Mapping between Images and Conceptual Spaces: Sketch-based Image Retrieval

SounakDey

Mapping between Images and Conceptual Spaces: Sketch-based Image Retrieval

A dissertation submitted by **Sounak Dey** at Universitat Autònoma de Barcelona to fulfil the degree of **Doctor of Philosophy**.

Bellaterra, September 30, 2020

Director
Dr. Josep Lladós
Universitat Autònoma de Barcelona
Dept. Ciències de la Computació
Centre de Visió per Computador

Co-Director
Dr. Umapada Pal
Indian Statistical Institute
Computer Vision and Pattern Recognition Unit

Thesis committee
Dr. Oriol Ramos Terrades
Universitat Autònoma de Barcelona
Dept. Ciències de la Computació
Centre de Visió per Computador

Dr. Francesc Moreno Noguer
Universitat Politécnica de Catalunya
Institut de Robótica i Informática Industrial (CSIC-UPC)

Dr. Eric Anquetil
Universitaire de Beaulieu
INSA Rennes, France

International evaluators
Dr. Anjan Dutta
University of Exeter
Exeter, United Kingdom

Dr. Partha Pratim Roy
Indian Institute of Technology Roorkee
Roorkee, India

This document was typeset by the author using LaTeX 2_{ε}.

The research described in this book was carried out at the Computer Vision Center, Universitat Autònoma de Barcelona.

ISBN 978-84-121011-8-8

Printed by Ediciones Gráficas Rey, S.L.

To my parents, sister and friends…

Agradecimientos

The research work presented in this thesis is an outcome of a long process of hard work and knowledge sharing through which I got this opportunity to collaborate with a number of extraordinary people. Undertaking this project has been a truly life-changing experience and it would not have been possible without the support and guidance that I received from some very special people. I would like to express my gratitude to them all and also to the organizations and institutions who have supported me during these years.

First of all I would like to thank the Centre de Visió per Computadorand the Universitat Autònoma de Barcelonafor allowing me to pursue my doctoral research and for providing me all the facilities. I would like to thank the Ag´encia de Gestio d'Ajuts Universitats i Recerca (AGAUR) for providing me a three years doctoral study scholarship (2011 FIB 01022, 2012FI B1 00174 and 2013FI B2 00074) and a mobility scholarship (2011 BE1 00169) for doing a three months research stay abroad, without that for sure it would have not been possible.

I would like to sincerely thank my supervisor Dr. Josep Llad´os, for giving me the opportunity to do research and also for his guidance, understanding and patience during this long journey.

I would also like to thank my co-supervisor Dr. Umapada Pal, from whom I got the inspirations for pursuing Ph.D. Thank you very much for encouraging me and introducing me to such opportunities.

My special thanks to Prof. Dr. Yi-Zhe Song for having lot of fruitful discussions, meetings, comments and ideas. Thank you very much Yi-Zhe, without you my thesis would have not been completed.

Allow me this opportunity to thank Dr. Anjan Dutta of University of Exeter, United Kingdom, with whom I spent many inspiring moments. With him I learnt a lot. He also was very instrumental in giving the right direction during critical stage of my PhD. Dr. Suman Kumar Ghosh was also very kind and nice towards me. He taught me many basic skills to live this PhD life.

During my research I was fortunate enough to have several collaborations. I would like to thank Dr. Juan Ignacio Toledo, for collaborating with me.

Thank you Pau Riba, Ali Furkan Biten, Andres Mafla Delgado, Ruben Perez Tito, Dr. Joan Mas, Dr. Dimosthenis Karatzas, Dr. Oriol Ramos Terrades, Dr. Alicia Fornes, Dr. Ernest Valveny of CVC for enlightening me with the fruitful collaborations. A very special thanks to Anguelos Nicolaou for collaborating with me from the beginning of

my days in CVC. Even my first paper from CVC was also resulted from the collaboration with you.

I express my gratitude to all my friends from CVC, especially, the ones who started doing doctoral study with me: David, Sanket, Arnau. Also Xavier, Prassana, Cesar, Papa, Samuel and Manuel with whom I have shared so many talks, lunches, coffees, parties. Thank you so much, for filling the void of my family. Also a huge shout to for Anguelos, Mohhamed, Cesear, Manuel, Victor, Sergi, Arindam, Pragna and Suman for their company during different phase of my life in Barcelona.

I cordially thank to all the administrative and technical supporting personnel of CVC for being patient in front of my extensive requirements, specially, Montse, Gigi, Claire and Marc for helping me out of any office and technical work. For sure without you it would have been much much difficult.

Finally my thanks, love and respect to those people, very close to my heart, my family. Heartfelt gratitude and regards to my parents for their unequivocal support throughout, as always, for which my mere expression of thanks would never be enough. My special thanks to Lucia and Samadrita for understanding me and extending their love, care and support everyday.

Abstract

The deluge of visual content on the Internet – from user-generated content to commercial image collections - motivates intuitive new methods for searching digital image content: how can we find certain images in a database of millions? Sketch-based image retrieval (SBIR) is an emerging research topic in which a free-hand drawing can be used to visually query photographic images. SBIR is aligned to emerging trends for visual content consumption on mobile touch-screen based devices, for which gestural interactions such as sketch are a natural alternative to textual input.

This thesis presents several contributions to the literature of SBIR. First, we propose a cross-modal learning framework that maps both sketches and text into a joint embedding space invariant to depictive style, while preserving semantics. The resulting embedding enables direct comparison and search between sketches/text and images and is based upon a multi-branch convolutional neural network (CNN) trained using unique training schemes. The deeply learned embedding is shown to yield state-of-art retrieval performance on several SBIR benchmarks.

