CLIP-based dual-encoder approach for Sketch-to-Image Retrieval

Ky Quoc Binh Le[1,2] and Thi-Ngoc-Hanh Le[1,2⋆]

[1] International University, VNU-HCM, Ho Chi Minh City, Vietnam
[2] Vietnam National University Ho Chi Minh City, Viet Nam
lkqb2708@gmail.com, ltnhanh@hcmiu.edu.vn

Abstract. Sketch-Based Image Retrieval (SBIR) enables users to search visual databases using freehand sketches, yet effective zero-shot retrieval remains challenging due to the large modality gap between sparse line drawings and natural images. This paper introduces an efficient dual-encoder SBIR framework built upon CLIP's ViT-B/32 backbone and a hybrid loss that combines InfoNCE contrastive alignment with hard-batch triplet mining. The resulting embedding space jointly captures global semantic structure and fine-grained distinctions, enabling robust retrieval even when sketches are abstract, incomplete, or stylistically diverse. Evaluated on the Sketchy benchmark, our approach converges rapidly-achieving $\geq$ 99% Recall100 in only 50 epochs, a 3× speed-up over Doodle2Search while also improving retrieval accuracy (mAP 81.4 vs. 74.2). Automatic mixed precision and selective fine-tuning further reduce computational cost, enabling larger batch sizes and more stable optimization on a single consumer GPU. Qualitative results show consistent retrieval across organic, geometric, and fine-structure categories, with improved resilience to noise and occlusion. Overall, our CLIP-guided dual-encoder framework delivers fast, scalable, and semantically aligned zero-shot SBIR, setting a strong foundation for future extensions to hierarchical semantics, generative augmentation, and large-scale indexing.

Keywords: SBIR, sketch, retrieval, sketch-to-image, contrastive

1 Introduction

Sketch-Based Image Retrieval (SBIR) has emerged as a powerful alternative to traditional text-based image search, enabling users to express visual intent directly through freehand sketches rather than natural-language descriptions. As digital content continues to proliferate across domains such as e-commerce, visual design, creative media, and medical imaging, the ability to retrieve images using intuitive, language-free queries has become increasingly important. Unlike keyword-based retrieval, which often struggles when users cannot precisely verbalize the visual concept they seek, SBIR bypasses linguistic ambiguity by

⋆ Corresponding author: ltnhanh@hcmiu.edu.vn

allowing queries to be conveyed as simple line drawings-making the modality both accessible and semantically expressive.

Despite its promise, SBIR remains a fundamentally challenging problem due to the stark domain gap between sparse abstract sketches and rich high-dimensional photographs. Early approaches based on hand-crafted descriptors were sensitive to drawing style and generalised poorly across object variations. More recent deep-learning methods have achieved significant progress by learning shared feature spaces for sketches and images, yet many continue to suffer from inefficiencies, high computational cost, and inadequate alignment when scaled to large datasets. In particular, existing systems often rely on convolutional networks trained from scratch and employ randomly sampled negatives during metric learning, resulting in slow convergence, suboptimal embedding separation, and substantial GPU memory usage.

At the same time, the practical importance of SBIR has grown rapidly. In retail applications, users can sketch rough silhouettes of shoes or garments on mobile devices to retrieve visually similar products without knowing brand names or precise terms. In law enforcement, composite forensic sketches can be matched against mugshot databases when photographs are unavailable. In clinical workflows, clinicians can outline anomalous structures to retrieve similar diagnostic images, supporting decision-making when textual descriptions fall short. These use cases highlight that SBIR is not merely an academic task - it is a versatile retrieval paradigm with impactful real-world implications.

