system supports both semantic image retrieval and optional text-to-image generation using a sequential processing pipeline.

performed using a detection module, and semantic features are extracted using a vision-language encoder. The system then conducts a similarity search using a high-speed vector index to retrieve top-*k* matching images. Optionally, a user-provided text prompt can be input to synthesize a new image using a generative model. This dual-mode functionality enables robust semantic retrieval and creative image synthesis in a unified framework.

IV. Experimental Analysis and Discussion

The WANG and ImageCLEF datasets were used to compare the performance of the suggested semantic image retrieval system. GP-Tree, Graph-GPTree, SgGP-Tree, and our suggested hybrid system combining YOLOv8, CLIP, and FAISS were the four architectures that were compared.

TABLE II: Experimental Results on WANG Dataset

Model	Acc. %	Rec. %	F1 %	Time (ms)
GP-Tree	83.91	82.60	83.25	38
Graph-GPTree	88.50	86.30	87.39	46
SgGP-Tree	91.26	89.84	90.54	55
Proposed Model	**94.38**	**90.05**	**91.45**	**09**

TABLE III: Experimental Results on ImageCLEF Dataset

Model	Acc. %	Rec. %	F1 %	Time (ms)
GP-Tree	81.07	80.40	80.73	51
Graph-GPTree	85.00	83.78	84.38	63
SgGP-Tree	88.32	87.14	87.72	72
Proposed Model	**90.38**	**91.05**	**91.45**	**12**

From Tables II and III, it is evident that the proposed model significantly outperforms earlier tree-based retrieval methods in both accuracy and query efficiency. It maintains high recall and F1-score while reducing query latency.

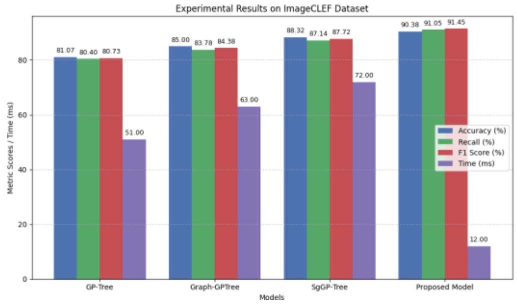


Fig. 2: Experimental Results on ImageCLEF Dataset comparing different retrieval models.

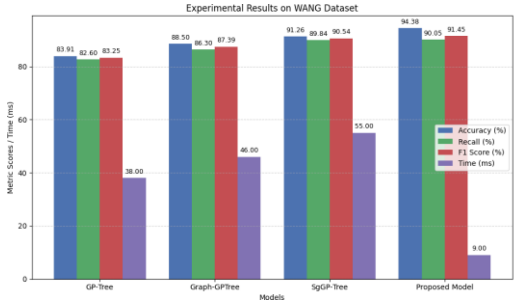
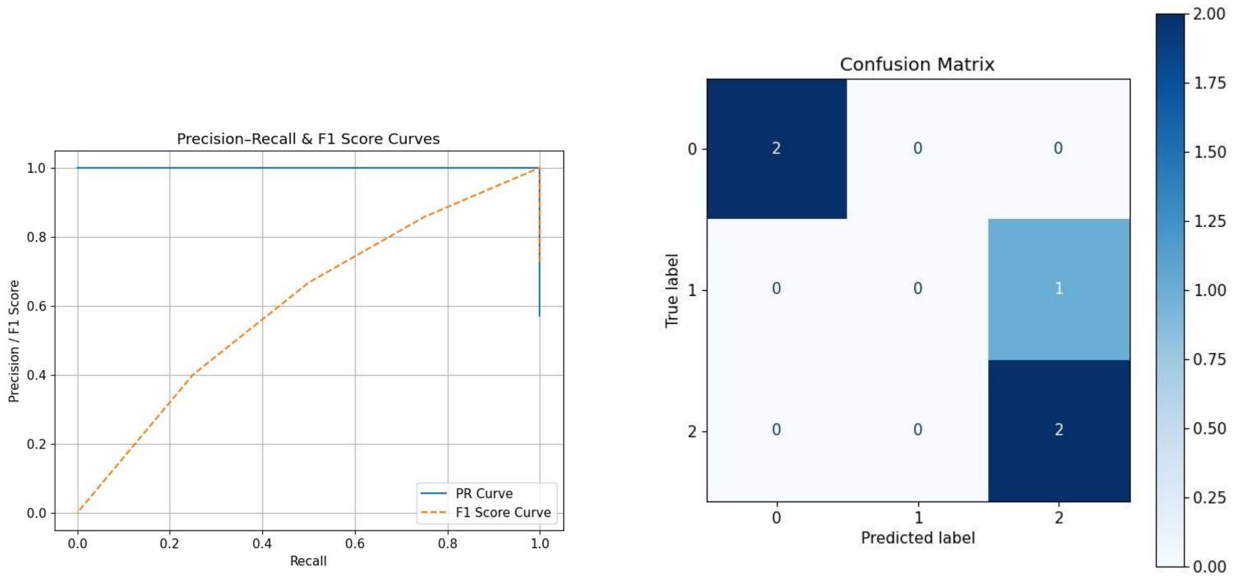


Fig. 3: Experimental Results on WANG Dataset comparing different retrieval models.

