Neural Computing and Applications (2023) 35:5395–5408
https://doi.org/10.1007/s00521-022-07978-9

ORIGINAL ARTICLE

# SBIR-BYOL: a self-supervised sketch-based image retrieval model

Jose M. Saavedra[1] · Javier Morales[2] · Nils Murrugarra-Llerena[3]

Received: 19 May 2022 / Accepted: 17 October 2022 / Published online: 3 November 2022


**Abstract**
Sketch-based image retrieval is demanding interest in the computer vision community due to its relevance in the visual perception system and its potential application in a wide diversity of industries. In the literature, we observe significant advances when the models are evaluated in public datasets. However, when assessed in real environments, the performance drops drastically. The big problem is that the SOTA SBIR models follow a supervised regimen, strongly depending on a considerable amount of labeled sketch-photo pairs, which is unfeasible in real contexts. Therefore, we propose SBIR-BYOL, an extension of the well-known BYOL, to work in a bimodal scenario for sketch-based image retrieval. To this end, we also propose a two-stage self-supervised training methodology, exploiting existing sketch-photo pairs and contour-photo pairs generated from photographs of a target catalog. We demonstrate the benefits of our model for the eCommerce environments, where searching is a critical component. Here, our self-supervised SBIR model shows an increase of over 60% of mAP.

**Keywords** Sketch-based image retrieval · Self-supervision · Deep-learning · Representation learning

## 1 Introduction

Sketch-based understanding plays an important role in visual perception systems and communication between humans that transcends language barriers. At the beginning of artificial intelligence, Hubel and Wiesel [1] showed how the biological visual cortex highly responds to edge patterns; and recently, Walther et al. [2] also showed the semantic power of image contours to interpret our environment. They found that the primary visual cortex produces similar responses when stimulated by a regular image or its corresponding contour map. This study infers that temporal information of sketch used to boost sketch representations [3] is not a critical component.

Sketch understanding is deeply connected to cognition development [4]. Infants draw sketches to understand the natural environment, sometimes through rough drawings, and people also externalize and communicate simple and complex ideas through them. Indeed, people draw schemes or maps to understand and unfold complex structures and processes. In this vein, Mukherjee et al. [5] studied how we effortlessly associate a drawing with objects in the world. The compositional nature of object concepts allows us to decompose objects and drawings into semantically meaningful parts. Furthermore, free-hand sketches go beyond representing shapes; they can express emotions [6], needs [7], or even dynamic scenes [8].

Due to the critical role of sketch understanding in visual perception and the ubiquitous use of touch-screen devices that makes sketching an entertaining and convenient mechanism, the computer vision community has started to pay special attention to this area. For instance, the main computer vision conferences already include workshops to promote research and applications on this topic. In this vein, influenced by the deep-learning bloom, we have seen

Jose M. Saavedra, Javier Morales and Nils Murrugarra-Llerena have contributed equally to this work.

✉ Jose M. Saavedra
jmsaavedrar@miuandes.cl

Javier Morales
javiermoralesr95@gmail.com

Nils Murrugarra-Llerena
nmurrugarrallerena@weber.edu

[1] Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Monseñor Alvaro del Portillo, Santiago 7620001, RM, Chile

[2] Department of Computer Science, University of Chile, Av. Blanco Encalada 2120, Santiago 8370459, RM, Chile

[3] School of Computing, Weber State University, 3848 Harrison Blvd, Ogden UT 84408, Utah, United States

advances in a diversity of tasks like sketch classification [9, 10], sketch representation learning [11, 12], sketch-guided object localization [13], sketch-based image and video retrieval [7, 8, 14–18], sketch-to-photo translation [19, 20], among others. However, these advances have been centered on supervised models, requiring a huge amount of labeled data for training, which limits its potential application to the industry.

Sketch-based image retrieval (SBIR) [21–23] is an attractive task, having the largest number of publications in the sketch-based understanding area. This task extends the well-known content-based image retrieval problem (CBIR), where the query is expressed by a simple free-hand drawing (the sketch query). Unlike CBIR, a SBIR model does not need an example photograph of what we need to trigger the search; we just express our thoughts or needs through drawings. Beyond being a natural manner of communication, sketching is also a convenient and comfortable modality, given the prevalent use of mobile devices.

A natural environment where we can leverage SBIR models is eCommerce. In eCommerce, there are stores with plenty of catalogs with diverse products. We have big players like marketplaces selling millions of products, each represented by a textual description and a set of images. We also have potential consumers looking to satisfy their needs and stores demanding more customers. Therefore, an effective search engine is critical to increasing the satisfaction of all the players in this business.

In eCommerce, traditional search engines have been based on textual queries, primarily keywords. Even though they convey highly semantic information, texts also limit the underlying expressiveness, missing fine details of our intentions. An effective manner to overcome the limitation of text-based queries is through visual search, which has started to call the attention of retailers, especially after the explosion of deep-learning models that have contributed to improving retrieval effectiveness. A visual search is instantaneous (!'just take a picture and search!), and it is fully expressive of image details.

However, visual search has a serious drawback. It requires an example of what a user wants, which sometimes is cumbersome. Commonly, one needs to search for something that we do not own. We have ideas, abstractions or thoughts about what we desire, but we do not necessarily have an example image representing them. Here, a free-hand drawing expressing our intentions is a plausible solution.

A SBIR engine is an attractive modality for eCommerce, but unfortunately, we have not seen a proliferation of sketch-based engines yet; why not? Although state-of-the-art SBIR models have shown growing effectiveness in public datasets, when the models are evaluated in real environments, the retrieval performance decreases to very poor levels. Indeed, in a recent work, Torres and Saavedra [7] reported a mAP of around 15% when a state-of-the-art SBIR model is applied to real eCommerce data. Moreover, they observed a low semantic level in the results. For instance, Figure 1 shows results of retrieval when the query represents a *cap*. We observe that the model cannot understand the concept behind the query; it is only trying to match shapes, as we can see in the resulting product of the first position. This phenomenon is known as the **semantic-gap**.

To better understand why the current models produce poor performance in real data, we must go deep into how these are trained. The state-of-the-art method for sketch-based image retrieval is based on incrementally training Siamese networks, from coarse-grained to fine-grained similarity [14]. To this end, the model requires a huge amount of labeled data in the form of sketch-photo pairs. We need to define positive and negative pairs. Actually, there are few amounts of data with fine-grained labels, and producing it in real environments is somehow prohibitive. Thus, the current models, trained on public datasets, do not consider the information from the real target products.