Motivated by both the scientific and practical value of SBIR, this work aims to develop an efficient and high-performance retrieval system that generalizes to previously unseen object categories. Building upon the dual-encoder architecture of Doodle2Search [3], we identify key limitations in prior work: (i) convolutional backbones trained from scratch lack strong semantic priors and consume significant VRAM, and (ii) random negative sampling in triplet-based optimization provides weak supervision on fine-grained distinctions. To address these issues, we introduce two complementary contributions. First, we replace convolutional encoders with CLIP's ViT-B/32 vision transformer, leveraging its large-scale image-text pretraining to provide robust semantic representations for both sketches and images. Second, we incorporate a hard-batch triplet mining strategy that selects the most challenging negative within each mini-batch for every anchor. This encourages stronger local separation and accelerates convergence by focusing learning on the most confusable examples.

Additionally, we adopt selective fine-tuning, freezing early transformer blocks and updating only later layers and projection heads for sketches, combined with automatic mixed precision to significantly reduce GPU memory consumption. These design choices allow substantially larger batch sizes and faster training while maintaining retrieval accuracy. Our experiments on the Sketchy zero-shot benchmark show that the proposed hybrid loss and CLIP-based architecture improve Recall@K and mean Average Precision, achieving superior zero-shot accuracy while reducing VRAM usage and per-epoch training time.

In summary, this paper presents a computationally efficient, semantically robust SBIR framework that advances the state of the art in zero-shot retrieval. By addressing both modality misalignment and training inefficiencies, our method delivers a practical and scalable solution for real-world sketch-driven search scenarios.

2 Related Work

Sketch-a-Net introduced CNN-based sketch understanding tailored to long-range stroke structures [32], marking one of the earliest demonstrations that deep models could effectively interpret the sparsity and abstraction of freehand drawings. Building on this foundation, earlier studies analyzed human sketch behavior and established the first comprehensive SBIR benchmarks [6], which provided crucial insights into stroke variability and category-level abstraction. Leveraging these datasets, subsequent CNN methods expanded cross-modal learning by incorporating multi-branch fusion architectures[17] and cross-modality representation learning strategies [30], enabling more robust alignment between sketches and photographs.

As research progressed, triplet-loss formulations emerged as a powerful mechanism for enhancing sketch–photo embedding alignment [1]. These efforts were complemented by broader advances in metric learning, such as FaceNet's hard-negative triplet mining [23], deep ranking networks [27], and hard-example mining for image retrieval [18], which collectively underscored the importance of discriminative feature learning for fine-grained retrieval. To further improve scalability, hashing-based representations were explored, with methods such as DSH [15], online sketch hashing [14], and center-similarity hashing [13] demonstrates that compact binary codes can deliver fast and memory-efficient retrieval at scale.

Attention-based architectures introduced another milestone by enabling networks to prioritize semantically important regions. AGS-Net [7] exemplifies how attention can guide cross-domain matching, while more general attention modules, including CBAM [28], Learn-to-Pay-Attention [11], and Strip Pooling [9], have been shown to enhance contextual reasoning and improve robustness to noisy or incomplete sketches. Parallel to these developments, generative models offered an alternative strategy for bridging the sketch–photo modality gap. Techniques such as Sketch2Image-GAN [20], built atop Pix2Pix [10], CycleGAN [34], and multimodal VAE frameworks [25], translate sketches into pseudo-photographic images, allowing retrieval to operate in a single visual domain. Although such methods depend heavily on synthesis quality, they highlight the potential of generative modeling for cross-modal alignment.

More recently, transformer architectures have significantly advanced SBIR performance. Seddati et al. [24] showed that transformers outperform CNNs in capturing global stroke dependencies—an insight supported by breakthroughs in transformer-based vision models such as ViT [4], DeiT [26], and DINO [2]. In parallel, large-scale vision–language models including CLIP [19] and ALIGN [12]

have provided powerful multimodal embeddings that generalize well across domains. Adaptation methods such as conditional prompt learning [33] and survey-driven insights into prompt transferability [16] have further enhanced the flexibility of these models. Building on this, CLIP-based SBIR frameworks [21] have achieved state-of-the-art zero-shot and fine-grained retrieval performance, demonstrating the utility of pretrained multimodal transformers in sketch understanding.

