Journal of Information Hiding and Multimedia Signal Processing ©202 ISSN 2073-4212
Ubiquitous International Volume **16**, Number **3**, September **2025**

A novel approach using the local sketch and its variations for image retrieval in education

Thien Le Quang[1,2]

[1]Faculty of Management Information Systems, Ho Chi Minh University of Banking, Ho Chi Minh City, Vietnam
[2]Faculty of Information Technology, Thuyloi University, Hanoi, Vietnam
lequangthien22@gmail.com

Dat Tran[2]

[2]Faculty of Information Technology, Thuyloi University, Hanoi, Vietnam

Quynh Nguyen Huu[3]

[3]Faculty of Information Technology, CMC University, Hanoi, Vietnam
nhquynh@cmc-u.edu.vn

Received April 17, 2025, revised July 5, 2025, accepted July 6, 2025.

ABSTRACT. *In the field of education, retrieving student information from classrooms is extremely important. While textual data such as dates and class schedules are commonly used for information retrieval tasks, the use of camera footage in classrooms is also widespread. However, image retrieval, particularly using sketches, is a new and complex technology. In this paper, we develop a sketch-based image retrieval system to extract information from classroom cameras. The final results allow for precise retrieval previously unattainable, enabling users to make increasingly detailed queries and incorporate attributes such as color and contextual hints from the sketches. To achieve this, we introduce a new framework that effectively integrates sketch images using pre-trained CLIP models, eliminating the need for detailed sketch descriptions. Lastly, our system extends to include sketch-based image retrieval applications, domain attribute transformation, and detailed image generation, offering solutions for various real-world scenarios.*
Keywords: local sketch retrieval, image retrieval, deep learning

1. **Introduction.** Information retrieval in classrooms within the education sector is quite common for some classes equipped with cameras. However, retrieval using sketch images is still novel as it heavily depends on resources and technological platforms [1, 2]. Nonetheless, transforming from sketch images to specific images has become a focal task. The more detailed the sketch captures information, the better the retrieval accuracy, and vice versa. Research in the field of sketch-based image retrieval is quite abstract and faces many challenges.

In this paper, we pose research questions on how a sketch image can be used to retrieve related images in the field of education, particularly concerning classroom issues. The output of this deep learning model will be images that are related or closely related to the sketch. From there, certain assessments about the quality of education, teaching, and student performance can be made. This problem serves as a contextual suggestion for a specific sketch image, such as finding a student using a phone from a sketch of a person using a phone.

Although sketch-based image retrieval has been studied before, it mainly focuses on a single scene, object, or category available in the dataset, rather than combining multiple scenes and objects. Combining multiple scenes and objects for image retrieval at different levels and with increasing detail will enhance retrieval accuracy. Indeed, for sketch-based image retrieval, the

higher the detail in the sketch, the more accurate the retrieval. Therefore, building a model that can retrieve images based on sketches to capture multiple scenes and objects is quite challenging.

The main challenge we aim to address is solving the problem of combining detailed features, specifically researching how sketches can supplement features to build detailed characteristics. Our goal is to maintain the semantics of scenes and objects. To address this challenge, we leverage popular backbones to create composite semantics to reconstruct image features, supporting the recognition of similarity between sketch and real images through scene and object features.

Our contributions are three folds and are summarized as follows:

- **Novel methodology**: We introduce a deep learning model capable of searching based on sketch images to support the retrieval of similar images in classrooms, named SIRE. To address these challenges, our work introduces a novel methodology with several key contributions to the SBIR field: (1) Local Sketch Retrieval: Our model is specifically designed to interpret basic, low-detail sketches, reducing the need for elaborate user input; (2) Complex Scene Understanding: Unlike methods focusing on single objects, our approach handles retrieval in complex scenes with multiple, potentially overlapping objects, tailored for the classroom environment; and (3) Hybrid Loss Function: We introduce a combined loss function ($L_{total} = L_{triplet} + L_{rec}$) that simultaneously learns shape similarity and reconstructs fine-grained visual details, leading to more robust retrieval performance.
- **New dataset**: We also introduce the dataset we collected in classrooms at Thuyloi University, named MLIC-Edu.
- **Analysis and evaluation**: We evaluate the proposed model against recent similar models. Additionally, we assess it on popular datasets to evaluate the proposed model.

The remainder of this paper is structured as follows. Section 2 discusses relevant previous studies. Section 3 presents our method. The experimental evaluation is shown in Section 4. Finally, some concluding remarks and a brief discussion are provided in Section 5.

