Grenze International Journal of Engineering and Technology, January Issue

Sketch-based Image Retrieval using Unsupervised Learning

Atharva Rathi[1] and Dr. Dipali Kasat[2]
[1]Shri Ramdeobaba College of Engineering and Management, Nagpur, India
Email: atharvarathi2004@gmail.com
[2]Sarvajanik College of Engineering & Technology, Surat, India
Email: dipali.kasat@scet.ac.in

***Abstract*— Sketch-Based Image Retrieval (SBIR) enables users to query image databases using freehand sketches, offering an intuitive alternative when textual input is ambiguous or unavailable. However, the practical deployment of SBIR systems is impeded by the considerable domain gap between abstract sketches and detailed natural images, the limited availability of annotated sketch-photo pairs, and the high variability inherent in human drawing styles. Recent advances in unsupervised learning have sought to address these limitations by leveraging unpaired and unlabeled data to learn transferable feature representations. This paper presents and evaluates contemporary unsupervised SBIR techniques, including instance-level contrastive learning, transformer-based sketch encoders, and style-invariant frameworks aimed at improving robustness to intra-class variations and noise. Furthermore, data-free and zero-shot retrieval approaches facilitate retrieval across unseen categories without requiring manual annotations. Additional strategies such as deep manifold alignment and sketch-to-photo synthesis further improve cross-modal alignment. Experimental results on benchmark datasets demonstrate the effectiveness of these methods in real-world, noisy, and unpaired scenarios. The review also explores future research directions, including multi-modal composite queries and meta-learning strategies, which aim to enable scalable and generalizable SBIR systems. Overall, unsupervised learning remains a central component in advancing SBIR by minimizing reliance on labeled data while enhancing retrieval performance.**

***Index Terms*— Sketch-Based Image Retrieval (SBIR), unsupervised learning, self-supervised learning, cross-modal embedding, zero-shot retrieval, feature representation, deep learning.**

I. Introduction

The increasing demand for intuitive and efficient visual search systems has led to growing interest in Sketch-Based Image Retrieval (SBIR). Unlike conventional image retrieval approaches that rely on textual descriptions or exemplar images, SBIR allows users to perform queries using freehand sketches. This modality offers a more natural and flexible form of interaction, especially in scenarios where users may lack access to reference images or precise keywords [14].

Despite its potential, SBIR presents several critical challenges. A major limitation arises from the substantial domain gap between sketches and natural images—while sketches are abstract, sparse, and monochromatic, photographs are dense in texture, color, and structure [1], [14].

Grenze ID: 01.GIJET.12.1.504_2

Bridging this cross-modal disparity requires advanced feature alignment strategies [16], and the lack of large-scale paired datasets further restricts the applicability of supervised learning techniques, which depend heavily on annotated data [1], [7]. Additionally, significant variability in sketching styles introduces intra-class diversity that hinders robust retrieval, even within the same category [3].
Recent advances in unsupervised and self-supervised learning have shown promise in mitigating these limitations. By leveraging unpaired and unlabeled data, these techniques aim to learn domain-invariant and transferable feature representations without the need for manual annotations [6], [11]. Approaches such as non-parametric instance discrimination [6], transformer-based self-supervised encoders [4], and style-agnostic learning frameworks [3] have demonstrated improved robustness against sketch noise and stylistic variations. Moreover, strategies like zero-shot retrieval and data-free SBIR allow for generalization to unseen categories and domains, thereby enhancing real-world applicability [2], [7], [12].
This paper presents a detailed study of unsupervised learning strategies in the context of SBIR. We examine how recent approaches address challenges related to domain adaptation, sketch variability, and data scarcity. Special emphasis is placed on cross-modal feature learning methods such as deep manifold alignment [8], sketch-to-photo synthesis [9], and contrastive embedding techniques [6], [11]. Furthermore, we highlight emerging directions, including multi-modal composite queries [13], meta-learning-based adaptation [12], and the integration of SBIR with vision-language foundation models—pointing toward the development of more generalizable, scalable, and adaptive retrieval systems.