Finally, zero-shot learning theory has strongly influenced modern SBIR design. Foundational work on semantic embedding models such as DeViSE [8], unified perspectives on zero-shot classification [31], and large-scale benchmark analyses [29] have shaped the principles behind cross-modal transfer. These ideas culminate in SBIR systems such as SEM-PCYC [5], which leverage semantic side information and cycle consistency to achieve any-shot retrieval without requiring paired training data. Together, these advances define a rich trajectory of SBIR research and contextualize the direction of our proposed CLIP-based dual-encoder approach.

3 Methodology

3.1 Overall Framework

Our proposed framework addresses limitations of prior systems by integrating a CLIP-based dual encoder with a hybrid supervision strategy that jointly optimizes global and local alignment. The system architecture is illustrated in Figure 1, in which the sketches and images are processed through modality-specific CLIP ViT-B/32 encoders that share the same architecture but maintain separate parameters to specialize for sparse sketches and natural photographs. Each encoder output is passed through a lightweight projection head to produce ℓ_2-normalized embeddings in a shared retrieval space.

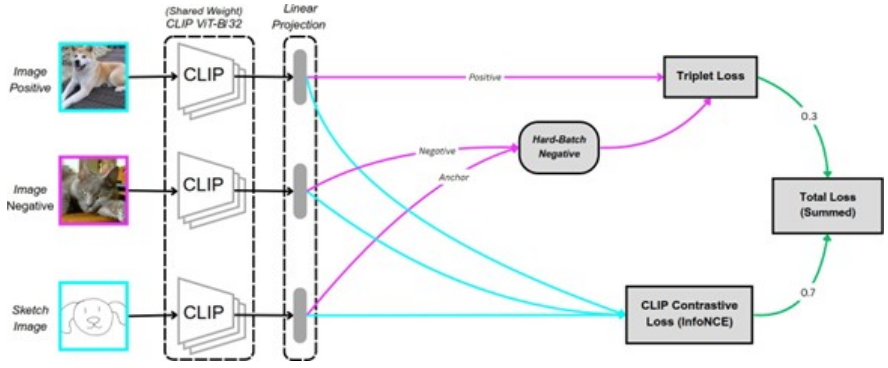

Fig. 1. The proposed framework

To enforce discriminative alignment across modalities, the system simultaneously applies: (1) a CLIP-style InfoNCE contrastive loss for global alignment across the batch, and (2) a triplet loss with hard-batch mining for fine-grained local separation. The total loss is a weighted sum of both objectives, enabling the model to benefit from global semantic structure and targeted refinement of hard negatives.

3.2 CLIP-Based Encoders

Sketch and Image Encoders. We adapt CLIP's ViT-B/32 vision transformer as the backbone for both sketches and images. Each sketch is rasterized or resized to 224×224, triplicated to three channels, and normalized using CLIP preprocessing statistics. Images are processed using the same normalization pipeline. The ViT architecture splits each input into 32×32 patches, embeds them, and processes them through a stack of self-attention layers to yield a 512-dimensional feature vector. While the sketch and image encoders share the ViT-B/32 architecture, their weights are not shared to accommodate the domain gap between sparse sketches and natural images. A two-layer projection head (`Linear` $\rightarrow$ `ReLU` $\rightarrow$ `LayerNorm`) refines representations before final ℓ_2-normalization.

Fine-Tuning Strategy. To stabilize training and preserve CLIP's semantic priors, we freeze early transformer blocks and fine-tune only the later layers along with the projection heads. Sketch-specific augmentations, including random horizontal flips, small rotations ($\pm 10^\circ$), and random erasing, improve robustness to drawing style variations. Training is performed with automatic mixed precision (AMP) to reduce GPU memory usage.

3.3 Loss Functions

CLIP Contrastive Loss (InfoNCE). For a batch of N sketch image pairs, we compute cosine similarities between ℓ_2-normalized sketch embeddings s_i and image embeddings v_j, forming a similarity matrix $M \in \mathbb{R}^{N \times N}$. A bidirectional InfoNCE loss is applied such that each sketch must identify its corresponding image among all others, and vice versa.