2. Related works.

In this section, we will survey some notable works on foundational image retrieval techniques, sketch-based image retrieval techniques, followed by a brief introduction to modern works in the field of image retrieval.

2.1. Image retrieval techniques.

Image retrieval methods are quite popular in determining the location of objects for similarity matching [3, 4], and are applied in problems such as object detection [5, 6], object recognition [7, 8], and object segmentation [9, 10]. Image retrieval can be performed through single-frame [11] or multi-frame methods [12] to locate objects over time and space. In practice, image retrieval can be achieved through global features [13] or local features [14] based on machine learning models, such as image retrieval models from dictionary learning [15] or image retrieval methods from feature synthesis [16]. Alsmadi [17] introduces an image representation method to partition into clusters that support image retrieval. Yang [18] enhances image retrieval accuracy by ranking and minimizing global average precision. Recently, Guan et al. [19] utilized a CNN backbone to aggregate positional features to speed up image retrieval. Xinfeng [20] has emerged with a method combining Transformer to aggregate features and build a dictionary to support image retrieval tasks.

2.2. Sketch-based image retrieval techniques.

Image retrieval based on sketch images originates from identifying various levels of image retrieval [21, 22], progressing to scenes [23, 24] and objects within the images [25, 26]. Previous deep learning methods [27, 28] often train based on the correlation distances in the embedding space of features [27] related to sketch images. Subsequently, Sangkloy [29] retrieves images from a similar image database with the query being a sketch. Bhunia [25] has developed a deep ranking framework to perform early retrieval from sketches, generalizing the catalog features of sketches, and extracting specific

features related to scenes and objects from the sketches. Extended research based on sketch images [10] has demonstrated effectiveness and enhanced the ability to query images. Most studies on image retrieval from sketches involve constructing mappings from regular images to sketches, which presents a significant challenge in building datasets.

2.3. **Discussion.** Most research on image retrieval focuses primarily on retrieving images through a detailed image or a sketch that contains quite detailed information. However, in reality, images with detailed information are relatively rare. Therefore, in this study, we aim to develop a model that can understand scenes and objects to facilitate image retrieval from sketches in the field of education. We focus on analyzing classroom objects such as whether students are using phones, whether they are paying attention, whether they are leaving their seats, etc. Addressing these challenges is one of the difficult tasks, and our proposed model makes an effort to retrieve images from basic sketches (local sketches).