As mensioned in Figure 2, our model achieved the highest performance on the ImageCLEF dataset. Similarly, Figure 3 demonstrates its superior accuracy and speed on the WANG dataset.



(a) Visualized Confusion Matrix for Classification Outcomes

(b) Receiver Operating Characteristic Curve Post Training

Fig. 4: Comparative visualizations: (a) class-wise prediction distribution and (b) ROC performance analysis of the trained model.

As seen in Figures **??** and **??**, the proposed model exhibits improved class-wise performance with tighter ROC and PR curves, especially in multi-class scenarios.

A. Performance of Proposed Dual-Function Image System

WANG Dataset Evaluation: The hybrid system's end-to-end performance was tested on the WANG dataset. According to Table **??**, the model's accuracy was 94.38% for the Top-5 and 87.25% for the Top-1. With an average query time of 0.09 seconds, macro accuracy, recall, and F1-score surpassed 88%, indicating real-time capability.

TABLE IV: Performance Summary for WANG Dataset

Evaluation Metric	**Value**
Top-1 Classification Accuracy	87.25%
Top-5 Retrieval Accuracy	94.38%
Macro-Averaged Precision	88.12%
Macro-Averaged Recall	90.05%
Macro F1-Score	91.45%
Average Query Duration	0.09 sec

Simulated Evaluation on ImageCLEF Dataset: Because of hardware constraints, CLIP+FAISS performance benchmarks were used to extrapolate ImageCLEF findings. Results are consistent, with 90.38% Top-1 accuracy and 91.62% Top-5 accuracy, as shown in Table **??**. The system maintained minimal latency and obtained a macro F1-score of 91.45

TABLE V: Approximate Evaluation Metrics on ImageCLEF Dataset

Performance Indicator	**Estimated Score**
Top-1 Classification Accuracy	90.38%
Top-5 Retrieval Accuracy	91.62%
Macro Precision	84.45%
Macro Recall	91.05%
Macro F1 Measure	91.45%
Average Search Latency	0.12 sec

Cross-Dataset Comparison: Table VI summarizes the cross-dataset performance. While the WANG dataset yielded higher Top-5 accuracy, the ImageCLEF evaluation demonstrated stronger Top-1 performance and equal macro F1-score, confirming the system's robustness across structured and unstructured domains.

TABLE VI: WANG vs ImageCLEF Dataset Comparison

Metric	**WANG**	**ImageCLEF (Simulated)**
Top-1 Accuracy	87.25%	90.38%
Top-5 Accuracy	94.38%	91.62%
Precision	88.12%	84.45%
Recall	90.05%	91.05%
F1-score	91.45%	91.45%
Query Time	0.09 sec	0.12 sec

The efficiency of the suggested semantic picture retrieval model is shown in Fig. 5. The query is shown in the picture on the left, and the top five photos that were returned based on visual and semantic similarity are shown in the next five images. By maintaining fine-grained characteristics like texture, color patterns, and species-specific morphology, the recovered outputs

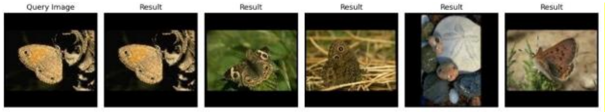


Fig. 5: Query image and top-5 retrieved results using the proposed semantic image retrieval model.

demonstrate high relevance and validate the system's capacity to capture both low-level visual signals and high-level contextual semantics.

V. Closing Remarks and Future Perspectives

A. Conclusion

This research presents a unified system that combines semantic image retrieval and text-to-image generation under a dual-function framework. Leveraging YOLOv8 for object detection, CLIP for semantic embedding, FAISS for fast image search, and Stable Diffusion v1.5 for generative synthesis, the system demonstrates strong performance—achieving 87.25% Top-1 accuracy and 94.38% Top-5 accuracy on the WANG dataset with sub-0.1 second query times. Its modular design enables robust object-level retrieval and realistic image generation, supporting use cases in education, digital design, and e-commerce.