Let τ be a learnable parameter. The contrastive loss is:

$$\mathcal{L}_{\text{clip}} = \frac{1}{2N} \sum_{a=1}^{N} \left[-\log \frac{\exp(s_a \cdot v_a/\tau)}{\sum_{b=1}^{N} \exp(s_a \cdot v_b/\tau)} - \log \frac{\exp(v_a \cdot s_a/\tau)}{\sum_{b=1}^{N} \exp(v_a \cdot s_b/\tau)} \right]. \quad (1)$$

This loss provides broad batch-level negative sampling, leading to stable gradients and strong global alignment.

Triplet Loss with Hard-Batch Mining. To complement InfoNCE, we incorporate a bidirectional triplet loss. After computing pairwise distances among all samples in a batch, we identify for each anchor the hardest negative sample, the non-matching item with the smallest Euclidean distance (or highest cosine similarity).

Given an anchor a, its positive p, and hardest negative n, the triplet loss is defined as:

$$\mathcal{L}_{\text{triplet}} = \max\left(\|a-p\|_2^2 - \|a-n^{\|_2^2+\alpha,\ 0}\right), \tag{2}$$

where the margin α is set to 0.3. This loss is applied symmetrically for sketch anchors and image anchors.
The final training objective is a weighted sum of global and local alignment terms:

$$\mathcal{L}_{\text{total}} = 0.7\,\mathcal{L}_{\text{clip}} \ + \ 0.3\,\mathcal{L}_{\text{triplet}}. \tag{3}$$

This balance enables fast convergence and strong generalization to unseen categories.

4 Experimental Results

4.1 Implementation details

All experiments were conducted in Python 3.9 using PyTorch 2.0 as the primary deep-learning framework. We adopt the official OpenAI CLIP repository to load the pretrained ViT-B/32 backbone for both sketch and image encoders. Supporting libraries (including NumPy, SciPy, scikit-learn, Pandas, and Matplotlib) were used for data preprocessing, analysis, and visualization, while TensorBoard was employed for monitoring training dynamics. Training was performed on a single NVIDIA RTX 3060 GPU with 24 GB of VRAM running Ubuntu 22.04 LTS, although the codebase is fully compatible with both Linux and Windows environments.
In terms of training data, we use the Sketchy dataset, introduced by Sangkloy et al. (2016) [22], which is a widely used benchmark for sketch-based image retrieval, designed to bridge the gap between abstract freehand drawings and real photographs. It contains crowd-sourced sketches collected via Amazon Mechanical Turk, with each object photo paired with multiple hand-drawn depictions that capture diverse drawing styles and abstraction levels. Zero-shot SBIR evaluations typically use the Sketchy-25 split, where 100 classes are used for training and 25 unseen classes for testing; an alternative Sketchy-21 split removes classes overlapping with ImageNet to ensure truly novel test categories [13]. Preprocessing is minimal, as vectorized sketches are rasterized for network input, and most studies use the full sketch set without heavy filtering.

4.2 Results and Evaluations

We test our model on several cases, which are exhibited in Figure 2. The qualitative results reveal that our model performs reliably on objects with clear geometric structures. For categories such as Alarm Clock or Elephant, sketches emphasizing distinctive circular or organic contours consistently retrieve correct photographic matches, indicating that the model effectively encodes dominant shape cues. Performance becomes more variable with moderately complex objects. Motorbike sketches generally return two-wheeled vehicles, yet incomplete

depictions, missing exhausts or handlebars, occasionally lead to confusion with bicycles or scooters. For more deformable shapes such as Jellyfish, the model handles morphological variability reasonably well, though overlap with squid or octopus images persists when tentacle detail is minimal. Finally, Gun retrievals correctly capture key structural parts-barrel, trigger, grip-though sketches resembling toys or tools occasionally introduce non-weapon results. This highlights the importance of context-aware safeguards in real-world deployment.