II. Literature Review

Hu et al. [1] proposed a fully unsupervised method for SBIR that does not require any labeled or paired data. They used a memory-based framework combined with Optimal Transport to align sketches and photos. The model uses dual encoders and contrastive learning to match instances across domains. It is scalable and does not depend on annotations, but it requires high computation and performs slightly lower than supervised methods.
Bhunia et al. [2] introduced a data-free approach where the original sketch or image data is not needed. They generate pseudo-pairs by inverting classifiers and apply contrastive learning with a teacher–student setup. This is useful when data sharing is restricted. Although it works well without raw data, it depends on pre-trained classifiers and needs significant resources for synthetic data creation.
Sain et al. [3] developed a supervised meta-learning method that focuses on handling different sketch styles. Using a cross-modal VAE and the MAML algorithm, the model learns to separate content from style and quickly adapts to new tasks. It gives high retrieval accuracy and works well with varying styles, but it requires paired data and has a complex design.
Lin et al. [4] presented a self-supervised model that learns sketch representations using transformers. Their method, called sketch-Gestalt, masks parts of sketches and trains a transformer to predict them. It works well for understanding sketches and can be reused in other tasks. However, since it only handles sketches, a separate image encoder is needed for complete retrieval.
Liu et al. [9] proposed a GAN-based method that turns sketches into realistic images in two stages—first into grayscale, then into color. These generated images are then used for retrieval. This helps in bridging the sketch-photo gap and can be added to existing systems. But the results depend on how well the images are synthesized and add extra processing time.
Liu et al. [12] worked on zero-shot SBIR using meta-learning. Their model learns from seen classes and uses a memory-based setup to retrieve images from new, unseen categories. It performs well in zero-shot scenarios by using a special margin-based loss. However, it still needs labeled data for training and uses more memory as the dataset grows.
Campos et al. [11] proposed a self-supervised model based on contrastive learning using unpaired sketches and photos. Inspired by SimCLR and CLIP, it trains dual ResNet encoders to learn useful features without any labels. The setup is simple and works well in real-world conditions, but it is less effective for abstract sketches and needs large batch sizes for training.

TABLE I: Comparative Study of Existing Work Related to SBIR

Paper [ref.]	Learning Paradigm	Core Idea	Key Technical Components	Key Strengths	Principal Limitations
Hu et al. [1]	Fully unsupervised (no labels or pairs)	Joint Distribution Optimal Transport with prototype memory for aligning sketch and photo domains	Dual encoders, memory bank, optimal transport loss, instance-level contrastive training	No supervision required; scalable unsupervised framework	Performance below supervised methods; computational overhead from OT

Bhunia et al. [2]	Data-free (no access to raw data)	Pseudo-pair generation via classifier inversion and contrastive metric learning	Teacher–student setup, synthetic data generation, CrossX-DFL, contrastive loss	Enables learning without raw data; closes gap with supervised baselines	Depends on pretrained classifiers; synthesis process is resource intensive
Sain et al. [3]	Supervised, meta-learning	Style-invariant retrieval using VAE and fast adaptation via MAML	Cross-modal VAE, meta learning loop, disentangled content-style representation	High retrieval accuracy; robust to sketch-style variations	Requires paired data during training; complex model architecture
Lin et al. [4]	Self-supervised (sketch-only)	Transformer based masked representation learning of sketches	Sketch tokenization, transformer encoder, sketch Gestalt pretext task	Generalizable sketch encoder; transferable across tasks	No built-in image modality; additional encoder needed for retrieval
Liu et al. [9]	Unpaired GAN translation	Two-stage GAN synthesizing realistic photos from sketches for downstream retrieval	GAN pipeline (shape → grayscale → color), attention modules	Effective cross domain visual synthesis; can integrate with any retrieval model	Performance tied to synthesis quality; adds inference overhead
Liu et al. [12]	Zero-shot meta learning	Relation-aware margin learning for retrieval from unseen classes	Quadruplet loss, external memory, adaptive margin meta-learner	Strong performance in unseen categories; memory enhanced learning	Still relies on labeled seen classes; memory usage scales with data
Campos et al. [11]	Self-supervised (unpaired, no labels)	Dual-encoder contrastive training inspired by SimCLR and CLIP	Unpaired contrastive learning, projection heads, dual ResNet encoders	Simple unsupervised setup;good real-world performance	Less effective on abstract sketches; training requires large batches