To exploit the vast amount of data without requiring labeling, we explore self-supervised regimens working under bimodal inputs. Hence, in this work, we propose SBIR-BYOL, an extension of BYOL that works in a bimodal environment for sketch-based image retrieval. We present a training methodology for SBIR-BYOL that leverages public sketch-photo pairs and unlabeled photographs from target domains to improve retrieval performance.

Our self-supervised proposal is easy to implement in real-world applications, and it shows to outperform SOTA methods, just leveraging photographs from a target catalog. We see an increment of 67% of precision when we evaluate searching in eCommerce catalogs.

This document is organized as follows. Section 3 presents the related work. Section 3 describes SBIR-BYOL in detail. Section 2 discusses the experimental evaluation and the achieved results. Section 5 describes the relevance of having a method like this proposal in other vision tasks. Finally, Section 6 describes the final remarks.

## 2 Related work

Computer-based sketch understanding is an emerging area within computer vision, with its own characteristics and challenges. Sketching represents a primitive means of communication between humans to transmit ideas and abstractions from ancient times [24]. Sketching is also a direct expression of creativity. For instance, architects and

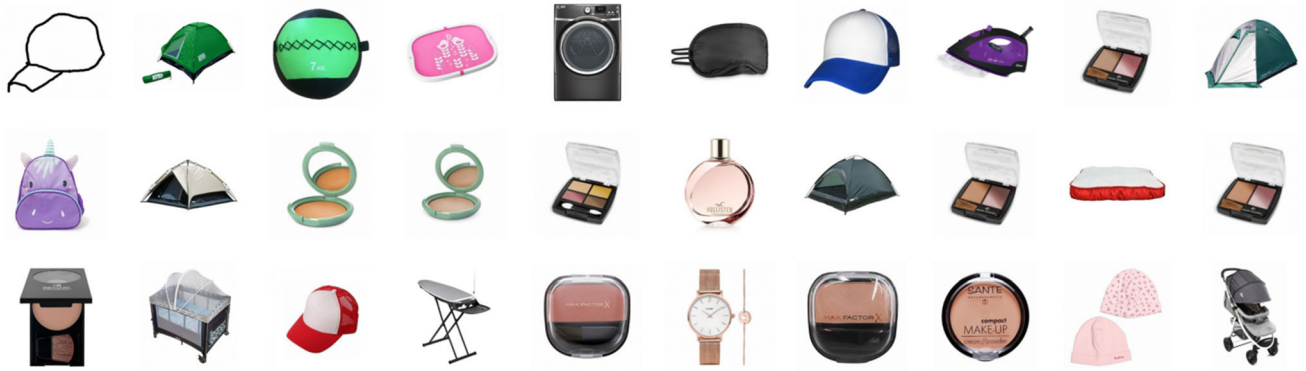


**Fig. 1** A SBIR result in eCommerce using a state-of-the-art SBIR model. The first image is the query sketch, and the next is the retrieved photographs ordered from the most similar to the less similar, from left to right and top to bottom

designers express creativity and conception by hand-sketching [25].

Moreover, sketching is strongly related to human cognitive development [4, 26, 27]. In this vein, Fernades et al. [27] showed that drawing improves memory and creativity in normal aging individuals and those with cognitive impairments. Furthermore, De Andrade et al. [26] showed the benefits of collaborative drawing for learning and collective thinking, enhancing social-cognitive interactions among persons.

Besides the relevant role of sketching in visual perception, the massification of touch-screen devices allows sketching to become an easy and convenient manner of interacting with machines [16, 17].

Indeed, there have been a proliferation of sketch-based applications related with classification [9, 10], representation learning [11, 28–30], sketch-based image synthesis [19, 20, 31, 32]. Among these tasks, sketch-based image retrieval represents the one with the greatest number of publications.

## 2.1 Sketch-based image retrieval

Sketch-based image retrieval (SBIR) is a growing field in computer vision that consists of retrieving a collection of photographs or images resembling a query sketch. The input query is formulated as a simple hand drawing composed uniquely of strokes like those of Fig. 2 to make the querying process as simple as possible.

Sketch-based image retrieval aims to find discriminative representations (feature vectors) from two kinds of images, hand-drawn sketches that serve as queries and photographs or regular images that represent the target catalog. These representations generate a shared feature space. We aim to produce a *semantic space*, where sketches and photographs representing the same concept fall close together, while those representing different meanings fall apart from each other.

We have seen a diversity of proposals addressing the mentioned problem. The first approaches were based on low-level features using histograms of orientations [9, 22, 23, 33, 34]. Regular images were previously converted into sketch-like images by an edge detection method [35, 36] to homologate the visual representations obtained from sketches and photographs. Other authors proposed mid-level representations to extract meaningful information from sketches [21, 37].

However, it was not until the explosion of deep learning that performance rates increased. The work of Yu et al. [38] and Sangloy et al. [39] was pioneers in using convolutional neural networks for SBIR. However, the training methodology proposed by Bui et al. [14] marked a significant step in performance. The authors proposed an incremental methodology for training, from coarse-grained to fine-grained similarity. In addition, they showed the benefits of combining a contrastive or triplet loss with cross-entropy. Thus, the model learns a semantic embedding space keeping the capability to discriminate between different classes.

When we transfer our solutions to real environments, it is essential to be aware of the efficiency of the models in terms of memory and processing time. Commonly, SBIR methods produce floating-point representations in high-dimensional feature space, which impact the required resources to implement a SBIR model. Recently, Torres and Saavedra [7] studied different compact representations in the context of sketch-based image retrieval. They leverage a local-structure preserving reduction technique like UMAP [40] to reduce the underlying feature space to a very low-dimensional space without reducing the effectiveness in retrieval.

Other works extend the traditional sketch-based image retrieval problem, as discussed above, to other problems. For instance, researchers are focusing on querying images by colored sketches [15] or using sketches to express motion and retrieve videos [8, 18]

**Fig. 2** A sample of sketches used as queries

Although we have seen a proliferation of sketch-based image retrieval proposals, all these share a critical drawback limiting the application to industry. All discussed methods are produced by a supervised learning methodology, requiring a data labeling process. Consequently, even though the SOTA methods show a high performance in public datasets, it does not follow the same direction in real applications. In fact, Torres and Saavedra [7] showed a large gap when a SOTA model was evaluated on real eCommerce data. They showed a reduction of almost 70%. But what is the reason for this phenomenon? The answer seems simple. Real applications do not have data appropriately labeled as required by supervised models (i.e., sketch-photo pairs). There is another critical problem with supervised models; they may only look at a subset of all data, which may not generalize well for unseen data.