3. Material and methods.

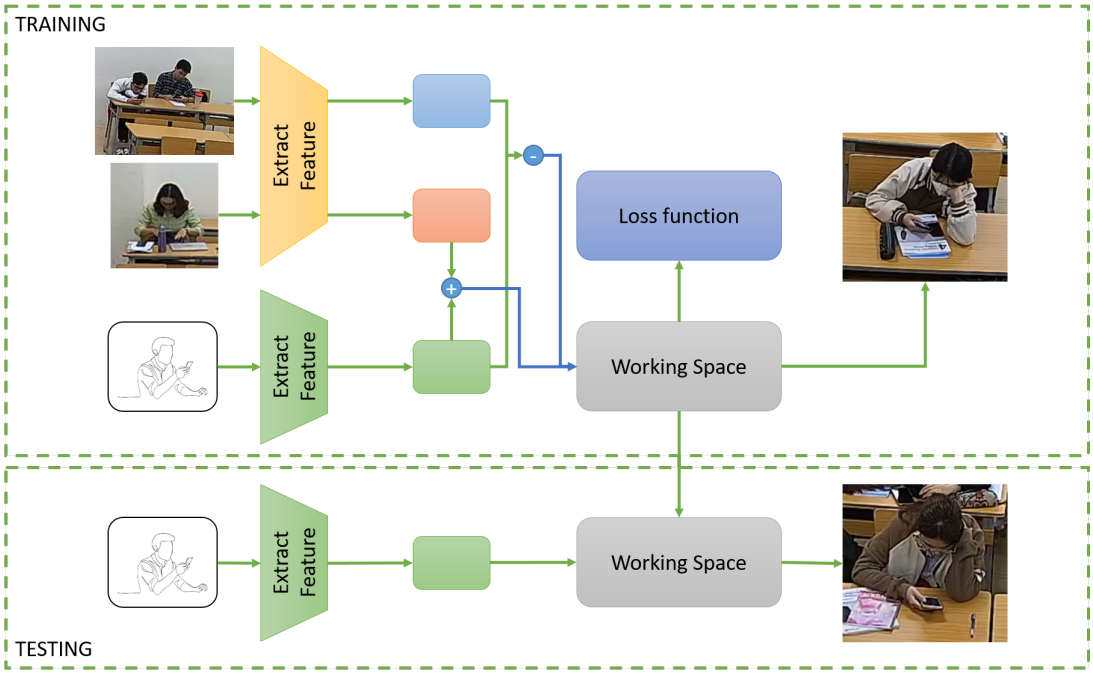

FIGURE 1. Our SIRE framework consists of two main phases: Training and Testing

3.1. **SIRE framework.** Combining structural understanding cues from sketch images and data images creates an effective query for image retrieval. Several studies [30, 31] use features extracted from sketches to generate distinct feature encoders and apply methods like feature mixing or summation to form unique query sets. The main objectives here are: (i) to describe the sketch image accurately; (ii) to map the sketch to real images to aid image retrieval. However, mapping sketches to real images is a challenging task requiring advanced image mapping techniques [32]. We leverage the concept of CLIP for this image mapping task to utilize the image space for encoding sketches, thus addressing the issue of image retrieval. Specifically, we bring sketches into a feature space closer to the real-world images being searched, facilitating image retrieval and reducing search time.

For each feature pair (s, i), each layer is visually described with the semantic image of self-attention and cross-attention units, along with position encoders. Each classifier has a confidence level, c, to determine when to cease the image retrieval inference process. If there are few

points with high confidence, the inference process continues until mismatched points are eliminated during training, allowing the model to predict confidently. Once confidence is achieved, the feature pairs (s, i) between sketches and real images are mapped into a near-similar space for each pair, ensuring the two images almost align within the same feature space. Our proposed architecture is illustrated in Figure 1.

3.2. **Extract Feature.** Traditional image retrieval methods typically use an image to query within a set of existing images [6]. In contrast, we convert a sketch into query tokens to learn from two sets of images: one set with high correlation to the sketch and another with low correlation to the sketch. This approach allows us to build two image sets that support the efficient retrieval of sketch images from a larger dataset.

Specifically, for an image P, we create its latent representation feature p as $p = V(P)$. However, for a sketch, we represent its equivalent feature as sp, where sp denotes the equivalent feature of the query sketch. We construct a feature extraction network from computer vision transformations using a three-layer MLP [33] with ReLU activation functions [34]. During training, we continuously extract features from three inputs: the sketch, the set of images highly correlated with the sketch as the positive feature (p^+), and the set of images with low correlation to the sketch as the negative feature (p^-). After processing these three inputs, we integrate them using a triplet loss function [35] to build continuous suggestion vectors for the image retrieval process. The triplet loss function aims to minimize the distance between a randomly chosen positive feature (p^+) and the sketch feature. We leverage the compositional feature properties of zero-shot learning [36] to aggregate feature encoders, so even when the sketch provides minimal information, the training process still has sufficient data to support learning.

3.3. **Working Space.** The model for feature extraction and integration into the workspace consists of four main modules, show in Figure 2. Module 1 calculates features that are designed to preserve structure, preventing the generation of non-informative features. Module 2, a self-attention module [37], identifies the most informative regions to facilitate local information matching. Module 3 is a cross-attention model that learns correlated features. Finally, Module 4 enables the linking of feature pairs to support the measurement and retrieval of matching feature pairs within the workspace.