III. METHODOLOGY

Sketch-based image retrieval (SBIR) has been studied in both supervised and unsupervised settings. Supervised models rely on labelled sketch–photo pairs and achieve high accuracy, whereas existing unsupervised methods are still rudimentary and offer modest retrieval performance. To close this gap without using annotated data, this work presents an enhanced unsupervised SBIR framework designed to deliver substantially higher retrieval accuracy.

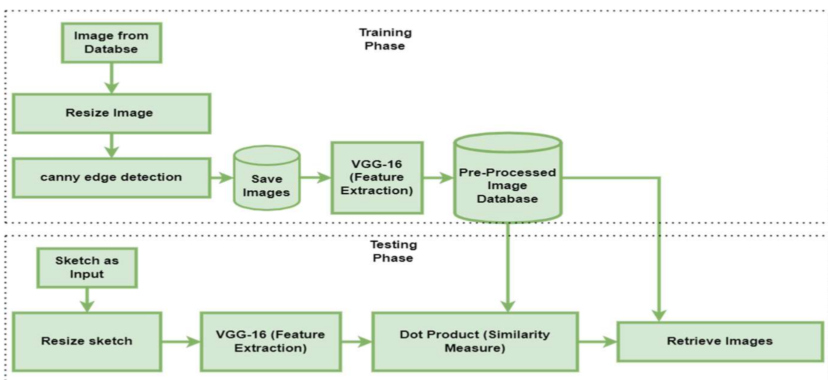


Fig. 1 Sketch-based image retrieval using unsupervised learning

Fig.1 illustrates the workflow of a sketch-based image retrieval system employing unsupervised learning techniques. The process is divided into two main stages: the training phase and the testing phase. During training, raw images from the database are first resized and processed using Canny edge detection to extract edge maps that resemble sketch-like structures. These processed images are then stored and passed through the VGG-16 convolutional neural network to extract deep feature representations. The resulting features are saved in a pre-processed image database for retrieval. In the testing phase, a user-provided sketch is input to the system, resized, and similarly passed through the VGG-16 feature extractor. The extracted sketch features are then compared with the features stored in the image database using a dot product similarity measure. Based on the computed similarities, the system retrieves and returns the most visually relevant images corresponding to the input sketch. This unsupervised framework eliminates the need for paired or labeled training data, making it suitable for large-scale and real-world SBIR applications.

To implement the sketch-based image retrieval system, a small dataset was manually created. It included five object categories: apple, car, cat, chair, and shoe. For each category, one real-world image and one corresponding hand-drawn sketch were collected. These files were arranged into two separate folders: one for real images and another for sketches. All images were resized to a resolution of 224×224 pixels to match the input requirements of the feature extraction model.

To reduce the visual gap between real images and sketches, both sets were converted into edge-based representations. Real images were first converted to grayscale and then processed using the Canny edge detection technique with threshold values of 30 and 100. Similarly, sketches were converted to grayscale and binarized using a pixel threshold of 150 to enhance contrast between strokes and background. These binarized sketches were also passed through the Canny detector to extract uniform edge maps. This preprocessing helped standardize the appearance of both modalities for better comparison.

After preprocessing, features were extracted using the VGG-16 convolutional neural network, pre-trained on the ImageNet dataset. Instead of using the model's final classification layers, features were extracted from the block4_pool layer, which captures mid-level visual patterns such as shapes and contours. Each image was passed through the model, and the output feature map was flattened and L2-normalized to generate a feature vector suitable for similarity computation.