Second, we propose an approach for multi-modal image retrieval in multi-labelled images. A multi-modal deep network architecture is formulated to jointly model sketches and text as input query modalities into a common embedding space, which is then further aligned with the image feature space. Our architecture also relies on a salient object detection through a supervised LSTM-based visual attention model learned from convolutional features. Both the alignment between the queries and the image and the supervision of the attention on the images are obtained by generalizing the Hungarian Algorithm using different loss functions. This permits encoding the object-based features and its alignment with the query irrespective of the availability of the co-occurrence of different objects in the training set. We validate the performance of our approach on standard single/multi-object datasets, showing state-of-the art performance in every SBIR dataset.

Third, we investigate the problem of zero-shot sketch-based image retrieval (ZS-SBIR), where human sketches are used as queries to conduct retrieval of photos from unseen categories. We importantly advance prior arts by proposing a novel ZS-SBIR scenario that represents a firm step forward in its practical application. The new setting uniquely recognizes two important yet often neglected challenges of practical ZS-SBIR, (i) the large domain gap between amateur sketch and photo, and (ii) the necessity for moving towards large-scale retrieval. We first contribute to the community a novel ZS-SBIR dataset, QuickDraw-Extended, that consists of $330,000$ sketches and $204,000$

photos spanning across 110 categories. Highly abstract amateur human sketches are purposefully sourced to maximize the domain gap, instead of ones included in existing datasets that can often be semi-photorealistic. We then formulate a ZS-SBIR framework to jointly model sketches and photos into a common embedding space. A novel strategy to mine the mutual information among domains is specifically engineered to alleviate the domain gap. External semantic knowledge is further embedded to aid semantic transfer. We show that, rather surprisingly, retrieval performance significantly outperforms that of state-of-the-art on existing datasets that can already be achieved using a reduced version of our model. We further demonstrate the superior performance of our full model by comparing with a number of alternatives on the newly proposed dataset.

Keywords – Computer Vision, Pattern Recognition, Deep Learning, Sketch-based Image Retrieval, Zero-shot learning, Cross-modal retrieval, Multi-object Multi-modal retrieval, Hungarian Loss, QuickDraw Extended Dataset

Resum

El diluvi de contingut visual a Internet –de contingut generat per l'usuari a col·leccions d'imatges comercials- motiva nous mètodes intuïtius per cercar contingut d'imatges digitals: com podem trobar determinades imatges en una base de dades de milions? La recuperació d'imatges basada en esbossos (SBIR) és un tema de recerca emergent en què es pot utilitzar un dibuix a mà lliure per consultar visualment imatges fotogràfiques. SBIR s'alinea a les tendències emergents de consum de contingut visual en dispositius mòbils basats en pantalla tàctil, per a les quals les interaccions gestuals com el croquis són una alternativa natural a l'entrada textual.

Aquesta tesi presenta diverses contribucions a la literatura de SBIR. En primer lloc, proposem un marc d'aprenentatge entre modalitats que mapi tant esbossos com text en un espai d'inserció conjunta invariant a l'estil representatiu, conservant la semàntica. L'incrustació resultant permet la comparació directa i la cerca entre esbossos / text i imatges i es basa en una xarxa neuronal convolutional multi-branca (CNN) formada mitjançant esquemes d'entrenament únics. S'ha demostrat que l'incorporació profundament obtinguda ofereix un rendiment de recuperació d'última generació en diversos punts de referència SBIR.

En segon lloc, proposem un enfocament per a la recuperació d'imatges multimodals en imatges amb etiquetes múltiples. Es formula una arquitectura de xarxa profunda multi-modal per modelar conjuntament esbossos i text com a modalitats de consulta d'entrada en un espai d'inscripció comú, que s'alinea encara més amb l'espai de funcions d'imatge. La nostra arquitectura també es basa en una detecció d'objectes destacables mitjançant un model d'atenció visual basat en LSTM supervisat, obtingut de funcions convolutives. Tant l'alineació entre les consultes com la imatge i la supervisió de l'atenció a les imatges s'obté generalitzant l'algoritme hongarès mitjançant diferents funcions de pèrdua. Això permet codificar les funcions basades en l'objecte i la seva alineació amb la consulta independentment de la disponibilitat de la coincidència de diferents objectes del conjunt d'entrenament. Validem el rendiment del nostre enfocament en conjunts de dades d'un sol objecte o amb diversos objectes, mostrant el rendiment més modern en tots els conjunts de dades SBIR.

En tercer lloc, investiguem el problema de la recuperació d'imatges basada en esbossos de zero (ZS-SBIR), on els esbossos humans s'utilitzen com a consultes per a la recuperació de fotografies de categories no vistes. Avancem de forma important les arts prèvies proposant un nou escenari ZS-SBIR que representi un pas endavant en la seva aplicació pràctica. El nou entorn reconeix exclusivament dos importants reptes importants, però sovint descuidats, de la pràctica ZS-SBIR, (i) la gran bretxa de domini entre el dibuix i la fotografia aficionats, i (ii) la necessitat d'avançar cap a una recuperació a gran escala. Primer cop aportem a la comunitat un nou conjunt de dades ZS-SBIR, QuickDraw-Extended, que consisteix en esbossos de 330.000 dòlars i 204.000 dòlars de fotos en 110 categories. Esbossos humans amateurs altament abstractes s'obtenen intencionadament per maximitzar la bretxa de domini, en lloc dels inclosos en conjunts de dades existents que sovint poden ser semi-fotorealistes. A continuació, formulem un marc ZS-SBIR per modelar conjuntament esbossos i fotografies en un espai d'inserció comú. Una nova estratègia per extreure la informació mútua entre dominis està dissenyada específicament per pal·liar la bretxa de domini. El coneixement semàntic extern s'incorpora més per ajudar a la transferència semàntica. Mostrem que, més aviat sorprenent, el rendiment de recuperació supera significativament el de l'última generació als conjunts de dades existents que ja es poden aconseguir mitjançant una versió reduïda del nostre model. Demostrem a més el rendiment superior del nostre model complet comparant amb diverses alternatives del conjunt de dades recent proposat.