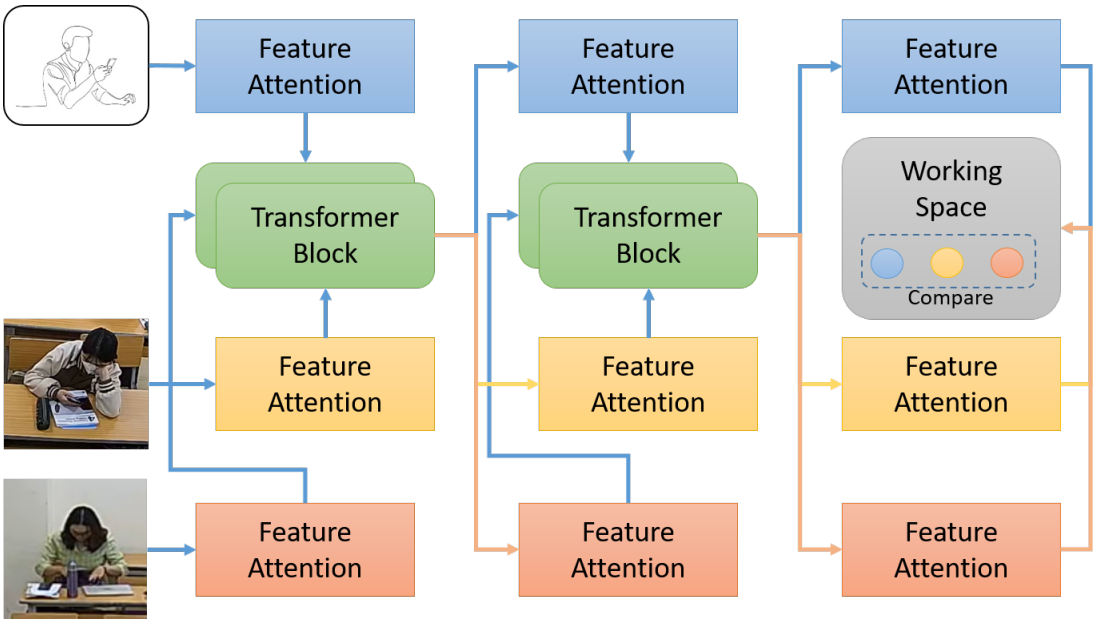

FIGURE 2. Network overview

Our method primarily relies on describing a sketch and pairing it with two support feature sets during the training process. During model usage, the user provides a sketch image to infer the desired visual information. The sketch is combined with two support feature sets to accurately retrieve images in the educational domain. For this purpose, we construct a workspace that includes embedded feature vectors to represent the discrepancy between the sketch and the target image, expressed as $DeltaF = |p^+ - e|$, and concatenate it with DeltaW. DeltaW is a token that stores the difference between the sketch and the actual image.

The support feature set [38] functions similarly to constructing fixed handcrafted features to enhance generalization on unseen images, which can improve performance on the seen training sets. Although these support feature sets will be continuously trained to replace handcrafted features, to generalize image retrieval in education based on sketches, we create a common workspace to embed both the features and these support sets. This also allows generalization beyond the training capability on seen sets. Specifically, at each training step, we randomly select a support feature set D and set the embedding step PD = Embedding(di) to generate a fixed representation. Then, we represent the sketch with the embedding representation.

3.4. **Loss function.** To further improve the connection between modules, we consider the CLIP visual encoder [40] to transform the input image I into features $pr = V(I)$ for each patch. To build region-based connection sets, we create high-correlation computation functions A between the global sketch features sTL, defined as $A = (pr \cdot sTL)$, where A is normalized with the softmax function across the patches. Each value a_i represents the combination of the global sketch retrieval and the image features to be retrieved for each patch. We then sum the weighted embeddings across all patches to obtain a regional image feature for retrieval: $ps = \sum(a_i \times p_i)$, where i ranges from 1 to T. We construct correlated feature embeddings (p^+ and p^-) to support the triplet loss function Ltriplet with a margin $\phi > 0$ as follows:

$$Ltriplet = max(0,\ \phi + \theta(sTL,\ p_s^+) - \theta(sTL,\ p_s^-)) \tag{1}$$

For sketch images, the triplet loss function typically focuses on matching detailed shapes [35], while neglecting finer details such as color or texture. To create a workspace for compositional image retrieval, we develop structural features within the visual domain. We design an objective function to reconstruct the sketch into a real-life image, requiring a Unet decoder [41] to achieve realistic image reconstruction using the pixel-level L2 reconstruction loss function. The reconstruction loss Lrec is thus formulated for real-life image reconstruction as follows:

$$Lrec = \left\| P^+ - G(sTL) \right\|_2 \tag{2}$$

Finally, the overall loss function becomes:

$$Ltotal = Ltriplet + Lrec \tag{3}$$

4. **Experiments.** In this section, we provide a detailed description of the implementation of the proposed model for image retrieval in education and the evaluation metrics used on three datasets: QMULShoeV2 [39], SketchyCOCO [32] and our dataset, MLIC-Edu. We also compare our model with state-of-the-art methods through both quantitative and qualitative summary studies to analyze the superiority of the proposed model. Additionally, we discuss the experiments conducted to assess the effectiveness of each module within the proposed model.