Therefore, in this work, we present the first self-supervised model for sketch-based image retrieval that is capable of exploiting the data available in the application domain. As a use case, we present results in the context of eCommerce, where a retrieval model is required. Our proposal increases the effectiveness up to 67% without requiring extra labeling. Our model can be easily adjusted to different eCommerce just by looking at the images in the target catalog.

Our proposed model was inspired by BYOL [41], a state-of-the-art strategy aiming to learn visual representations in a self-supervised manner. It was proposed to work with single-modal data, as in the case of images. The model is trained under a teacher-student architecture, where the student branch (a.k.a. online branch) tries to learn representations close to those provided by the teacher branch (a.k.a. target branch). Incrementally, the student gets improved representations. Furthermore, as the student improves, the teacher slowly updates its weights through the *exponential moving average* mechanism, using the learned student knowledge.

BYOL is trained over positive pairs only, and its stop-gradient strategy on the teacher branch seems to prevent learned representations from collapsing. Therefore, we extend BYOL to work in a bimodal environment to handle sketches and images in a shared domain. We also present a methodology to build sketch-photo pairs under a self-supervised regimen.

## 3 SBIR-BYOL

In this section, we present an adaptation of the BYOL model, named SBIR-BYOL, to work in a bimodal context for sketch-based image retrieval (SBIR). Our proposal also produces sketch-photo pairs in a self-supervised manner. The scheme of our BYOL-based proposal is depicted in Fig. 3.

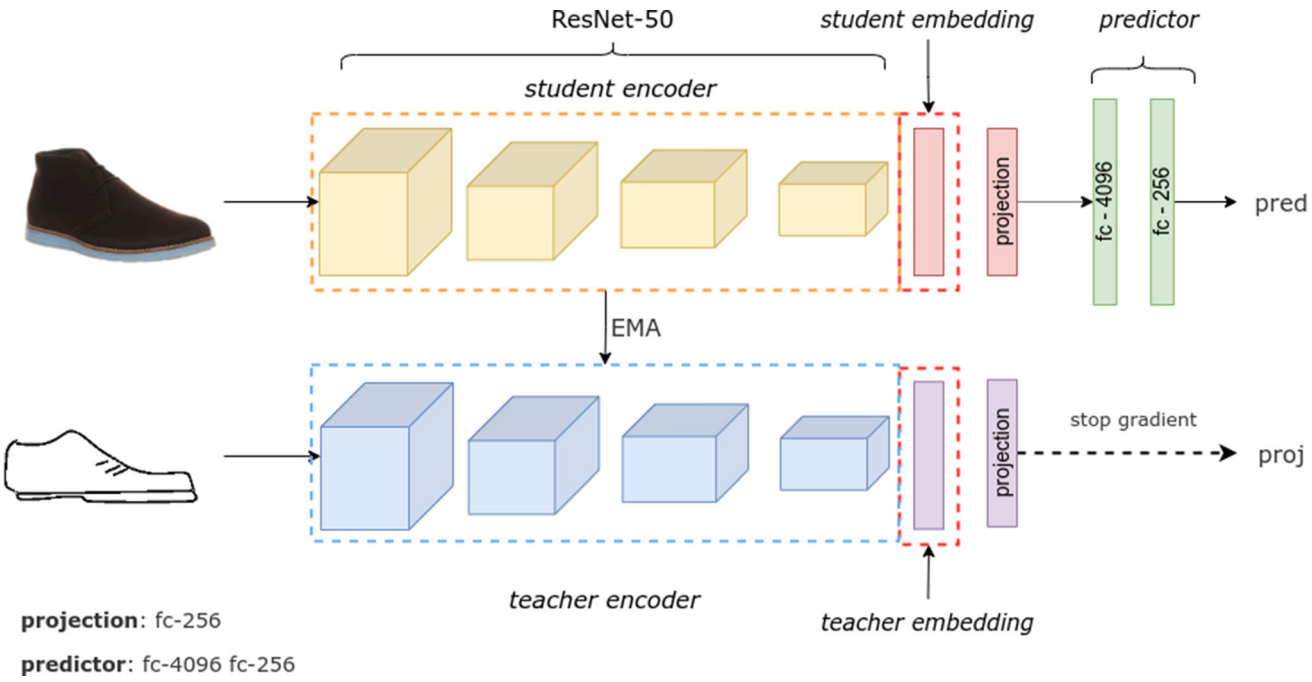



**Fig. 3** Scheme of SBIR-BYOL, we assign a single modality to each of BYOL's branches. The student network learns to extract features from photographs in the same way the teacher does for sketches

Even though our proposal is based on BYOL, there are still some critical questions that we should answer:

- What modality (image or sketch) should go to the online and what should go to the target branch?
- How should the backbones be initialized?
- How should sketch-photo pairs be built?
- How to teach the abstraction level of sketches to the model?

We address the aforementioned questions in the following paragraphs.

### 3.1 Teacher or student branch

A critical decision in our proposal is determining what branch of the model should be connected with the sketch and photograph inputs. We experimentally determined that the best configuration is to assign the input photograph to the student branch (the online one) and the input sketch to the teacher branch (the target one). In this manner, the model is teaching the online network to interpret photographs in the same way the target network interprets sketches.

The intuition behind preferring the target branch for the sketches is related to the amount of information a photograph conveys with respect to a sketch. A photograph conveys more information than a sketch does. Thus, it would be easier for the model to go from photograph to sketches than in a reverse manner. As the student tries to mimic the teacher, putting a photograph in the student branch allows the model to converge more easily. If we connect a sketch in the student branch, this does not have enough information to mimic the photograph's representations, making learning more difficult.

### 3.2 Backbone initialization

The model strongly depends on a good initialization of the target branch, the teacher. As this process sketches, we pretrained the corresponding backbone in a self-supervised manner using the BYOL model trained on a sketch dataset as discussed in [30].

### 3.3 Classical data augmentation

Data augmentation is the core of discriminative self-supervised models such as BYOL. Indeed, BYOL applied data augmentation operations over the images to produce inputs to the two involved branches in the network. However, it does not seem to be a good strategy for sketch-based image retrieval because different transformations on the sketches and their corresponding photographs may produce incompatibilities that negatively affect the learning process. Therefore, in this proposal, we do not apply data augmentation to the inputs. Even though in the next sections, we will present some results of using transformation over the inputs.