Paraules Clau – Computer Vision, Pattern Recognition, Deep Learning, Sketch-based Image Retrieval, Zero-shot learning, Cross-modal retrieval, Multi-object Multimodal retrieval, Hungarian Loss, QuickDraw Extended Dataset

Resumen

El diluvio de contenido visual en Internet, desde contenido generado por el usuario hasta colecciones de imágenes comerciales, motiva nuevos métodos intuitivos para buscar contenido de imágenes digitales: ¿cómo podemos encontrar ciertas imágenes en una base de datos de millones? La recuperación de imágenes basada en bocetos (SBIR) es un tema de investigación emergente en el que se puede usar un dibujo a mano libre para consultar visualmente imágenes fotográficas. SBIR está alineado con las tendencias emergentes para el consumo de contenido visual en dispositivos móviles con pantalla táctil, para los cuales las interacciones gestuales como el boceto son una alternativa natural a la entrada de texto.

Esta tesis presenta varias contribuciones a la literatura de SBIR. En primer lugar, proponemos un marco de aprendizaje multimodal que mapea tanto los bocetos como el texto en un espacio de incrustación conjunto invariante al estilo representativo, al tiempo que conserva la semántica. La incrustación resultante permite la comparación directa y la búsqueda entre bocetos / texto e imágenes y se basa en una red neuronal convolucional de múltiples ramas (CNN) entrenada utilizando esquemas de entrenamiento únicos. La incrustación profundamente aprendida muestra un rendimiento de recuperación de última generación en varios puntos de referencia SBIR.

En segundo lugar, proponemos un enfoque para la recuperación de imágenes multimodales en imágenes con etiquetas múltiples. Una arquitectura de red profunda multimodal está formulada para modelar conjuntamente bocetos y texto como modalidades de consulta de entrada en un espacio de incrustación común, que luego se alinea aún más con el espacio de características de la imagen. Nuestra arquitectura también se basa en una detección de objetos sobresalientes a través de un modelo de atención visual supervisado basado en LSTM aprendido de las características convolucionales. Tanto la alineación entre las consultas y la imagen como la supervisión de la atención en las imágenes se obtienen generalizando el algoritmo húngaro utilizando diferentes funciones de pérdida. Esto permite codificar las características basadas en objetos y su alineación con la consulta independientemente de la disponibilidad de la concurrencia de diferentes objetos en el conjunto de entrenamiento. Validamos el rendimiento de nuestro enfoque en conjuntos de datos estándar de objeto único / múltiple, mostrando el rendimiento más avanzado en cada conjunto de datos SBIR.

En tercer lugar, investigamos el problema de la recuperación de imágenes basadas en bocetos de disparo cero (ZS-SBIR), donde los bocetos humanos se utilizan como consultas para llevar a cabo la recuperación de fotos de categorías invisibles. Avanzamos de manera importante en las técnicas anteriores al proponer un nuevo escenario ZS-SBIR que representa un firme paso adelante en su aplicación práctica. El nuevo entorno reconoce de manera única dos desafíos importantes pero a menudo descuidados de la práctica ZS-SBIR, (i) la gran brecha de dominio entre el boceto aficionado y la foto, y (ii) la necesidad de avanzar hacia la recuperación a gran escala. Primero contribuimos a la comunidad con un nuevo conjunto de datos ZS-SBIR, QuickDraw -Extended, que consta de bocetos de 330,000 y fotos de 204,000 que abarcan 110 categorías. Los bocetos humanos aficionados altamente abstractos se obtienen a propósito para maximizar la brecha de dominio, en lugar de los incluidos en los conjuntos de datos existentes que a menudo pueden ser semi-fotorrealistas. Luego formulamos un marco ZS-SBIR para modelar conjuntamente bocetos y fotos en un espacio de incrustación común. Una estrategia novedosa para extraer la información mutua entre dominios está específicamente diseñada para aliviar la brecha de dominio. El cono -cimiento semántico externo está aún más integrado para ayudar a la transferencia semántica. Demostramos que, sorprendentemente, el rendimiento de recuperación supera significativamente el del estado de la técnica en los conjuntos de datos existentes que ya se pueden lograr utilizando una versión reducida de nuestro modelo. Además, demostramos el rendimiento superior de nuestro modelo completo comparándolo con varias alternativas en el conjunto de datos recientemente propuesto.

Palabras Clave – Computer Vision, Pattern Recognition, Deep Learning, Sketch-based Image Retrieval, Zero-shot learning, Cross-modal retrieval, Multi-object Multi-modal retrieval, Hungarian Loss, QuickDraw Extended Dataset

Contents

List of Tables

List of Figures

Chapter 1

Introduction

You can't do sketches enough. Sketch everything and keep your curiosity fresh.
– Paintings, Drawings, Watercolors; (1970), by John Singer Sargent

Sketch-based image retrieval (SBIR) has been studied since the early 1990s and has drawn more and more interest recently. To address some important questions like: What are the objectives of SBIR, and what is the general methodology of SBIR? The different perspectives can be observed, common values can be discovered and new ideas can be inspired.

1.1 Visual media upheaval and management issues

Digital imagery awash the internet recently. The rate of content generation is staggering, with Cisco forecasting that by 2021, the Internet traffic [83] will account for 82% of visual content. Millions of photos (Facebook 15m [1], Instagram 2.5m [2]), and video (YouTube 500 video hours per minute [3]) are published every hour. These figures are only set to increase as behavioural trends such as vlogging gains traction, and social media catches up with developing nations. The boom of visual content is also assisted by the growth in popularity of image capturing devices such as cameras and CCTVs. Anyone would agree that billions of pictures accessible these days make a dense sampling of the visual world.