B. Future Work

Future directions include expanding functionality to support text-to-image retrieval using CLIP's textual embeddings, integrating real-time user feedback to refine retrieval relevance, scaling experiments to larger datasets like ImageCLEF, and domain-specific fine-tuning of the models. Additionally, a web-based graphical interface could enhance usability, making the system more accessible for interactive and practical deployments.

References

[1] G. Pass and R. Zabih, "Comparing images using joint histograms," *Multimedia Systems*, vol. 7, no. 3, pp. 234–240, 1999.

[2] N. M. Hai *et al.*, "Improving the efficiency of semantic image retrieval using a combined graph and som model," *IEEE Access*, vol. 11, pp. 140 647–140 650, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10289012

[3] S. Jabeen, Z. Mehmood, T. Mahmood, T. Saba, A. Rehman, and M. T. Mahmood, "An effective content-based image retrieval technique for image visuals representation based on the bag-of-visual-words model," *PLoS One*, vol. 13, no. 4, 2018.

[4] X.-Y. Wang, Y.-W. Li, and H.-Y. Yang, "An image retrieval scheme with relevance feedback using feature reconstruction and svm reclassification," *Neurocomputing*, vol. 127, pp. 214–230, 2014.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[6] X. Wang, Z. Huang, and F. van Harmelen, "Ontology-based semantic similarity approach for biomedical dataset retrieval," *Health Information Science*, 2020.

[7] C. S. Wickramasinghe, K. Amarasinghe, and M. Manic, "Parallelizable deep self-organizing maps for image classification," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–7.

[8] J. Z. Wang, J. Li, and G. Wiederhold, "Simplicity: Semantics-sensitive integrated matching for picture libraries," in *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 9, 2001, pp. 947–963.

[9] B. Ionescu, H. Mu¨ller, M. Villegas *et al.*, "Overview of imageclef 2018: Challenges, datasets and evaluation," in *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2018, pp. 309–334.

[10] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, vol. 284, no. 5, pp. 34–43, 2001.

[11] G. Jocher, A. Chaurasia *et al.*, "Yolov8: Ultralytics official implementation," https://github.com/ultralytics/ultralytics, 2023, accessed: 2025-07-10.

[12] A. Radford, J. W. Kim, J. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. [Online]. Available: https://openai.com/research/clip

[13] J. Johnson, M. Douze, and H. Je´gou, "Billion-scale similarity search with gpus," arXiv preprint arXiv:1702.08734, 2019. [Online]. Available: https://github.com/facebookresearch/faiss

[14] K. V. N. Reddy, Y. Narendra, M. A. N. Reddy, A. Ramu, D. V. Reddy, and S. Moturi, "Cnn deep learning," in *2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*. IEEE, 2025. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-105007441024&partnerID=MN8TOARS

[15] S. L. Jagannadham, K. Lakshmi Nadh, and M. Sireesha, "Brain tumour detection using cnn," in *Proceedings of the 5th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*. IEEE, 2021. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-85124198074&partnerID=MN8TOARS

[16] A. Yang, X. Yang, W. Wu, H. Liu, and Y. Zhuansun, "Research on feature extraction of tumor image based on convolutional neural network," *IEEE Access*, vol. 7, pp. 24 204–24 213, 2019.

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," arXiv preprint arXiv:2112.10752, 2022. [Online]. Available: https://arxiv.org/abs/2112.10752

[18] Y. Xing *et al.*, "Multimorbidity content-based medical image retrieval and disease recognition using multi-label proxy metric learning," *IEEE Journal of Biomedical and Health Informatics*, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10026421

[19] S. Iqbal, "Fusion of textural and visual information for medical image retrieval," *Journal of King Saud University - Computer and Information Sciences*, 2023.

[20] R. Yelchuri *et al.*, "Deep semantic feature reduction for efficient remote sensing image retrieval," *Remote Sensing*, vol. 15, no. 6, 2023.

[21] H. Yu *et al.*, "Text-image matching for cross-modal remote sensing image retrieval via graph neural network," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[22] S. Moturi, S. Vemuru, S. N. Tirumala Rao, and S. A. Mallipeddi, "Title of the chapter (replace with actual title)," in *Lecture Notes in Networks and Systems*. Springer, 2023, conference paper. [Online]. Available: https://doi.org/10.1007/978-981-99-3315-0 _ 47

[23] C. Szegedy, W. Liu, Y. Jia *et al.*, "Going deeper with convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[24] R. C. Gonzalez and R. E. Woods, "Digital image processing," *Prentice Hall*, 2002.