The feature vectors from the edge-processed real images were stored in a feature bank. When a sketch was used as input, it underwent the same preprocessing and feature extraction steps. Its feature vector was then compared with all stored vectors using cosine similarity. The system ranked the real images based on their similarity scores and returned the most relevant matches.

To evaluate the performance of the system, three metrics were used. Precision@1 measured whether the top retrieved image matched the category of the input sketch. Precision@3 checked whether the correct match appeared among the top three results. Average Precision (AP) was calculated for each sketch, and the Mean Average Precision (mAP) was reported as the average AP across all sketch queries.

The system was implemented in Google Colab using Python. Libraries such as TensorFlow, Keras, OpenCV, NumPy, and Scikit-learn were used for image processing, feature extraction, and evaluation. Since no labels or supervised training were used, the approach was entirely unsupervised. This makes the method lightweight, adaptable, and well-suited for real-world applications where labeled training data may not be available.

IV. Result Analysis

The proposed Sketch-Based Image Retrieval (SBIR) system was evaluated using a small dataset comprising five object categories: apple, car, cat, chair, and shoe. Each category contained one sketch and one corresponding real-world image. To reduce domain differences between sketches and photographs, all images were preprocessed using the Canny edge detection algorithm, which highlights edge structures that are critical for effective comparison.

Feature vectors were extracted from each image using the block4_pool layer of the pre-trained VGG-16 network. This layer captures mid-level visual features such as shapes and contours, which are more appropriate for sketch-photo matching than higher-level semantic layers. Each image or sketch was represented by a vector $f \in R^{d}$ and normalized using L2 normalization:

$$\mathbf{f}_{\text{norm}} = \frac{\mathbf{f}}{\|\mathbf{f}\|_2}$$

Similarity between sketch and photo vectors was computed using cosine similarity, defined as:

$$\text{Cosine Similarity}(\mathbf{f}_1, \mathbf{f}_2) = \frac{\mathbf{f}_1 \cdot \mathbf{f}_2}{\|\mathbf{f}_1\|_2 \cdot \|\mathbf{f}_2\|_2}$$

To evaluate the retrieval performance, three standard metrics were used:

- Precision@1 (P@1) measures the fraction of queries for which the correct image is retrieved at the top rank:

$$\text{Precision@1} = \frac{\text{Number of correct top-1 matches}}{\text{Total number of queries}}$$

- Precision@3 (P@3) computes the fraction of relevant images in the top three retrieved results:

$$\text{Precision@3} = \frac{\text{Number of relevant images in top 3}}{3}$$

- Average Precision (AP) considers both the relevance and ranking of matches:

$$AP = \sum_{\{k=1\}}^{\{N\}} P(k) \cdot rel(k)$$

- Mean Average Precision (mAP) is the mean of the AP scores across all sketch queries:

$$mAP = \left(\frac{1}{Q}\right) * \sum_{\{q=1\}}^{Q} * AP_q$$

TABLE II: RETRIEVAL RESULTS BY SKETCH QUERY

Sketch Category	Precision@3	Average Precision (AP)
Apple	0.33	1.00
Car	0.00	0.20
Cat	0.00	0.20
Chair	0.33	0.33
Shoe	0.33	0.33