### 3.4 Making sketch-photo pairs

Making pairs are the bottleneck of a supervised training methodology. We address this problem by leveraging high-level contour maps extracted from photographs. This strategy was previously used in the task of image generation from sketches by Chen et al. [19]. To this end, we take advantage of the *Pixel Difference Network* (PiDiNet), a recent model for edge detection combining traditional edge-detection operators with modern convolutional neural networks [42]. An example of a contour map produced by PiDiNet is shown in Fig. 4. The generated sketch shows different levels of detail, which are exploited by our SBIR-BYOL.

### 3.5 Leveraging real sketches

One drawback of using contour maps is that these do not represent the abstraction of sketches. For instance, Fig. 5 shows the differences between contour maps and real sketches for *a shoe*. Sketches are diverse and unique. Humans draw sketches uniquely to represent the most salient parts of objects. To reduce restricted creativity from contour maps, we leverage the Sketchy dataset [39], which provides a set of sketch-photo pairs. In this manner, we propose a two-stage training process, where we start to train our model with Sketchy's pairs and then with contour-photo pairs.

## 4 Experimental evaluation

This section describes the datasets and protocols used in our experimental evaluation. Afterward, in the last part of this section, we will present a detailed discussion of the achieved results.

### 4.1 Datasets

We take advantage of a diversity of datasets commonly used in the context of sketch-based image retrieval. As we aim to boost the performance of the models in real environments, we evaluate our approach in the context of eCommerce in the wild. In the following lines, we provide more details of each used dataset.

**Fig. 4** Example of a contour map produced by PiDiNet

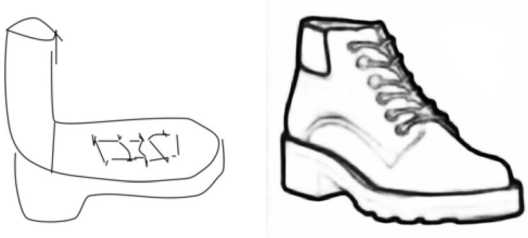


**Fig. 5** A comparison between real sketches and contour maps. Sketches are subjective and human-dependent. Each person has a unique drawing style to represent an object

#### 4.1.1 Sketchy

The Sketchy dataset [39] is a collection of sketch-photo pairs, having 75,471 different hand-drawn sketches, where each sketch is associated with one of 12,500 photographs coming from 125 categories.

The Sketchy dataset is composed of fine-grained pairs. Thus, a sketch is associated with one photograph sharing details like the pose, besides the own category. However, many photographs can be associated with one sketch, producing approximately five matching sketches for each photograph.

We leverage this dataset to allow the model to incorporate the abstraction of real sketches. In this manner, Sketchy's sketch-photo pairs are used during the first stage of our training methodology. Our experiments show that this stage helps the model to increase its performance.

#### 4.1.2 Flickr25K

This dataset was proposed for training purposes. It is an extension of the sketch dataset proposed by Eitz et al. [23], which consists of 20,000 different sketches distributed into 250 categories. Flick25K adds 25,000 photographs distributed in the same 250 categories. Thus, it is possible to build sketch-photo pairs by randomly picking a sketch and a photograph from the same category.

Flickr25K is commonly used along with Flickr15K, where the former is used for training and the last for evaluation. In our experiments, we use the same methodology when Flickr15K is used for evaluation.

#### 4.1.3 Flickr15K

This is a traditional dataset to evaluate SBIR models. Flickr15K consists of 15,000 photographs, mostly outdoor scenes, distributed into 33 different categories [22]. This dataset also includes 330 sketches used for querying, distributed into the same 33 classes, having ten sketches per class. The categories of this dataset include animals, plants,

famous landmarks, and everyday objects. We use this dataset just for evaluation purposes.

#### 4.1.4 eCommerce

This dataset was recently proposed by Torres et al. [7], aiming to have a closer approximation of the performance of SBIR models in real environments. This eCommerce dataset consists of two non-overlapping sets. The first one is a collection of 50701 photographs of diverse products representing what an eCommerce sells. This collection is used for training only. It is important to note that this set does not contain any query sketch. The second set consists of 5665 photographs and 666 real query sketches. This second set is used for evaluation, where the collection of photographs is the target catalog for searching. Both sets contain images distributed among 141 product categories.

A summary of the datasets used in our experiments is shown in Table 1.

### 4.2 Training protocol

SBIR-BYOL is trained in a self-supervised fashion for sketch-based image retrieval. The goal is to improve the retrieval performance just by looking at the collection of photographs where the search takes place. This methodology fits the eCommerce scenario, where a SBIR model should provide high effectiveness across several eCommerce sites.

As described in Sect. 3, we leverage contour maps extracted from photographs to build sketch-photo pairs. Here, the training sketch is a pseudo-sketch.

### 4.3 Discussion of results

Table 2 presents the results on the eCommerce dataset after training SBIR-BYOL with three different strategies to build sketch-photo pairs. The first two strategies are based on producing pseudo-sketches from each photograph of the target catalog. The pseudo-sketches are obtained by a contour-map extraction method. We try with Canny [35], a low-level approach, and PiDiNet [42], a SOTA method based on deep-learning. The third strategy does not see any photograph of the target catalog, training the model with Sketchy's pairs.

**Table 2** Retrieval results on the eCommerce dataset. The first two approaches are trained with eCommerce training photographs using pseudo-sketches produced by Canny and PiDiNet, respectively. The last row trains the model with Sketchy's pairs

| Strategy | mAP |
|---|---|
| SBIR-BYOL + Canny | 0.144 |
| SBIR-BYOL + PiDiNet | **0.190** |
| SBIR-BYOL + Sketchy | 0.178 |

The bold text is the best result in that evalution

Our results indicate that using Canny for generating pseudo-sketches from target photographs produces the worst results since it is a low-level method very sensitive to noise. The contour maps produced by Canny can be easily distorted by spurious information. However, the performance of this Canny-based model is comparable with that reported in the work of Torres and Saavedra [7].

On the other hand, using Sketchy's pairs produces better results, even though the training does not consider information from the target catalog. In the end, the best results are obtained by PiDiNet, achieving a mAP equal to 0.19. The quality of the contour maps and the target catalog information are the main ones responsible for these results.

The problem with PiDiNet is that the produced contour maps do not represent the abstraction of real sketches well. In contrast, Sketchy is a dataset built with real sketches, which is good for representing such a level of abstraction.