Figure 3 presents the visual comparison between our results and the baseline model Doodle2Search. Qualitative inspection of top-five retrievals shows that the model aligns structure and semantics well, even when sketches are abstract or lack fine detail. Organic categories like Apple achieve strong consistency with only mild subtype confusion, while geometric objects often reach near-perfect precision. However, classes with large intra-class variation still exhibit weaker top-1 accuracy in later epochs, suggesting that global alignment alone does not fully separate subtle hard negatives in the embedding space.

We further evaluate retrieval performance of our model on the Sketchy dataset under a zero-shot setting, where test classes are never observed during training. Following standard SBIR practice, we report two metrics: Recall@K and mean Average Precision (mAP). Recall@K measures whether the correct image appears among the top-K retrieved results for each sketch, while mAP captures the overall ranking quality across all relevant images. RecallK is defined as:

$$\text{Recall@K} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\big(y_i \in \text{TopK}(i)\big), \tag{4}$$

where N is the total number of test sketches, y_i is the ground-truth class label for sketch i, and TopK(i) denotes the indices of the top-K retrieved images ranked by cosine similarity. The indicator function returns 1 if the correct label is present in the top-K results and 0 otherwise. The **mAP** metric (mean Average Precision) evaluates the full ranking of retrieved images by averaging the precision at all relevant positions in the ranked list. Both metrics are computed on the held-out test classes of Sketchy to ensure reliable zero-shot generalization performance.

Table 1 shows above evaluation. Epochs to $\geq 99\%$ Recall100 measures the number of training epochs required for the model to achieve at least 99% retrieval accuracy within the top 100 retrieved images. Lower values indicate faster convergence and higher training efficiency. The dual-loss design, combining InfoNCE with hard-batch triplet mining plays a central role in our model's efficiency and retrieval accuracy. By selecting the hardest negatives within each batch rather than sampling them randomly, the model learns to separate closely related instances early in training. This targeted supervision enables our system to reach $\geq 99\%$ Recall100 in only 50 epochs, a $3\times$ faster convergence compared to Doodle2Search, which requires around 150 epochs to achieve the same threshold.