Sadly, technology for the management of visual content has not kept pace with its generation. Still textual queries are used to search visual media predominantly by current search engines (e.g. Google). Text-based image search is often unreliable as its average accuracy is reportedly just 15% [46]. It performs poorly probably due to following reason (i) word polysemes e.g. "mouse" can be a rodent or a hand-held pointing device and (ii) more importantly, visual content cannot be searched directly. While the former is a linguistic problem and could be improved by using more detailed text queries e.g. "mouse (hand-held pointing device)" or "mouse (rodent)", the latter is more systematic. Image retrieval by matching text queries against keywords has been used by most search engines, either as a secondary meta-data stream example like user-tagged keywords or with adjacent text on web page images. This indirect search often leads to many unrelated images among the returned results. Another problem of text-based image search is the unparalleled expressive power between text and image. Whilst text efficiently conveys semantic concepts (e.g. find me an image containing a shoe) it is neither intuitive nor concise to describe appearance in this manner (e.g. find me a shoe that looks like this, or a video containing movement like this). Rather, a visual query "paints a thousand words" and can communicate such concepts efficiently. Content-based Image Retrieval (CBIR, also referred as visual search) addresses these shortcomings, enabling direct search on image content rather than tags [32, 153]. The most common form of CBIR, query by visual example (QVE), encodes image content into vectors so called descriptors or fingerprints. However, users often wish to search for an image that they do not currently have, making QVE impractical when the visual example exists only within "the mental space".

SBIR is an emerging research topic within CBIR/QVE that enables users to provide a free-hand sketch as the query to search for similar photo images. This concept unifies the advantages of both CBIR (the illustrative power of visual queries) and text search (the effortless procedure of query generation). Sketching is an innate human ability; an efficient means for abstracting and communicating ideas since pre-historic times. It has been used from ancient times to today, comes naturally to children before writing, and transcends language barriers. Sketches in the form of hieroglyphics were employed in some of the oldest language systems such as Sumerian and Egyptian hieroglyphs Figure 1.1.

1.2 Sketches as graphical communication

Sketch is an universal communication and art modality that transcends barriers to link human societies. Different from other related forms of expression such as professional sketch, forensic sketch, cartoons, technical drawing, and oil paintings, it requires no training and no special equipment. As such sketch is not bound by age, race, language, geography, or national boundaries. It can be regarded as an expression of the brain's internal representation of the world, whether perceived or imagined. Smiling faces for example, are always recognized by humans.

Sketches can convey many words, or even concepts that are hard to convey at all in words. Figure 1.1 shows several examples covering ancient and contemporary; literal and emotional; iconic and descriptive; abstract and concrete; and different media of drawing. A sketch can be illustrative, despite its highly concise and abstract nature,

Figure 1.1: Diverse sketches in human daily life.

making it useful in various scenarios such as communication and design Figure 1.2.

Figure 1.2: Sketches illustrating various communication and design.

Therefore, sketch has been widely studied in computer vision and pattern recognition [72, 79, 191, 224, 225], computer graphics [46, 160], and human computer interaction [64, 81, 97, 178] communities. In particular, early research can be traced back to the 1960s and 1970s [74, 179].

However, sketch is fundamentally different to natural photos. Sketch images provide a special data modality/domain that has both domain-unique challenges (e.g., highly sparse, abstract, artist-dependent), as well as advantages (e.g., lack of background, use of iconic representation). It is also unique in that, because its source is a dynamic 'pen' movement, free-hand sketch can be stored and processed in multiple representations. These include static pixel space (when rendered as an image), dynamic stroke coordinate space (when considered as a time series), and geometrical graph space (when considered topologically) – as discussed in Chapter 2. Thus from a pattern recognition or machine intelligence perspective, these unique traits of sketch often lead to sketch-specific model designs in order to exploit sketch specific data properties and overcome sketch specific challenges when analyzing sketches for recognition, generation, and so on.

1.3 Sketch Based Image Retrieval

Sketching can be defined as the clairvoyance for the expression of thoughts through the panorama of human history and sketch-like petroglyphs that date back to primeval times before the dawn of text [46]. In the computer vision and graphics community, sketches have gone on to see numerous applications, ranging from sketch recognition to sketch synthesis to the ever-exciting sketch-based image retrieval. For the sketch recognition domain,it effectively solves the recognition of free-hand sketches [46], professional sketches (e.g., faces) [63, 182], symbols [162] and other line-rich objects [17, 163] which can be illustrated by a sketch-like format. Sketch synthesis [111] on the other hand deals with the application of amalgamating different aesthetic styles and artistic effects for synthesizing sketches. As for SBIR, it also works to provide some alternative and correlative searching approaches.

Input sketch witnesses a very high level of abstraction and roughly approximates the holistic shape and salient local shapes of the searched object/scene. The principal challenge faced by SBIR is to retrieve those images from the carnival which have some object/scene with similar holistic shape and salient local details as the input sketch. A number of substantial articles addressing the problem of SBIR has featured since the 1990s. The given Figure 1.3 illustrates the defined SBIR system. A universal and generic framework comprising all the necessary modules of SBIR has been depicted, underlining the key strategies, techniques and solutions that helps to coin an extensive and complete overview of the methodology. SBIR works into two different categories based on whether the user is sketching the whole scene of the image or an(several) object(s) in the image. The former category is denoted as sketching-the-scene type while the latter is called sketching-the-object type. Sketch differs from images on three

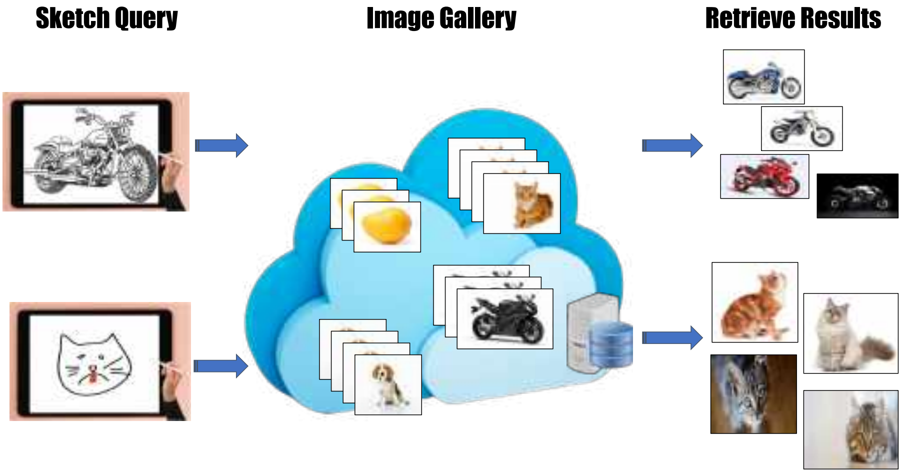


Figure 1.3: The concept of the sketch-based image retrieval system.