Therefore, we propose to merge the abstraction level provided by the Sketchy dataset with the information from a target catalog enriched with pseudo-sketches by PiDiNet. Our experiments showed that the best strategy to merge these two worlds is using a two-stage methodology, training with Sketchy's pairs in the first stage and then training with the target catalog in the second stage.

Table 3 shows the results achieved by the two-stage proposal on eCommerce dataset and Flickr25K. Here, it is important to note that in the case of Flickr25K, the model was previously trained with Flickr15K's photographs.

**Table 1** Datasets used for training and evaluation

| Dataset | Train | Evaluation |
|---|---|---|
| Sketchy | ✓ | – |
| Flickr25K | ✓ | – |
| Flickr15K | – | ✓ |
| eCommerce | ✓ | ✓ |

**Table 3** Results on the two evaluation dataset of the two-stage SBIR-BYOL. First, the model is trained with Sketchy and then with pairs produced by PiDiNet using the target catalog

| Evaluation dataset | mAP |
|---|---|
| eCommerce | 0.253 |
| Flickr15K | 0.360 |

**Table 4** Evaluation of the effect of using data augmentation during training. The SBIR-BYOL models were trained using Sketchy and evaluated with eCommerce. Each row shows the effect of using a different transformation. These transformations were applied with a probability of 0.5

| Transformation | mAP |
|---|---|
| No transforms | 0.143 |
| Rotations between -30° and 30° | 0.123 |
| Random sized crops with minimum size of 0.7 times the size original image | 0.122 |
| Random blank patches with a diameter of 0.3 times the size of the original image | 0.132 |
| Horizontal flip | 0.135 |

**Table 5** Evaluation of the effect of using data augmentation during training. The SBIR-BYOL models were trained using Sketchy and evaluated on Flickr15K dataset. Each row shows the effect of using a different transformation. These transformations were applied with a probability of 0.5

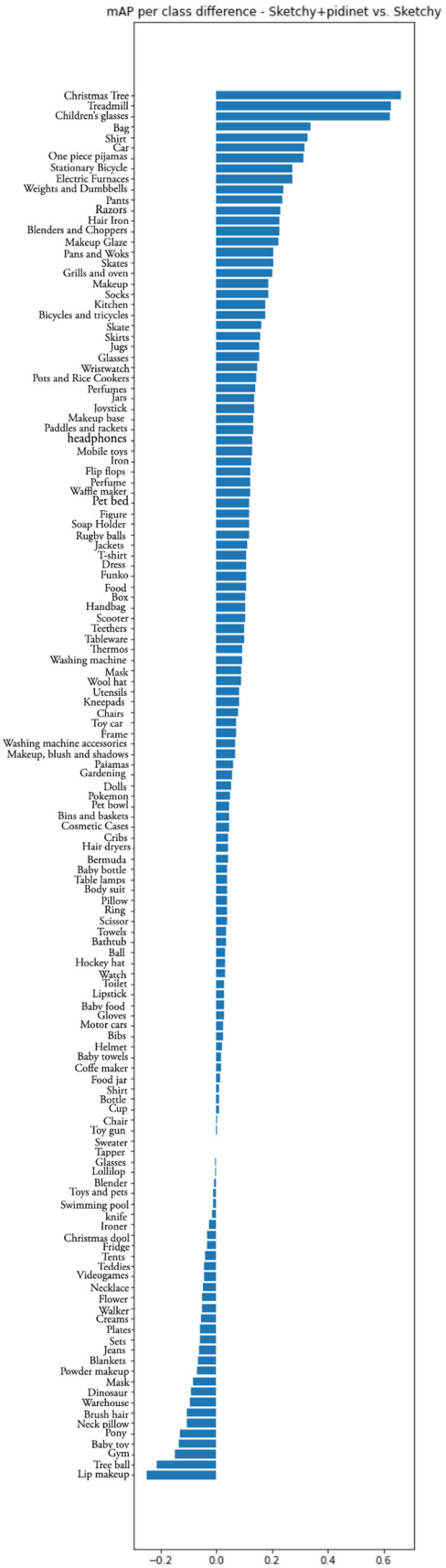
| Transformation | mAP |
|---|---|
| No transforms | 0.311 |
| Rotations between -30° and 30° | 0.241 |
| Random sized crops with minimum size of 0.7 times the size original image | 0.246 |
| Random blank patches with a diameter of 0.3 times the size of the original image | 0.289 |
| Horizontal flip | 0.291 |

**Table 6** Effect of cosine decay + EMA in our SBIR-BYOL when evaluated on the Flickr15K dataset

| Training setting | mAP |
|---|---|
| $\tau = 0.99$ without cosine decay | 0.311 |
| $\tau = 0.99$ with cosine decay | 0.351 |

**Table 7** Effect of cosine decay + EMA in our SBIR-BYOL when evaluated on the eCommerce dataset

| Training setting | mAP |
|---|---|
| $\tau = 0.99$ without cosine decay | 0.143 |
| $\tau = 0.99$ with cosine decay | 0.178 |




◀**Fig. 6** Performance gain per class, in terms of mAP, when refining our self-supervised SBIR model with PiDiNet on the eCommerce dataset

### 4.4 On the importance of transformations

SBIR-BYOL does not apply any data augmentation transformation on the inputs. We evaluated the impact of different transformations. However, none of them showed a positive impact on the final result. Tables 4 and 5 show the retrieval performance of using diverse transformations on the eCommerce and Flickr15K datasets, respectively. In both cases, the best performance was achieved without any transformation strategy.

### 4.5 Other settings

SBIR-BYOL applies the EMA (exponential moving average) strategy on the teacher side using the expression of Eq. 1.

$$\theta_t = \tau\theta_t + (1-\tau)\theta_s \tag{1}$$

where $\theta_t$ and $\theta_s$ represent the parameters of the teacher and the student branches, respectively. Here, we use $\tau = 0.99$.