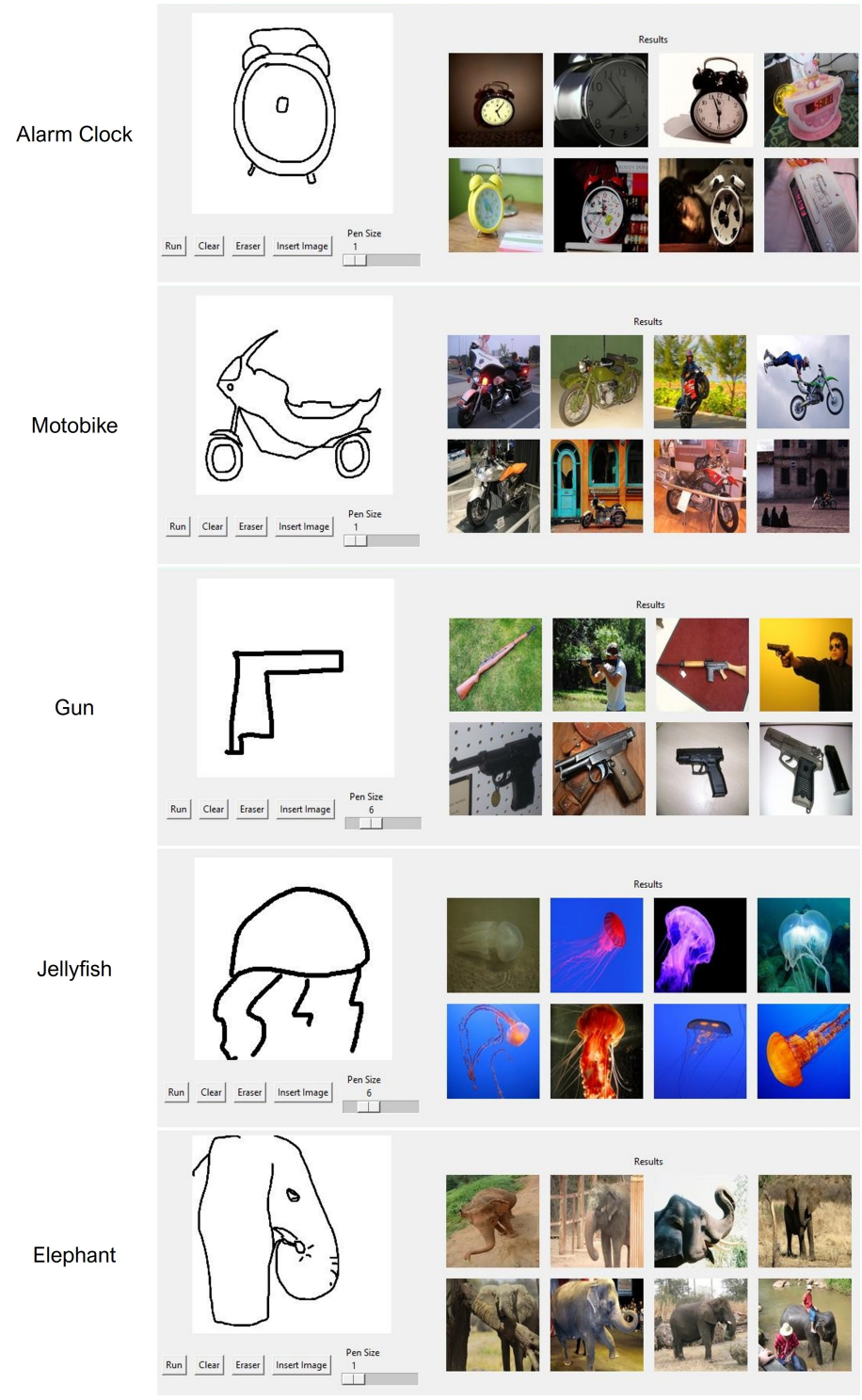

Fig. 2. Results generated by our model compete with the baseline model

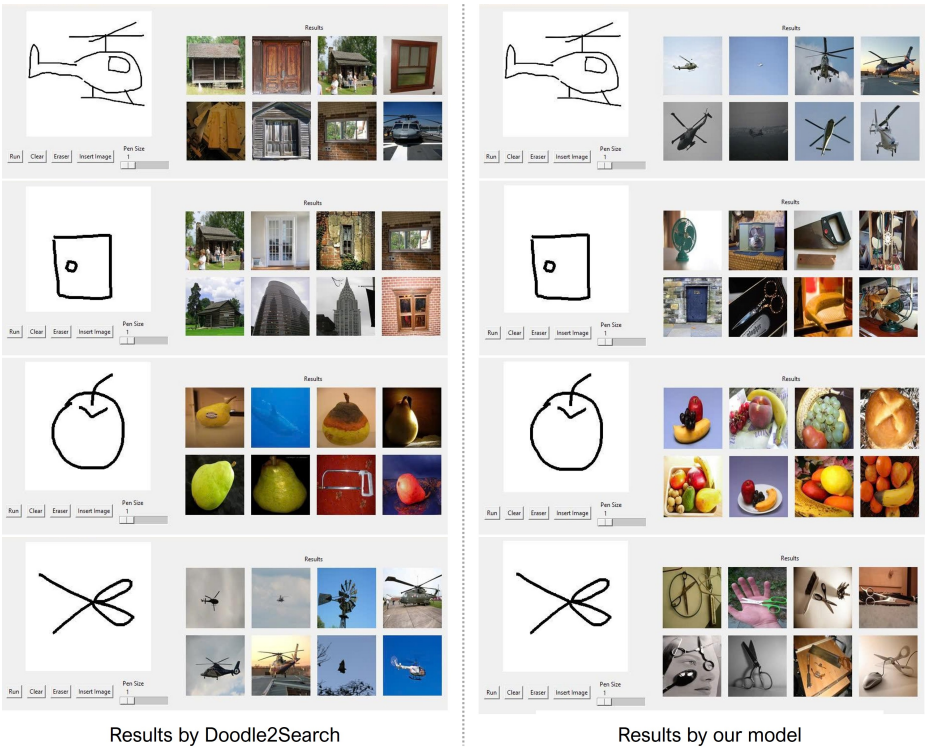

Fig. 3. Results generated by our model compete with the baseline model

Our use of CLIP's pretrained ViT-B/32 encoder further strengthens this advantage. The transformer's large-scale vision-language pretraining provides strong semantic priors that transfer effectively to freehand sketches, allowing the model to interpret abstraction and stylistic variation despite being trained primarily on photographic imagery. As a result, the model achieves a higher Recall@100 at epoch 50 (0.9992 vs. 0.9810) and delivers improved mAP. Finally, sketch-specific augmentations (such as random erasing and rotational jitter) help expand the variability of the training distribution, improving robustness and preventing overfitting to limited sketch styles. Together, these components explain the substantial improvement in both learning speed and retrieval quality observed in our results.

Table 1. Comparison of retrieval performance and training efficiency.

Model	Epochs to $\geq$ 99% Recall@100	Recall@100 @ Epoch 50	Relative Speed-Up	mAP (%)
Our Model	50	0.9992	3×	81.4
Doodle2Search	150	0.9810	1×	74.2

5 Conclusion

In this work, we presented an efficient SBIR framework that integrates dual CLIP ViT-B/32 encoders with a hybrid InfoNCE and hard-batch triplet loss. By leveraging CLIP's multimodal pretraining and targeted negative mining, the model learns embeddings that are both semantically aligned and fine-grained. Experiments on the Sketchy benchmark show rapid convergence (achieving near-perfect Recall@100 by epoch 50) and strong zero-shot performance across diverse object categories. Efficiency is a notable strength: automatic mixed precision