relevant aspects:

1. Visual cue imbalance sketches only have the holistic shape and salient local shapes (and sometimes symbolic colors), while images have plentiful details on shape, color and texture.
2. Content imbalance sketches typically do not contain any background, while images can have cluttered background.
3. The abstraction gap in a sketch even when a sketch and an edge map illustrates precisely the same object/scene, they still have predominantly different abstraction levels.

The visual cue imbalance and the abstraction gap has been observed in both categories of SBIR, while the content imbalance is mainly associated with sketching-the-object type.

1.4 Deep Learning

Deep learning approaches started gaining momentum after the success of AlexNet in the 2012 and Imagenet Large Scale Visual Recognition Challenges (ILSVRC) [96] which are still trendy till date. Deep learning aims to learn high level features from raw image pixels via an end-to-end framework. The most popular type of deep networks for visual classification and retrieval are Convolutional Neural Networks (CNN). CNN consists of multiple cascading layers with each layer's output being the input of the next one. A typical CNN layer contains a bank of linear filters (convolution) followed by a non-linear activation function (sigmoid, ReLU). Bottom layers often have additional spatial sub-sampling functions (max/average pooling) to compress the activated output, and sometimes normalisation functions (LRN, batch-norm) to prevent certain filters/neurons from dominating the others. Following the top convolution layer (fully-connected) is a loss layer that controls the learning objectives. The layer weights are learned by back-propagating the loss's gradient from top to bottom using chain rules. A CNN model often has millions of parameters e.g. 60M for AlexNet [96], 4M for GoogleNet [180] and 138M for VGG [171]. Therefore huge training data with data augmentation are explicitly required, along with weight regularisation and Dropout as implicit methods to combat over-fitting. CNN works well with stochastic gradient descent in many tasks but more advanced optimisation techniques such as RMSprop and Adam [92] are also algorithms of choice. During the training of a CNN model, the convolution layers learn activations to certain input signals that resemble hierarchical distributed representations. As shown in Deconvolution Net by Zeiler et al. [226], the bottom layers learn image-level features such as colour and edge patterns while the top layers can capture higher levels of concepts like eyes, wheel, face, etc. Thanks to its capability of encoding high level concepts, deep CNN models trained for classification tasks can be employed directly for CBIR by taking outputs of one of the top layers as image descriptors [149, 6]. More often the extracted features are refined with similarity learning; or the whole models are finetuned on new datasets for better generalisation as shown by Wan et al. [189]. Rather than taking the highest layer outputs for descriptors, Tolias et al. [184] proposed R-MAC descriptors that aggregate the max-pooled activations of multiple image regions at different scales across multiple CNN layers, leading to a significant performance boost in object retrieval and localisation tasks. Objective functions based on regression such as contrastive loss and triplet loss prove successful in achieving better image representations, as shown in [68, 196, 167, 200]. It is still unclear whether or not triplet loss performs better than contrastive loss, with Hoffer et al. [76] supporting the former while Radenovic et al. [145] backing the latter. Other authors adopted a mid-level extraction framework for CNNs, for example Gong et al. [65] proposed a multi-scale order less pooling (MOP-CNN) approach similar to spatial pyramid matching that extracts CNN features at multiple scales and aggregates using Vector of Locally Aggregated Descriptors (VLAD). There have been attempts for binarising deep features as well as unsupervised learning for CBIR. Lin et al. [113] set the first stone on deep learning of binary hash code by applying a sigmoid function after the latent layer, although this simple trick can only produce hash-like representation i.e. output values approximate {0,1} and no additional constraint is set to regularise the hashes. As the results, the hash-like representations was only used for coarse-retrieval which is followed by a refining search using outputs of a previous layer as the secondary descriptors. [113] went further on completely unsupervised deep hashing that utilises only label-less objective functions such as minimal quantisation loss (make CNN outputs close to binaries), even distributed code (make '1s' and '0s' in the binary codes evenly distributed) and rotation invariant (bring representations of rotation variants of the same images together). Gao et al. [62] and Liu et al. [116] integrated CNNs with binary hashing in an end-to-end framework for multimedia retrieval. Lin et al. Doersch et al. [41] studied an alternative unsupervised approach using spatial context to train a model predicting relative positions of two patches in an image.

Overall, approaches based on deep CNNs have become favourable in recent years, significantly outperforming traditional approaches in most CBIR tasks. A drawback of deep CNNs, as opposed to shallow-learning approaches, is that it requires large amount of training data. A part of this thesis is dedicated to find out the effects of training data size on performance of a retrieval system.

1.5 Scope and Research Questions

SBIR inherits all the classical challenges from CBIR, in addition to the challenges of cross-domain matching as well as unique problems such as ambiguity arising from the medium of sketch itself. This gives rise to several core research questions addressed by this thesis:

Research Question 1: Can we develop and efficiently implement deep neural network based models for discriminative and compact representation in sketch-based image retrieval datasets?

Objective: A unified framework for cross-modal image retrieval that can perform both sketch to image or text to image retrieval based on learning a common embedding between text and images and between sketches and images.

Contribution: A retrieval framework for images having multiple salient objects that leverages an attention model to select the subset of image features relevant to each query object.