**Fig. 7** Retrieval results on the eCommerce dataset with SBIR-BYOL trained only by the first stage

**Fig. 8** Retrieval results in the eCommerce dataset using our two-stage SBIR-BYOL

**Fig. 9** Retrieval results on the eCommerce dataset with SBIR-BYOL trained only with the Sketchy dataset. In this case

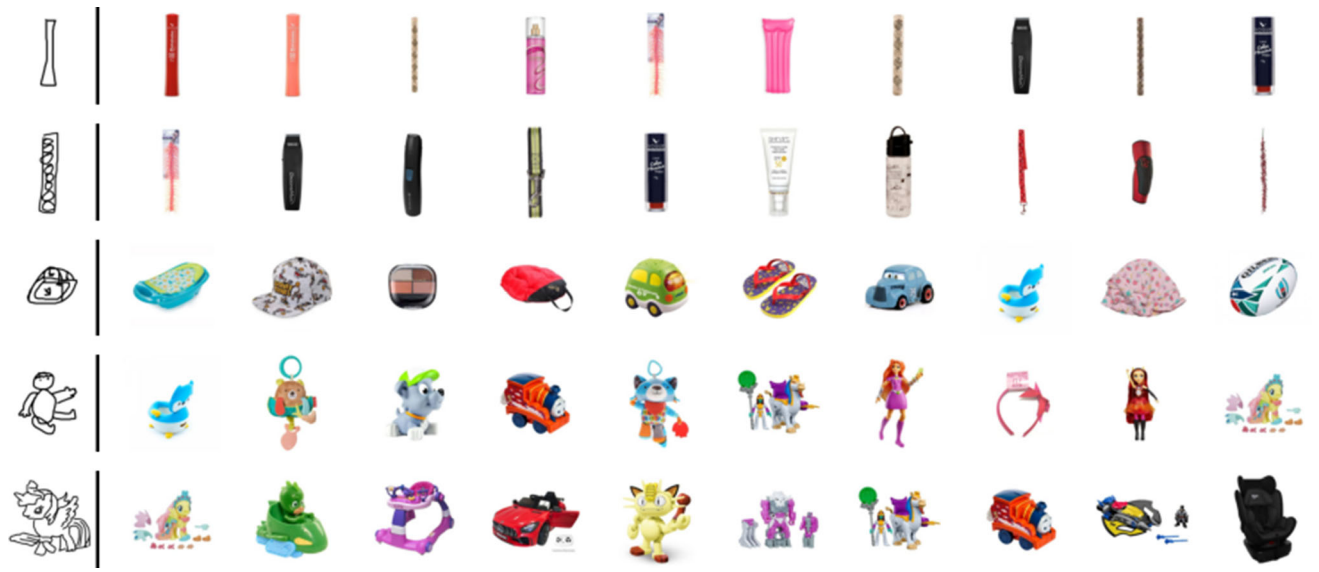


**Fig. 10** Retrieval results on the eCommerce dataset with SBIR-BYOL trained with Sketchy and then PiDiNet. The results do not show significant improvements with respect to the morel trained only with the first stage

We also use a cosine decay schedule that showed a gain up to 4% in both datasets as shown in Tables 7 and 6.

### 4.6 Per-class evaluation

In this section, we assess the performance gain of using our second training stage based on PiDiNet to leverage the target catalog's visual information. Figure 6 shows this gain per each class. The vast majority of classes improve after this second stage.

To better understand the improved performance, we present some results comparing the retrieval results after the first and second stages.

Figures 7 and 8 show results using five query sketches. The retrieval results using only the first stage are shown in Fig. 7. We observe low-semantic results. For instance, the query sketch showing *a tree* does not retrieve any relevant result; the same happens with the second query, *a tool*. The fourth query, *baggage*, returns a perfume in the first position. The model was probably confused by other objects with a similar shape.