4.1. **Dataset.** We utilized three datasets related to sketch images: QMULShoeV2 [39], SketchyCOCO [32], and MLIC-Edu (our dataset). The QMULShoeV2 dataset contains 2,000 sketches and 6,730 real-world images. The SketchyCOCO dataset consists of 27,683 images, of which 18,869 are real-world images and 5,512 are sketches. The MLIC-Edu dataset includes 10,235

images, consisting of 400 hand-drawn sketches, 2,367 sketches generated using the SketchyGAN model [32], and 7,468 real-world images collected by classroom cameras at Thuyloi University. Additionally, the MLIC-Edu dataset includes 10 education-related labels: dozing off, using a phone, turning sideways, turning vertically, fighting, hugging, raising a hand, opening a book, reading, and taking notes (see Figure 3); each label is assigned nearly equivalent amounts of data. Real-world image data were collected over a period of 6 months from five different classrooms at Thuyloi University. All recording activities were approved by the university administration and consented to by participating students. Cameras were positioned at various angles to capture diverse perspectives and lighting conditions. The original resolution of the captured images was 1920×1080 pixels prior to processing.

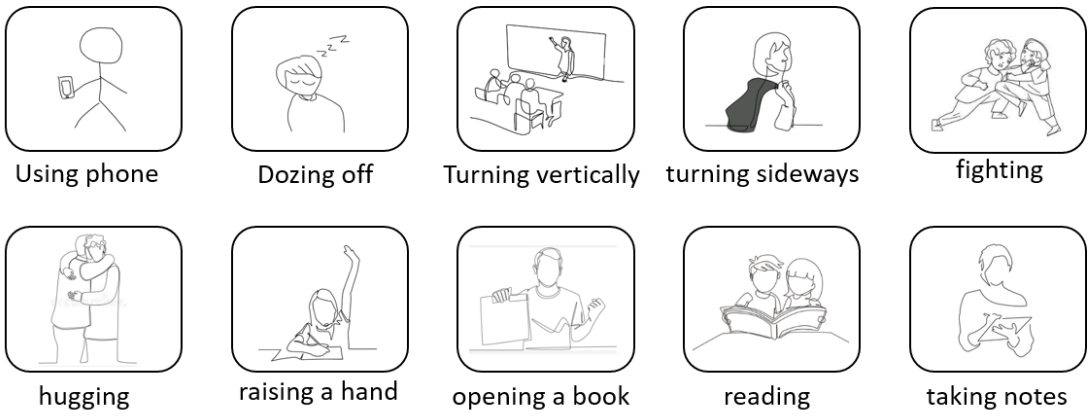

FIGURE 3. Images of sketches for each label in the MLIC-Edu dataset

TABLE 1. Comparison between MLIC-Edu and SketchyCOCO datasets

Feature	**MLIC-Edu (Ours)**	**SketchyCOCO**
Scope	Domain-specific: Education, classroom monitoring	General-purpose: Common objects and scenes
Application	Fine-grained analysis of student behavior	General scene-level image generation and retrieval
Label Semantics	Describes specific student actions and states (e.g., 'raising a hand', 'dozing off')	Describes common objects (e.g., 'person', 'car', 'dog') and their arrangement
Image Context	Consistent: Real-world images from indoor classroom cameras	Diverse: Images from various indoor and outdoor settings

As shown in Table 1, while SketchyCOCO is valuable for general scene understanding, MLIC-Edu provides a focused, context-rich resource essential for developing and evaluating SBIR systems tailored to the specific challenges of the educational domain.

4.2. **Performance metrics and training setup.** To extract features from images, we used the CLIP ViT-L/14 image encoder as the backbone pretrain [40] in all our experiments. The models were trained with a learning rate of 0.00001. The Unet decoder [41] and the image converter from sketch to real-life images were trained with learning rates of 0.0001 and 0.001, respectively. We trained the model for 120 epochs using the AdamW optimizer [42] with a batch size of 256. For data splitting, we allocated 90% of the data for the training set and 10% for the test set. We used the [25] accuracy metric to evaluate the percentage of sketches matching real-life images within the top 10 retrieved images.

TABLE 2. Results for fine-grained object-level composed retrieval.

Methods	QMULShoe-V2	SketchyCOCO	MLIC-Edu
TASK-former [29]	0.441	0.347	0.425
SceneTrilogy [45]	0.462	0.402	0.4467
SketchyS [43]	0.750	0.786	0.7143
Triplet-SN [44]	0.7156	0.735	0.6834
SIRE (ours)	**0.7646**	**0.8053**	**0.7455**

4.3. **Experiment setup.** We design our extensive empirical study to answer the following three key research questions (RQs):

- RQ1: How is the SIRE model better compared to other deep learning methods with the same concept?
- RQ2: How does each situation in SIRE contribute to accurate deep learning?
- RQ3: How close is the prediction of the SIRE model to the ground truth?

In RQ1, we showcase the experiments conducted on the three foundational network baselines. For RQ2, we carried out a total of three distinct scenarios. Furthermore, for RQ3, we used the SIRE model to make predictions and provided some of the model's prediction results. The results will be averaged over experimental runs on three datasets.