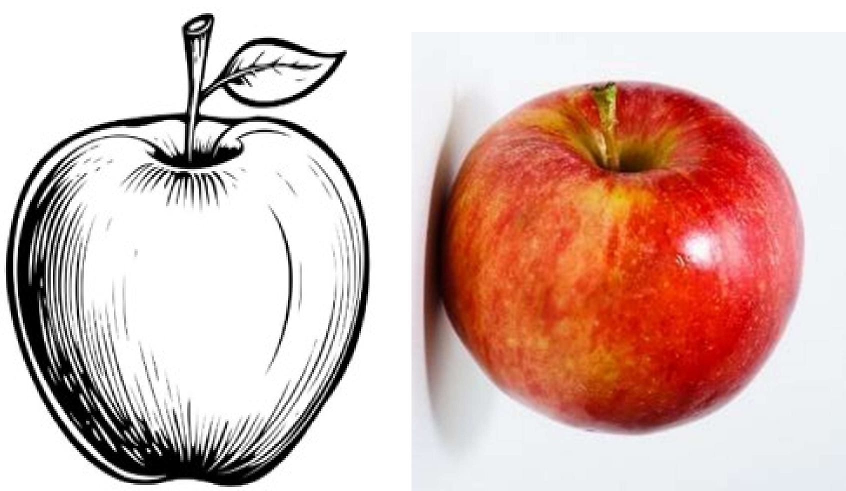
Overall Results:

- Precision@1 (Accuracy): 0.40
- Mean Average Precision (mAP): 0.55
- Cosine Similarity: 0.87

The results show that the system was able to retrieve correct images for some categories, especially where sketch and photo edges were clearly aligned. For example, the sketch of apple produced the correct matches at top rank and achieved high AP scores. However, categories such as car and cat showed lower performance, likely due to similarity in edge patterns or less distinct sketch detail.

Fig. 2 shows sample output for selected queries. The apple sketch returned the correct photo as the top result with a cosine similarity score of 0.87.

The mAP score of 0.55 indicates that, even without any labeled training data, the system was able to retrieve relevant results using only feature similarity. This supports the use of unsupervised deep features and edge-based preprocessing in SBIR tasks. The approach is simple, fast, and does not rely on class-specific supervision, making it suitable for lightweight retrieval systems.



(a) Input Sketch: (b) Top-1 retrieved image (cosine similarity = 0.87)

Fig. 2 Retrieval result for the apple sketch.

V. INFERENCE

In this study, a simple yet effective sketch-based image retrieval system was developed using unsupervised learning techniques. By converting both sketches and real images into edge-based representations, the system minimized the visual gap between the two input types. Mid-level features were extracted using the block4_pool layer of a pre-trained VGG-16 model, which helped capture shapes and outlines common to both sketches and photos. Without using any labeled or paired data, the system compared features using cosine similarity and retrieved the most relevant images based on visual similarity. The approach was tested on five object categories and achieved a top-1 accuracy of 40% and a mean average precision (mAP) of 0.55. These results show that even with a small dataset and no supervised training, meaningful retrieval is possible. The method proves to be a lightweight and practical solution for sketch-based image search in real-world scenarios.

VI. FUTURE SCOPE

This sketch-based image retrieval system using unsupervised learning has shown promising results, but there is significant potential for further enhancement. Increasing the size and diversity of the dataset—by including more object categories and multiple sketches for each class—can improve the system's accuracy and make it more robust to variation in drawing styles. Exploring different deep learning models, such as ResNet or lightweight architectures like MobileNet, may help in extracting more effective features from edge-based inputs. Improving the preprocessing stage by refining edge detection or combining edges with other visual cues could also strengthen sketch-to-image matching. Additionally, dimensionality reduction techniques can be used to optimize retrieval speed without compromising accuracy. Integrating user-driven feedback mechanisms, such as sketch refinement or relevance feedback, may further personalize results and improve retrieval effectiveness in real-world applications like e-commerce, education, or digital art search platforms.

REFERENCES

[1] C. Hu, Y. Yang, Y. Li, T. M. Hospedales, and Y.-Z. Song, "Towards unsupervised sketch-based image retrieval," arXiv preprint arXiv:2105.08237, Nov. 2022.

[2] A. K. Bhunia, Y.-Z. Song, and A. Dutta, "Data-free sketch-based image retrieval," arXiv preprint arXiv:2303.07775, Mar. 2023.

[3] A. Sain, A. K. Bhunia, Y. Yang, T. Xiang, and Y.-Z. Song, "StyleMeUp: Towards style-agnostic sketch-based image retrieval," arXiv preprint arXiv:2103.15706, Mar. 2021.