Our major focus to address this question is the development of novel SBIR algorithms that exploit machine learning to create search embeddings. The ability to distil the information in a sketch, or image, to a point in a metric search embedding is fundamental to scalability, since identifying the nearest neighbours to a query in such a space (ranking these on distance using a metric such as L2 norm) underpins most contemporary search algorithms. For visual search to be practical, it must operate over diverse object classes; a user sketching a particular object expects to see similar objects return. This requirement fits well under class-level SBIR which aims to match sketches and images at category level i.e. a returned image is considered as relevant if it belongs to the same category as the query sketch's. Discriminative representations underpin accurate visual search, since similar sketch- image and image-image pairs are brought together (and conversely dissimilar pairs separated) in a discriminative search embedding. Compact representations are often achieved in scalable search pipelines via dimensionality reduction. In a traditional shallow SBIR framework the two processes are independent to each other with feature extraction being followed by dimensionality reduction (if available). In a deep network it is usual to sample activations from late fully connected (FC, or dense) layers, as a high dimensional search embedding — and dimensionality reduction is often omitted. The embedding must retain local visual similarity between samples i.e. similar images are close in the embedding space, yet being invariant to depiction i.e. sketches and images of the same conceptual structure are close to each other regardless of its visual appearance. Use text as an additional or complementary input modality. It should also be able to deal with multi-object scenarios. Our main contribution to address this question is a novel neural network model for learning cross-modal deep embeddings for multi-object image retrieval using both text and sketch. They improve upon earlier work in the community on so-called GN-Triplet [160].

Research Question 2: Can neural networks be utilized for multi object image search using query comprised of multiple sketch and text instances?

Objective: To create a common semantic space among text and sketches, obtained through word2vec representation of both input modalities, and aligned with the image feature space.

Contribution: A visual attention model that automatically detects salient objects from an image, that is trained in a supervised way in order to minimize the assignment cost between attention output and object bounding boxes.

To resolve ambiguity in sketch via incorporating multiple modalities within the query. In two independent studies we tackle fusion of sketched shape with two separate modalities that have not been explored in previous SBIR work. To effectively search an large image space the query should be flexible enough to comprise of either sketch or text or both. Allowing to express queries that can refer to multiple objects. Differences in drawing styles also lead to great variance in the level of abstraction – some users like drawing very detailed objects while others tend to draw only silhouette contours, causing the resulted sketch to match multiple objects. We introduce two extensions to the previous model to address this question in Chapter 4. The model provides more expressiveness to the search language since users can construct queries consisting of different concepts that are aligned to the salient objects of the target images.

Research Question 3: Can deep neural networks retrieve images for sketch classes it has never seen before in large scale scenario?

Objective: To do a zero shot sketch based image retrieval in a real-world scenario.

Contribution: A novel cross-domain zero-shot embedding model that mines the mutual information among domains that is specifically engineered to alleviate the domain gap.

The sketch ambiguity. Although sketching is an intuitive and convenient means of describing visual content, an average sketcher is not a faithful artist. Sketches are usually messy and non-linearly deformed due to either limitation in users' drawing skill, or the nature of the casual throwaway act of a posing a search query limiting the time a user spends on generating a sketch. Other forms of sketch deformation include caricature (certain features are exaggerated e.g. a cat has long whiskers), anthropomorphism (animals/static objects have human traits e.g. smiling spider) and simplification (e.g. limbs and legs are represented as single strokes). To design a model which can tolerate such deformations yet still being able to capture subtle, fine-grain details in a sketched query. SBIR is still an emerging technology, and not as prevalent as photographic visuals search. While photo datasets can reach millions in size and thousands of categories, it is costly to create a high quality and diverse collection of example sketches. This limitation particularly affects the training of SBIR models that involves machine learning, such as the algorithms contributed in this thesis. It is interesting to explore how the training sketch volume influences the retrieval performance, and explore how best to augment or otherwise build successful training methodologies for SBIR given the limitations of existing sketch datasets. We introduce the Doodle to Search (D2S) model in Chapter 5.

Research Question 4: How does the synthetic generated sketches boost the existing sketch based image retrieval methods?

Objective: Synthetically generated sketches are good enough to remove the necessity of having to acquire multiple sketches of different type.

Contribution: We have a started to a preliminary work to understand more how human-level concept learning for sketching.

To address this question, we start a preliminary work on human concept learning of sketches which helps to create a synthetic dataset that further allow us to do study on the data hungry deep learning algorithms of SBIR.

1.5.1 New datasets

As an additional contribution, this thesis contributes two new datasets to the SBIR literature. The first, **modified MS-COCO**, We use the MS-COCO dataset for constructing a database of images containing multiple objects. As the label number for each image also varies considerably, rendering MS-COCO is even more challenging. We use the class names of the Sketchy dataset and take all possible combinations by taking two, three, four, five class names. Afterwards, we download the images belonging to these combined classes, and use them for training and retrieval (used in Chapters 4). The second one, **QuickDraw-Extended**, which is also novel large-scale dataset especially made for zero shot sketch based image retrieval (used in Chapter 5).

1.6 Thesis Structure/ Outline

The thesis structure is outlined as follows, along with a brief description of contributions of each:

Chapter 2 – Literature review

In this Chapter we present a comprehensive survey of existing work in the field of SBIR. We identify common trends in the field spanning three different ages of SBIR algorithm, from classical optimization based approaches to shallow-learning to the most recent deep learning based approaches.

Chapter 3 – Learning Cross-Modal Deep Embeddings for Multi-Object Image Retrieval using Text and Sketch

In this work we introduce a cross-modal image retrieval system that allows both text and sketch as input modalities for the query. A cross-modal deep network architecture is formulated to jointly model the sketch and text input modalities as well as the image output modality, learning a common embedding between text and images, and between sketches and images. In addition, an attention model is used to selectively focus the attention on the different objects of the image, allowing for retrieval with multiple objects in the query. Experiments show that the proposed method performs the best in both single and multiple object image retrieval in standard datasets.