4.4. **Results and discussion.**

4.4.1. *Comparison With Four Baselines (RQ1).* Table 2 presents the quantitative results compared to methods related to the sketch-based image retrieval problem. In the image retrieval setup (Table 2), our method significantly outperformed both the baselines and state-of-the-art methods across all datasets, demonstrating the effectiveness of the proposed approach in sketch-based image retrieval. This achievement may be attributed to the support of feature sets and the backbone functions that help identify regions in the sketches. Recent competitors like SketchyS [43] and Triplet-SN [44] attempted to retrieve sketch images by combining inverse networks but did not achieve better results. The challenge with sketches lies in representing detailed information within the sketches and reverse retrieval in real-life images. However, thanks to the enhanced learning capabilities through advanced features and backbones, our model surpassed other methods with an accuracy of 74.55% on the MLIC-Edu educational dataset that we collected.

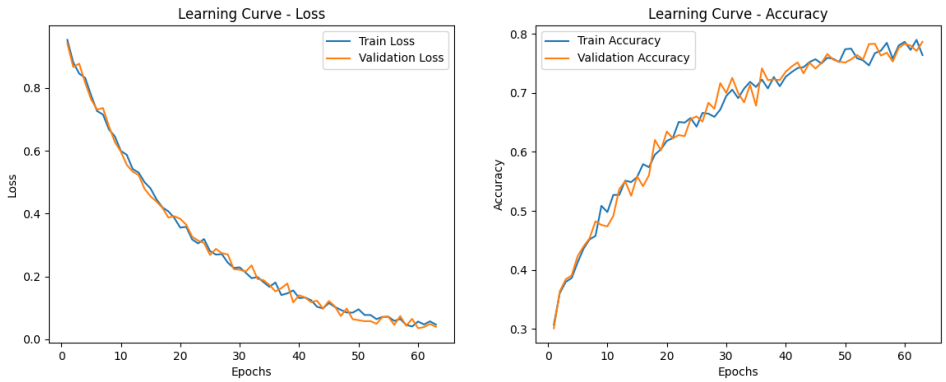

FIGURE 4. Training and loss progress.

Due to early stopping techniques, the training process was halted after 63 epochs. Figure 4 illustrates the model's training progress over time in terms of accuracy and loss for the proposed

model. The narrow gap between the curves indicates that overfitting did not occur. Both training and validation accuracy increased, while training and validation loss decreased as the number of training iterations increased.

4.4.2. *Applicability to Scenarios (RQ2).* To evaluate the proposed method and the effectiveness of combining features and backbones, we conducted five experiments, as presented in Table 3. First, we separately evaluated the positive features and negative features in block 01. Second, we assessed the standard backbone encoder against the CLIP ViT-L/14 image encoder in block 02. Finally, we evaluated the standard backbone decoder against the Unet decoder in block 03. Additionally, we also evaluated the AdamW optimizer in this block 03.

TABLE 3. Five experiments with different inputs and networks

Experiments	**Block 01**	**Block 02**	**Block 03**
Scenario 01	Positive feature	-	-
Scenario 02	Positive feature + Negative feature	-	-
Scenario 03	Positive feature + Negative feature	CLIP ViT-L/14 image encoder	-
Scenario 04	Positive feature + Negative feature	CLIP ViT-L/14 image encoder	Unet decoder
Scenario 05 (SIRE)	Positive feature + Negative feature	CLIP ViT-L/14 image encoder	Unet decoder + AdamW

Table 4 also illustrates the effectiveness of combining feature sets with backbones, which can be roughly understood as multimodal data. In general, the integration of different modalities leads to improvements over using individual modalities. The accuracy increased from the first experiment (only positive feature set) to the third experiment (positive feature set + negative feature set + CLIP ViT-L/14 image encoder), with an average increase of 3% to 6%. Similarly, in the fourth and fifth experiments, accuracy was significantly enhanced with the addition of Unet decoder support and the AdamW convergence function. Furthermore, we observed that selecting the right, sufficiently robust backbone helps the model converge more quickly and achieve higher results.