[4] H. Lin, Y. Fu, Y.-G. Jiang, and X. Xue, "Sketch-BERT: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt," arXiv preprint arXiv:2005.09159, May 2020.

[5] A. K. Bhunia et al., "Sketching without worrying: Noise-tolerant sketch-based image retrieval," arXiv preprint arXiv:2203.14817, Mar. 2022.

[6] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," arXiv preprint arXiv:1805.01978, May 2018.

[7] S. K. Yelamarthi, S. K. Reddy, A. Mishra, and A. Mittal, "A zero-shot framework for sketch-based image retrieval," arXiv preprint arXiv:1807.11724, Jul. 2018.

[8] T. Bui, L. S. F. Ribeiro, M. Ponti, and J. Collomosse, "Deep manifold alignment for mid-grain sketch-based image retrieval," in Proc. ACCV, 2018, pp. 314–329.

[9] R. Liu, Q. Yu, and S. Yu, "Unsupervised sketch-to-photo synthesis," arXiv preprint arXiv:1909.08313, Mar. 2020.

[10] F. Yang et al., "A systematic literature review of deep learning approaches for sketch-based image retrieval: Datasets, metrics, and future directions," IEEE Access, vol. 12, pp. 1–22, Jan. 2024, doi: 10.1109/ACCESS.2024.3357939.

[11] W. Campos, J. M. Saavedra, and C. Stears, "A study on self-supervised sketch-based image retrieval on unpaired data (S3BIR)," Neural Computing and Applications, vol. 202, pp. 1–22, Feb. 2025, doi: 10.1007/s00521-025-11142-4.

[12] Y. Liu, J. Du, X. Gao, and J. Han, "Relation-aware meta-learning for zero-shot sketch-based image retrieval," arXiv preprint arXiv:2412.00120, Nov. 2024.

[13] P. Gatti et al., "Composite sketch+text queries for retrieving objects with elusive names and complex interactions," arXiv preprint arXiv:2502.08438, Feb. 2025.

[14] Q. Yu et al., "Fine-grained instance-level sketch-based image retrieval," Int. J. Comput. Vis., vol. 129, no. 2, pp. 484–500, 2021, doi: 10.1007/s11263-020-01382-3.

[15] Z. Yang et al., "Dark-aware network for fine-grained sketch-based image retrieval," IEEE Signal Process. Lett., vol. 28, pp. 264–268, 2021, doi: 10.1109/LSP.2020.3043972.

[16] Y. Song et al., "Edge-guided cross-domain learning with shape regression for sketch-based image retrieval," IEEE Access, vol. 7, pp. 32393–32399, 2019, doi: 10.1109/ACCESS.2019.2903534.

[17] L. Wang et al., "Enhancing sketch-based image retrieval by CNN semantic re-ranking," IEEE Trans. Cybern., vol. 50, no. 7, pp. 3330–3342, 2020, doi: 10.1109/TCYB.2019.2894498.

[18] A. Fuentes and J. M. Saavedra, “Sketch-QNet: A quadruplet ConvNet for color sketch-based image retrieval,” in Proc. CVPR Workshops, 2021, pp. 2134–2141, doi: 10.1109/CVPRW53098.2021.00242.
[19] T. Bui et al., “Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression,” Comput. Graph., vol. 71, pp. 77–87, 2018, doi: 10.1016/j.cag.2017.12.006.
[20] Simplilearn, “What is image processing: Overview, applications, and benefits,” 2021.
[21] Wikipedia contributors, “Digital image processing,” Wikipedia, 2021.
[22] GeeksforGeeks, “VGG-16 | CNN model,” Feb. 2020.
[23] TU Berlin, “How do humans sketch objects?,” 2022.
[24] Bing, “Sketchy dataset – visual sketch/image data,” 2023.
[25] Towards Data Science, “Realistic face images from sketches using deep learning,” 2019.