Chapter 4 – A Multi-modal and Multi-object Image Retrieval Framework

In this chapter we propose an approach for multi-modal image retrieval in multi-labelled images. A multi-modal deep network architecture is formulated to jointly model sketches and text as input query modalities into a common embedding space, which is then further aligned with the image feature space. Our architecture also relies on a salient object detection through a supervised LSTM-based visual attention model learned from convolutional features. Both the alignment between the queries and the image and the supervision of the attention on the images are obtained by generalizing the Hungarian Algorithm using different loss functions. This permits encoding the object-based features and its alignment with the query irrespective of the availability of the co-occurrence of different objects in the training set. We validate the performance of our approach on standard single/multi-object datasets, showing state-of-the art performance in every dataset.

Chapter 5 – Doodle to Search: Practical Zero-Shot Sketch-based Image Retrieval

In this chapter, we investigate the problem of zero-shot sketch-based image retrieval (ZS-SBIR), where human sketches are used as queries to conduct retrieval of photos from unseen categories. We importantly advance prior arts by proposing a novel ZS-SBIR scenario that represents a firm step forward in its practical application. The new setting uniquely recognizes two important yet often neglected challenges of practical ZS-SBIR, (i) the large domain gap between amateur sketch and photo, and (ii) the necessity for moving towards large-scale retrieval. We first contribute to the community a novel ZS-SBIR dataset, QuickDraw-Extended, that consists of 330,000 sketches and 204,000 photos spanning across 110 categories. Highly abstract amateur human sketches are purposefully sourced to maximize the domain gap, instead of ones included in existing datasets that can often be semi-photo realistic. We then formulate a ZS-SBIR framework to jointly model sketches and photos into a common embedding space. A novel strategy to mine the mutual information among domains is specifically engineered to alleviate the domain gap. External semantic knowledge is further embedded to aid semantic transfer. We show that, rather surprisingly, retrieval performance significantly outperforms that of state-of-the-art on existing datasets that can already be achieved using a reduced version of our model. We further demonstrate the superior performance of our full model by comparing with a number of alternatives on the newly proposed dataset.

Chapter 6 – Conclusion

We summarise the contributions of this thesis including our key findings and identify ongoing research questions for future study.

Chapter 2

Related Works

The past is behind, learn from it. The future is ahead, prepare for it.
The present is here, live it.
– In Search of Treasure, by Thomas S. Monson

Sketch based image retrieval has been an important research area of Computer Vision and a large variety of approaches in this field are documented in the literature. The contributions of this thesis are within the fields of Information Retrieval (IR), Computer Vision (CV) and Machine Learning (ML), particularly into the sub-fields of Content Based Image Retrieval(CBIR) and Sketch Based Image Retrieval(SBIR). In this chapter we present a comprehensive review of techniques used in SBIR with primary focus on sketch-related algorithms. This Chapter begins with a review of free-hand sketches and its unique challenges, followed by a more specific focus on SBIR. The review then tackles other sketch-related work on zero shot learning. Finally, a small survey on existing commercial sketch applications available till date.

Sketch research and applications in both industry and the scholarly world have exploded as of late because of the pervasiveness of touchscreen gadgets (e.g., cell phone, tablet) that make getting sketch information a lot simpler than any time in recent memory as well as the quick advancement of profound learning procedures that are accomplishing state-of-the-art performance in diverse artificial intelligence tasks.

This boom has occurred on several fronts:

- Some classic research topics (e.g., sketch recognition, sketch-based image retrieval, sketch-based 3D shape retrieval) have been re-studied in deep learning contexts [151, 160, 191, 220, 225, 226] resulting in significant performance improvements.

- Some brand-new topics have been proposed based on deep learning, e.g., deep learning based sketch generation/synthesis [72], sketch-based model generation [77], reinforcement learning based sketch abstraction [151], adversarial sketch based image editing [142], graph neural network based sketch recognition [214], graph convolution-based sketch semantic segmentation [217], and sketch based software prototyping [178].

- Beyond global representation based tasks (e.g., sketch recognition), and more fine-grained tasks have been further studied or proposed, e.g., instance-level sketch-based image retrieval [224], and deep stroke-level sketch segmentation [144].

- Compared with the conventional approach of representing sketches as static images [79], the trends of touchscreen acquisition and deep learning have underpinned progress on designing deep network architectures to exploit richer representations of sketch. Thanks to works such as SketchRNN [72], the sequential nature of free-hand sketches is now widely modeled by recurrent neural network (RNN).

- More sketch based applications have appeared, i.e., the online sketch game QuickDraw [72], sketch-based commodity search engine [175, 224].

- Some large-scale sketch datasets have been collected, e.g., Sketchy [160] and Google QuickDraw [72] – a million-scale sketch dataset (50M+).

Table 2.1: Notation definitions and abbreviated terms

Notations	Descriptions
$\mathbf{M},\mathbf{M}^T$	matrix **M** and its transpose
$\mathscr{X} = \{\mathbf{X}_n\}_{n=1}^N$	sketch sample set
$\mathbf{X}_m,\mathbf{X}_n$	m -th and n -th sketch samples in the sketch sample set $\mathscr{X}$
$\mathscr{Y} = \{y_n\}_{n=1}^N$	associated label set of $\mathscr{X}$
y_n	label of $\mathbf{X}_n$
$\mathscr{L}$	loss function
Θ	learnable parameters of neural network
$\mathscr{F}(\cdot)$	function mapping or feature extraction
$\mathscr{F}_\Theta(\cdot)$	neural network feature extraction, parameterized by Θ
$\mathscr{D}(\cdot,\cdot)$	distance metric, $e \cdot g \cdot, \ell_2$ distance
λ	weighting factor
$\sum$	summation
α,β,γ	hyper parameters by manually setting

Abbreviated Terms	Descriptions
CNN	convolutional neural network
GNN	graph neural network
RNN	recurrent neural network
TCN	temporal convolutional neural network

2.1 Notation

This section, in table 2.1, provides a reference for the most commonly used notation across this thesis. Individual chapters introduce additional notation where necessary.