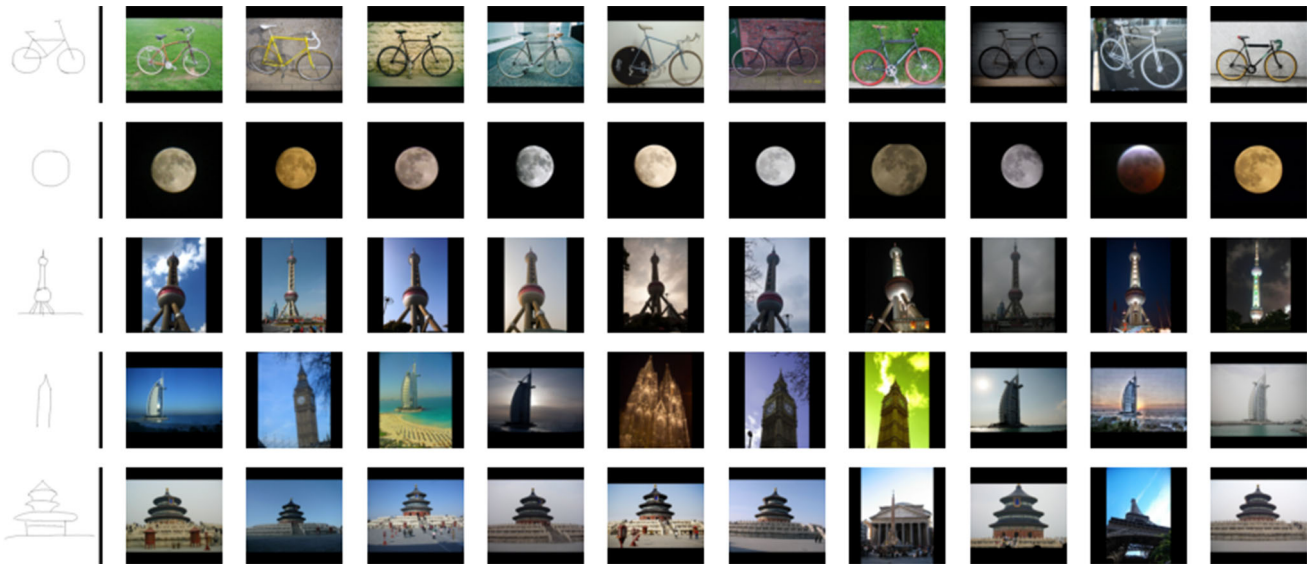


**Fig. 11** Examples of queries in the Flickr15K dataset with SBIR-BYOL trained with Sketchy's pairs

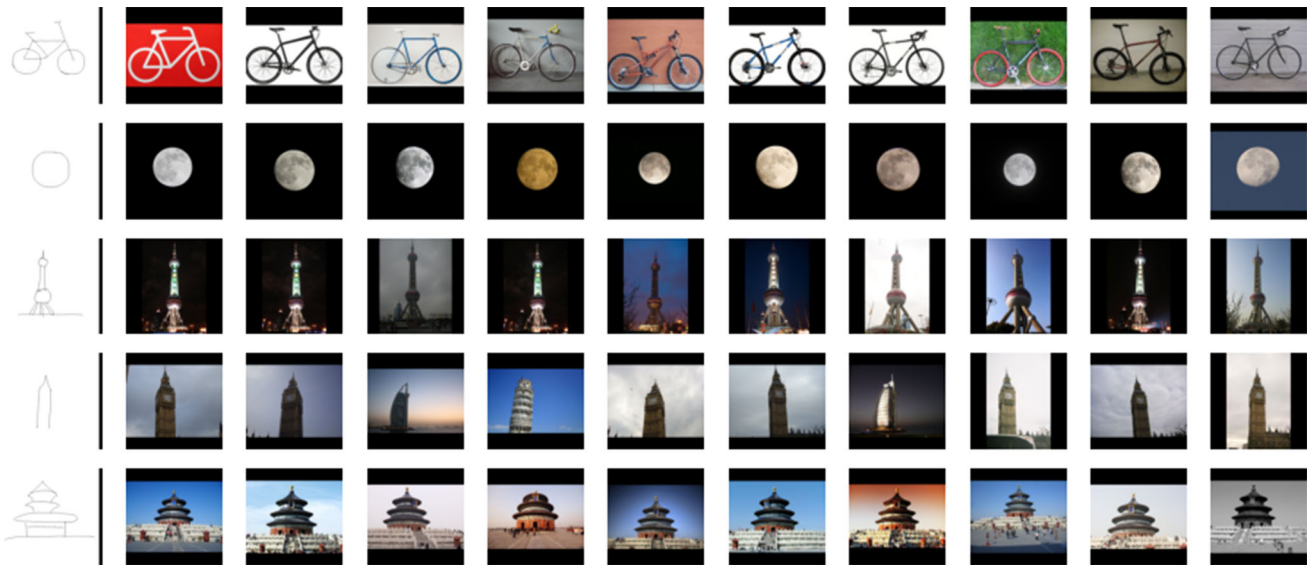


**Fig. 12** Examples of queries in the Flickr15K dataset with SBIR-BYOL trained with Sketchy and then PiDiNet edge maps

In contrast, Fig. 8 shows the retrieval results using the two-stage SBIR-BYOL that shows clear improvements. We see, for instance, that the model retrieves trees in the first query. The model retrieves tools in the first position for the second query, very close to the query's shape. In the case of the *baggage*, the model retrieves objects semantically similar to the query in the first three positions.

In our qualitative analysis, we also include cases where the two-stage model does not provide any improvement. Figures 9 and 10 show the retrieval results produced by the first stage only and the complete two-stage SBIR-BYOL, respectively. Although we do not observe significant improvement with the two-stage approach, this model still retrieves results close to the query in terms of shape.

Finally, we present results on the Flickr15K dataset. Figure 11 depicts the results of our proposal after training with the first stage only, and Fig. 12 shows the results using the two-stage training approach. We can observe that the last approach produces results closer to the query with respect to the first method. For instance, looking at the last row of both figures, we can note that our proposal produces more consistent results; that is, the resulting images are visually close to each other.

**Fig. 13** Retrieval results produced by SBIR-BYOL on a catalog of shoes. For each row, the first image is the sketch query, and then retrieved images are ordered from the most to the less similar

**Fig. 14** Example of a generated image from an input sketch. The generation model was trained with pairs produced by a sketch-based image retrieval model trained in a self-supervised manner

## 5 Beyond SBIR

We show how our approach can be useful for other sketch-based understanding tasks, such as sketch2photo translation. For instance, sketch2photo translation requires a huge amount of sketch-photo pairs. However, we address this problem using our proposed Bimodal BYOL for sketch-based image retrieval. We train this model in a self-supervised manner with a catalog of 13,000 shoe images. Figure 13 depicts example results from our SBIR-BYOL model on a shoe catalog.

We leverage this retrieval model to generate pairs between the query sketch with the K nearest images to train a conditional generative model. In our experiments, we use K = 3. The generated results are shown in Fig. 14.

Here, we showed how the proposed self-supervised SBIR-BYOL could be leveraged for other visual tasks where making sketch-photo pairs is critical.

An implementation of our SBIR-BYOL approach can be found in https://github.com/javier-op/bimodal-byol-shoes/tree/main/data.

## 6 Conclusions

This work presents the first self-supervised sketch-based image retrieval model, SBIR-BYOL, an extension of BYOL. Our proposal is a bimodal model for sketch-based image retrieval. Here, we present a self-supervised 2-stage training methodology leveraging sketch-photo pairs and contour-photo pairs. The last set of pairs is obtained directly from the target catalog. We demonstrate the impact of our model in real environments like eCommerce, where a search engine is a critical component. In this context, our results showed improvements in 67% , compared to classical models trained in a supervised manner.

Our models can be easily applied to any target catalog in diverse scenarios. In addition, we also showed how our solution is used for other sketch-based tasks like sketch2-photo translation or any other task where making sketch-photo pairs represents a bottleneck.

In future work, we will explore how self and cross-attention mechanisms can be incorporated into our SBIR-BYOL. We hope to add interpretability with attention to understand better what our model learns.

## Declarations

**Conflict of interest** The authors have no relevant financial or non-financial interests to disclose.The authors have no competing interests to declare that are relevant to the content of this article.All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this

manuscript.The authors have no financial or proprietary interests in any material discussed in this article.