4.4.3. *Qualitative study (RQ3).* The qualitative results are presented in Figure 5. In the setup for sketch image retrieval in the educational domain, our method significantly outperforms other approaches and state-of-the-art methods on sketch image datasets (as demonstrated earlier), highlighting the effectiveness of combining image retrieval with support feature sets and advanced model backbones. This achievement can be attributed to the fine-tuning of parameters and loss functions to recognize regions and our generative support feature sets. The biggest challenge in sketch image retrieval within the educational field is identifying tiny objects, as a

TABLE 4. Experiment results

Methods	**QMULShoe-V2**	**SketchyCOCO**	**MLIC-Edu**
Scenario 01	0.6443	0.6363	0.6032
Scenario 02	0.7074	0.6897	0.6453
Scenario 03	0.7306	0.7386	0.6753
Scenario 04	0.7554	0.7955	0.7338
Scenario 05 (SIRE)	**0.7646**	**0.8053**	**0.7455**

FIGURE 5. Top-10 fine-grained retrieval result comparison on MLIC-Edu. GT photos are green-bordered. (Zoom-in for best-view)

classroom typically contains many objects, some of which may be obscured by others. However, thanks to the enhanced interaction capabilities provided by the support feature sets to aid the inference and retrieval process, our method achieves relatively stable results that assist in analyzing and evaluating student focus in the classroom.

5. **Conclusion.** Exploring the detailed representation capabilities of sketch images combined with their support feature sets marks a significant advancement in the field of image retrieval. By harmoniously integrating sketch images with real-life feature sets, we introduce a modern and innovative approach to sketch image retrieval within the educational domain. The introduction of an improved image retrieval model drives the training of large language models and supports the annotation of image information. Equally important, we provide an educational dataset for diverse fields such as detailed image generation based on sketches and educational domain-based image retrieval. In the near future, we plan to integrate additional data sources such as text and audio to enhance the efficiency of image retrieval.

REFERENCES

[1] Thanh Nguyen Van, Quynh Nguyen Huu*. Using graph convolutional networks to improve accuracy of image retrieval with relevance feedback method Journal of Information Hiding and Multimedia Signal Processing, Vol. 16, No. 2, pp. 678-689 June 2025.

[2] Mohammad Anwar Assaad*, Maral Ismael Saleh, Rania Mahrousseh. A Novel Framework for Accurate Brain Tumor Detection in MRI Scans Using CNN, MLP, and KNN Techniques Journal of Information Hiding and Multimedia Signal Processing, Vol. 16, No. 2, pp. 590-602, June 2025.

[3] LEE, Seongwon, et al. Correlation verification for image retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. p. 5374-5384.

[4] DIAO, Haiwen, et al. Similarity reasoning and filtration for image-text matching. In: Proceedings of the AAAI conference on artificial intelligence. 2021. p. 1218-1226.

[5] Raja, Rohit, Sandeep Kumar, and Md Rashid Mahmood. "Color object detection based image retrieval using ROI segmentation with multi-feature method." Wireless Personal Communications 112.1 (2020): 169-192.