and in-batch hard negative mining significantly reduce memory use and training time compared to CNN-based baselines like Doodle2Search, while also improving mAP. These properties make the system well suited for scalable retrieval. Remaining challenges include handling highly variable fine-grained classes. Future work may incorporate hierarchical semantic constraints, generative augmentation for harder negatives, auxiliary modalities, and large-scale indexing with user feedback to further enhance precision and deployment readiness.

References

1. Bui, T., Ribeiro, L., Ponti, M., Collomosse, J.: Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network. Computer Vision and Image Understanding **164**, 27–37 (2017). https://doi.org/10.1016/j.cviu.2017.06.007
2. Caron, M., Touvron, H., Misra, I., et al.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021). https://doi.org/10.1109/ICCV48922.2021.00953
3. Dey, S., Riba, P., Dutta, A., Llados, J., Song, Y.Z.: Doodle to search: Practical zero-shot sketch-based image retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2179–2188 (2019)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
5. Dutta, A., Akata, Z.: Semantically tied paired cycle consistency for any-shot sketch-based image retrieval. International Journal of Computer Vision **128**(10–11), 2684–2703 (2020). https://doi.org/10.1007/s11263-020-01350-x
6. Eitz, M., Hays, J., Alexa, M.: How do humans sketch objects? ACM Transactions on Graphics **31**(4), 1–10 (2012). https://doi.org/10.1145/2185520.2185527
7. Feng, Z., Huang, S., Lai, J.: Attention-guided siamese network for clothes-changing person re-identification. In: Image and Graphics (Lecture Notes in Computer Science, vol. 12889). pp. 314–325. Springer (2021). https://doi.org/10.1007/978-3-030-87358-5_25