## References


1. Hubel DH, Wiesel TN (2004) Brain and Visual Perception: The Story of a 25Year Collaboration, Illustrated. Oxford University Press, London
2. Walther DB, Chai B, Caddigan E, Beck DM, Fei-Fei L (2011) Simple line drawings suffice for functional mri decoding of natural scene categories. Proceed Natl Acad Sci 108(23):9661–9666
3. Yu Q, Yang Y, Liu F, Song Y-Z, Xiang T, Hospedales TM (2017) Sketch-a-net: A deep neural network that beats humans. Int J Comput Vis 122:3
4. Forbus K, Usher J, Lovett A, Lockwood K, Wetzel J (2011) Cogsketch: sketch understanding for cognitive science research and for education. Topi Cognit Sci 3(4):648–666
5. Mukherjee K, Hawkins RXD, Fan JW (2019) Communicating semantic part information in drawings. In: Goel AK, Seifert CM, Freksa C (eds.) Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada. 24-27: 2413–2419
6. Kearney KS, Hyle AE (2004) Drawing out emotions: the use of participant-produced drawings in qualitative inquiry. Qualitat Res 4(3):361–382
7. Torres P, Saavedra JM (2021) Compact and effective representations for sketch-based image retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, Virtual, June 19-25, 2021, pp. 2115–2123. IEEE
8. Yu Q, Song J, Song Y-Z, Xiang T, Hospedales TM (2021) Fine-grained instance-level sketch-based image retrieval. Int. J. Comput. Vis 129(2):484–500
9. Eitz M, Hays J, Alexa M (2012) How do humans sketch objects? ACM Trans. Graph. (Proc. SIGGRAPH) 31(4):44–14410
10. Yu Q, Yang Y, Liu F, Song Y-Z, Xiang T, Hospedales TM (2017) Sketch-a-net: A deep neural network that beats humans. Int J Comput Vis 122(3):411–425
11. Xu P, Huang Y, Yuan T, Pang K, Song Y-Z, Xiang T, Hospedales TM, Ma Z, Guo J (2018) Sketchmate: Deep hashing for million-scale human sketch retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
12. Xu P, Hospedales TM, Yin Q, Song Y-Z, Xiang T, Wang L (2022) Deep learning for free-hand sketch: A survey. IEEE Transact Patt Analy Mach Intell 1:109
13. Tripathi A, Dani RR, Mishra A, Chakraborty A (2020) Sketch-guided object localization in natural images. In: Vedaldi, A, Bischof, H, Brox, T, Frahm, J (eds) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI. Lecture Notes in Computer Science vol 12351 pp 532–547
14. Bui T, Ribeiro L, Ponti M, Collomosse J (2018) Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression. Comput Graph 71:109
15. Fuentes A, Saavedra JM (2021) Sketch-qnet: a quadruplet convnet for color sketch-based image retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, Virtual, June 19-25, 2021, pp. 2134–2141. IEEE
16. Murrugarra-Llerena N, Kovashka A (2018) Image retrieval with mixed initiative and multimodal feedback. Brit Mach Vis Confer BMVC 207:103–204
17. Murrugarra-Llerena N, Kovashka A (2021) Image retrieval with mixed initiative and multimodal feedback. Computer Vision and Image Understanding 207:103204
18. Collomosse J, McNeill G, Qian Y (2009) Storyboard sketches for content based video retrieval. pp. 245–252
19. Chen W, Hays J (2018) Sketchygan: towards diverse and realistic sketch to image synthesis. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9416–9425
20. Sangkloy P, Lu J, Fang C, Yu F, Hays J (2017) Scribbler: Controlling deep image synthesis with sketch and color. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6836–6845
21. Saavedra JM, Barrios JM (2015) Sketch based image retrieval using learned keyshapes (LKS). In: Proceedings of the British Machine Vision Conference 2015, BMVC 2015. Swansea, UK, September 7-10, 2015, pp. 164–116411
22. Hu R, Collomosse J (2013) A performance evaluation of gradient field hog descriptor for sketch based image retrieval. Comput Vis Image Understand 117(7):790–806
23. Eitz M, Hays J, Alexa M (2012) How do humans sketch objects? ACM Trans. Graph. (Proc. SIGGRAPH) 31(4):44–14410
24. Hoffmann DL, Standish CD, García-Diez M, Pettitt PB, Milton JA, Zilhão J, Alcolea-González JJ, Cantalejo-Duarte P, Collado H, de Balbín R, Lorblanchet M, Ramos-Muñoz J, Weniger G-C, Pike AWG (2018) U-th dating of carbonate crusts reveals neandertal origin of iberian cave art. Science 359(6378):912–915
25. Li Y, Xu W (2022) Using cyclegan to achieve the sketch recognition process of sketch-based modeling. In: Yuan, PF, Chai, H, Yan, C, Leach, N (eds) Proceedings of the 2021 DigitalFUTURES. Springer: London pp. 26–34
26. de Andrade V, Freire S, Baptista M, Shwartz Y (2022) Drawing as a space for social-cognitive interaction. Educat Sci 12:45
27. Fernandes MA, Wammes JD, Meade ME (2018) The surprisingly powerful influence of drawing on memory. Curr Direct Psycholog Sci 27(5):302–308
28. Ha D, Eck D (2018) A neural representation of sketch drawings. In: International Conference on Learning Representations. https://openreview.net/forum?id=Hy6GHpkCW
29. Xu P, Huang Y, Yuan T, Xiang T, Hospedales TM, Song Y-Z, Wang L (2021) On learning semantic representations for large-scale abstract sketches. IEEE Transact Circuits Syst Video Technol 31(9):3366–3379
30. Morales J, Murrugarra-Llerena N, Saavedra JM (2022) Leveraging unlabeled data for sketch based understanding. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR-SketchDL Workshop. IEEE
31. Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251
32. Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976
33. Saavedra JM (2014) Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In: 2014 IEEE International Conference on Image Processing (ICIP). pp. 2998–3002
34. Saavedra JM (2017) Rst-shelo: sketch-based image retrieval using sketch tokens and square root normalization. Multimed Tools Appl 76(1):931–951
35. Canny J (1986) A computational approach to edge detection. IEEE Transact Patt Analy Mach Intell PAMI 8(6):679–698
36. Lim JJ, Zitnick CL, Dollár P (2013) Sketch tokens: A learned mid-level representation for contour and object detection. In:

2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3158–3165
37. Saavedra JM, Bustos B (2013) Sketch-based image retrieval using keyshapes. Multimed Tools Appl 73(3):2033–2062
38. Yu Q, Liu F, Song Y, Xiang T, Hospedales TM, Loy CC (2016) Sketch me that shoe. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 799–807
39. Sangkloy P, Burnell N, Ham C, Hays J (2016) The sketchy database: Learning to retrieve badly drawn bunnies. ACM Transactions on Graphics (proceedings of SIGGRAPH)
40. McInnes L, Healy J, Saul N, Großberger L (2018) UMAP: uniform manifold approximation and projection. J Open Sour Soft 3(29):861
41. Grill J-B, Strub F, Altché F, Tallec C, Richemond P, Buchatskaya E, Doersch C, Avila Pires B, Guo Z, Gheshlaghi Azar M, Piot B, kavukcuoglu k, Munos R, Valko M, (2020) Bootstrap your own latent - a new approach to self-supervised learning. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in Neural Information Processing Systems, vol 33. Curran Associates Inc, London, pp 21271–21284
42. Su Z, Liu W, Yu Z, Hu D, Liao Q, Tian Q, Pietikäinen M, Liu L (2021) Pixel difference networks for efficient edge detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5117–5127