[6] Dubey, Shiv Ram. "A decade survey of content based image retrieval using deep learning." IEEE Transactions on Circuits and Systems for Video Technology 32.5 (2021): 2687-2704.
[7] Chu, Kai, and Guang-Hai Liu. "Image Retrieval Based on a Multi-Integration Features Model." Mathematical problems in engineering 2020.1 (2020): 1461459.
[8] Bansal, Monika, Munish Kumar, and Manish Kumar. "2D object recognition: a comparative analysis of SIFT, SURF and ORB feature descriptors." Multimedia Tools and Applications 80.12 (2021): 18839-18857.
[9] Raja, Rohit, Sandeep Kumar, and Md Rashid Mahmood. "Color object detection based image retrieval using ROI segmentation with multi-feature method." Wireless Personal Communications 112.1 (2020): 169-192.
[10] Liu, Fang, et al. "Scenesketcher: Fine-grained image retrieval with scene sketches." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16. Springer International Publishing, 2020.
[11] Portillo-Quintero, Jesús Andrés, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. "A straightforward framework for video retrieval using clip." Mexican Conference on Pattern Recognition. Cham: Springer International Publishing, 2021.
[12] Jiang, Chen, et al. "Learning segment similarity and alignment in large-scale content based video retrieval." Proceedings of the 29th ACM International Conference on Multimedia. 2021.
[13] Yan, Chenggang, et al. "Deep multi-view enhancement hashing for image retrieval." IEEE Transactions on Pattern Analysis and Machine Intelligence 43.4 (2020): 1445-1451.
[14] Cao, Bingyi, Andre Araujo, and Jack Sim. "Unifying deep local and global features for image search." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. Springer International Publishing, 2020.
[15] Öztürk, Şaban. "Convolutional neural network based dictionary learning to create hash codes for content-based image retrieval." Procedia Computer Science 183 (2021): 624-629.
[16] Liu, Xihui, et al. "More control for free! image synthesis with semantic diffusion guidance." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[17] Alsmadi, Mutasem K. "Content-based image retrieval using color, shape and texture descriptors and features." Arabian Journal for Science and Engineering 45.4 (2020): 3317-3330.
[18] Yang, Min, et al. "Dolg: Single-stage image retrieval with deep orthogonal fusion of local and global features." Proceedings of the IEEE/CVF International conference on Computer Vision. 2021.
[19] Guan, Anna, et al. "Precision medical image hash retrieval by interpretability and feature fusion." Computer Methods and Programs in Biomedicine 222 (2022): 106945.
[20] Dong, Xinfeng, et al. "Hierarchical feature aggregation based on transformer for image-text matching." IEEE Transactions on Circuits and Systems for Video Technology 32.9 (2022): 6437-6447.
[21] Lee, Junsoo, et al. "Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[22] Ribeiro, Leo Sampaio Ferraz, et al. "Sketchformer: Transformer-based representation for sketched structure." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[23] Qi, Anran, et al. "Toward fine-grained sketch-based 3D shape retrieval." IEEE Transactions on Image Processing 30 (2021): 8595-8606.
[24] Xu, Fang, et al. "Mental retrieval of remote sensing images via adversarial sketch-image feature learning." IEEE Transactions on Geoscience and Remote Sensing 58.11 (2020): 7801-7814.
[25] Bhunia, Ayan Kumar, et al. "Sketch less for more: On-the-fly fine-grained sketch-based image retrieval." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[26] Hosseinzadeh, Mehrdad, and Yang Wang. "Composed query image retrieval using locally bounded features." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[27] Osahor, Uche, et al. "Quality guided sketch-to-photo image synthesis." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020.
[28] Kampelmuhler, Moritz, and Axel Pinz. "Synthesizing human-like sketches from natural images using a conditional convolutional decoder." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2020.
[29] Sangkloy, Patsorn, et al. "A sketch is worth a thousand words: Image retrieval with text and sketch." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.
[30] Liu, Bingchen, et al. "Self-supervised sketch-to-image synthesis." Proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 3. 2021.
[31] Chowdhury, Pinaki Nath, et al. "Fs-coco: Towards understanding of freehand sketches of common objects in context." European conference on computer vision. Cham: Springer Nature Switzerland, 2022.

[32] Gao, Chengying, et al. "Sketchycoco: Image generation from freehand scene sketches." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[33] Zhou, Zixia, Md Tauhidul Islam, and Lei Xing. "Multibranch CNN with MLP-mixer-based feature exploration for high-performance disease diagnosis." IEEE Transactions on Neural Networks and Learning Systems (2023).
[34] Xu, Jin, et al. "Reluplex made more practical: Leaky ReLU." 2020 IEEE Symposium on Computers and communications (ISCC). IEEE, 2020.
[35] Dehghan, Alireza, et al. "TripletMultiDTI: multimodal representation learning in drug-target interaction prediction with triplet loss function." Expert Systems with Applications 232 (2023): 120754.
[36] Wang, Qingsheng, et al. "Learning conditional attributes for compositional zero-shot learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[37] Pan, Xuran, et al. "On the integration of self-attention and convolution." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[38] Naeem, Hamad, et al. "Development of a deep stacked ensemble with process based volatile memory forensics for platform independent malware detection and classification." Expert Systems with Applications 223 (2023): 119952.
[39] Zhang, Yafei, et al. "Cross-compatible embedding and semantic consistent feature construction for sketch re-identification." Proceedings of the 30th ACM International Conference on Multimedia. 2022.
[40] Shan, Xiangheng, et al. "Open-Vocabulary Semantic Segmentation with Image Embedding Balancing." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[41] Bougourzi, Fares, et al. "PDAtt-Unet: Pyramid dual-decoder attention Unet for Covid-19 infection segmentation from CT-scans." Medical Image Analysis 86 (2023): 102797.
[42] Yao, Zhewei, et al. "Adahessian: An adaptive second order optimizer for machine learning." proceedings of the AAAI conference on artificial intelligence. Vol. 35. No. 12. 2021.
[43] Chowdhury, Pinaki Nath, et al. "Partially does it: Towards scene-level fg-sbir with partial input." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[44] Koley, Subhadeep, et al. "How to Handle Sketch-Abstraction in Sketch-Based Image Retrieval?." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[45] Chowdhury, Pinaki Nath, et al. "SceneTrilogy: On Human Scene-Sketch and its Complementarity with Photo and Text." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.