8. Frome, A., Corrado, G.S., Shlens, J., et al.: Devise: A deep visual-semantic embedding model. In: NeurIPS (2013)
9. Hou, Q., Zhang, L., Cheng, M.M., Feng, J.: Strip pooling: Rethinking spatial pooling for scene parsing. In: CVPR. pp. 4003–4012 (2020). https://doi.org/10.1109/CVPR42600.2020.00407
10. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 5967–5976 (2017). https://doi.org/10.1109/CVPR.2017.632
11. Jetley, S., Lord, N.A., Lee, N., Torr, P.: Learn to pay attention. In: ICLR (2018)
12. Jia, C., Yang, Y., Xia, Y., et al.: Scaling up visual and vision-language representation learning with contrastive learning. In: ICML (2021)
13. Jiang, W., Wang, X., Jiang, S., et al.: Deep sketch hashing with center similarity. IEEE Transactions on Image Processing **29**, 5314–5327 (2020). https://doi.org/10.1109/TIP.2020.2971443
14. Li, J., et al.: Online sketching hashing. In: AAAI. pp. 7105–7112 (2018). https://doi.org/10.1609/aaai.v32i1.12090
15. Liu, L., Shen, F., Shen, Y., Liu, X., Shao, L.: Deep sketch hashing: Fast free-hand sketch-based image retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2298–2307 (2017). https://doi.org/10.1109/CVPR.2017.247
16. Liu, Z., et al.: A survey on visual prompt learning. arXiv preprint arXiv:2307.XXX (2023)
17. Qi, Y., Song, Y.Z., Zhang, H., et al.: Sketch-based image retrieval via cnn and fusion. In: ACM Multimedia. pp. 69–73 (2016). https://doi.org/10.1145/2964284.2967308
18. Radenović, F., Tolias, G., Chum, O.: Cnn image retrieval learns fine-grained details. In: CVPR. pp. 1955–1963 (2016). https://doi.org/10.1109/CVPR.2016.215
19. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
20. Ramy, A., Barakat, N., Barakat, N.H.: Sketch to image using generative adversarial networks (gan) (2022). https://doi.org/10.13140/RG.2.2.16797.38889
21. Sain, A., Bhunia, A.K., Chowdhury, P.N., Koley, S., Xiang, T., Song, Y.Z.: Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2765–2775 (2023). https://doi.org/10.1109/CVPR52729.2023.00271
22. Sangkloy, P., Burnell, N., Ham, C., Hays, J.: The sketchy database: learning to retrieve badly drawn bunnies. ACM Transactions on Graphics **35**(4), 1–12 (2016). https://doi.org/10.1145/2897824.2925954
23. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. pp. 815–823 (2015). https://doi.org/10.1109/CVPR.2015.7298682
24. Seddati, O., Dupont, S., Mahmoudi, S., Dutoit, T.: Transformers and cnns both beat humans on sbir. arXiv preprint arXiv:2209.06629 (2022). https://doi.org/10.48550/arXiv.2209.06629
25. Sohn, K., Lee, H., Yan, X.: Learning structured latent representations with deep generative models. In: NeurIPS. pp. 3040–3048 (2015)
26. Touvron, H., Cord, M., Douze, M., et al.: Training data-efficient image transformers. In: ICML (2021)
27. Wang, J., Zhou, F., Wen, S., Liu, X.: Learning fine-grained image similarity with deep ranking. In: CVPR. pp. 1386–1393 (2014). https://doi.org/10.1109/CVPR.2014.180

28. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: ECCV. pp. 3–19 (2018). https://doi.org/10.1007/978-3-030-01234-2_1
29. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning — the good, the bad and the ugly. In: CVPR. pp. 3077–3086 (2017). https://doi.org/10.1109/CVPR.2017.330
30. Xu, P., Hospedales, T.M.: Sketch-a-class: Cross-modality feature learning. In: IJCAI. pp. 3163–3169 (2017)
31. Yang, Y., Hospedales, T.M.: A unified perspective on zero-shot image classification. In: CVPR (2015). https://doi.org/10.1109/CVPR.2015.7299154
32. Yu, Q., Yang, Y., Song, Y.Z., Xiang, T., Hospedales, T.: Sketch-a-net that beats humans. arXiv preprint arXiv:1501.07873 (2015). https://doi.org/10.48550/arXiv.1501.07873
33. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR. pp. 16816–16825 (2022). https://doi.org/10.1109/CVPR52688.2022.01637
34. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232 (2017). https://doi.org/10.1109/ICCV.2017.244