2.2 Intrinsic Traits and Sketch Domain-Unique Challenges

2.2.1 Unique Challenges and Opportunities

The unique challenges of free-hand sketch based computer vision problem can be summarized as follows:

1. Highly Abstract: Humans use sketch to depict an object or event in very few strokes, reflecting the high-level semantics of a mental image. As shown in Figure 2.1, a pyramid can be depicted as a triangle in sketch, and several strokes can depict a fancy handbag.
2. Highly Diverse: Different persons have different drawing styles. For example, a near 'realistic' (close to photo edge-map) sketch image could be portrayed in different ways as exaggerated (c.f. caricatures), iconic (where details are omitted and the sketch is near symbolic), or artistic. Depending on subjective opinion about salience, different parts may also be included or omitted in a sketch. For instance, given a "dog", people differ on choice of drawing with/without body (see Figure 2.1). Finally, there is the mental viewpoint of different users, e.g., whether they imagine an orthographic or perspective projection image (Figure 2.2). In Figure 2.1, we can see that different persons draw differing perspective views of an identical slipper.
3. Highly Sparse: No matter the representation, free-hand sketch is a highly sparse signal compared to photographs.
4. Finally, there are some unique challenges when collecting sketch, which will be discussed in detail in following
    i) Sketch Collection Strategies
    ii) Sketch Collection Challenges
        a) Time-Sequence Nature
        b) Cross-Modal Pairing
        c) Demographic Information

Sketch also provides some unique opportunities compared to photos:

1. As a counterpoint to the sparsity challenge, sketch often lacks distracting background clutter compared to photos, which can benefit automated analysis [225].

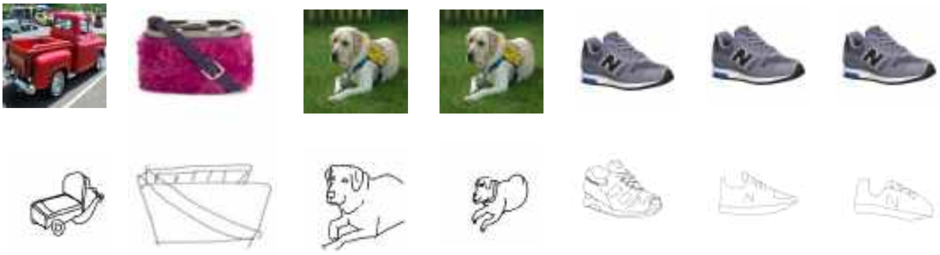


Figure 2.1: Illustrations of the major domain-unique challenges of free-hand sketch. Each column is a photo-sketch pair. Sketch is highly abstract. A truck can be depicted as a wierd rectangle in sketch, and a few strokes depict a fancy handbag. Sketch is highly diverse. Different persons draw distinctive sketches when given the identical reference, due to subjective salience (head vs. body), and drawing style.

2. If captured appropriately, the sequential nature of sketch generation can further be exploited to benefit analysis compared to static images [165].

3. The sparse and sequential nature of sketch also provides opportunities for high quality sketch generation, where image generation is hampered by the need to fill in pixel detail [72, 173].

4. Sketches can serve as a computer-interaction modality in a way that photos cannot [72, 165], due to the intuitive way humans can generate them without training.

Given these unique challenges and opportunities, it is often beneficial to design sketch-specific models to obtain best performance in various sketch-related applications.

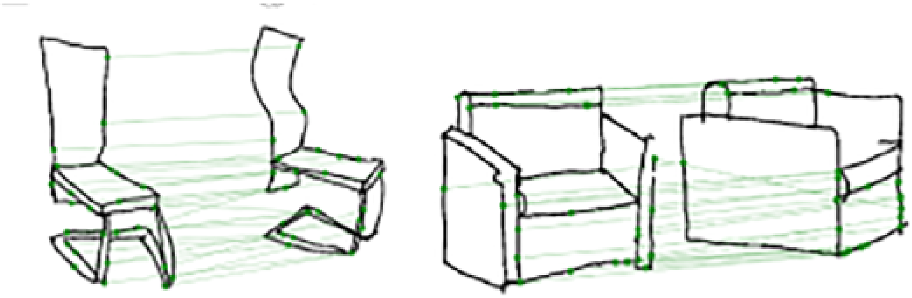
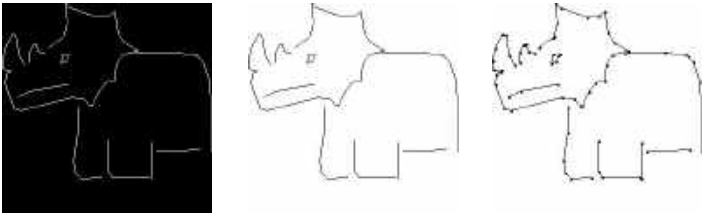


Figure 2.2: Multi-view sketch pairs selected from [221]. Different users can imagine and draw the same object from different perspectives.

2.2.2 Representation

Free-hand sketch is a special kind of visual data, intrinsically different to natural photos that are the pixel-perfect copies of the real world. For efficient storage and fast calculation, free-hand sketch can be saved as a sparse matrix, or as a black and white image that ignores its sparsity. Since sketch generation is a dynamic process, suitably captured sketches can also be represented as a sequence of strokes or pen coordinates (Figure 2.3, right images), which are represented in Euclidean space. In this regard, sketches share similarities with hand-written characters, yet are fundamentally different for their highly abstract and free-style nature (hand-writings are subject to specific rules and learning process). From another perspective, free-hand sketches can also be modeled as a sparsely connected graph where lines are the edges in the graph. Compared with a sequence of euclidean coordinates, topological representation as a graph can provide a more flexible and abstract representation. As a result of this diversity of possible representations, various deep learning paradigms can be used to process sketches including CNNs, RNNs, GCNs, and TCNs.



```
[  // First stroke
   [x0, x1, x2, x3, ...],
   [y0, y1, y2, y3, ...],
   [t0, t1, t2, t3, ...]
 ],
 [  // Second stroke
   [x0, x1, x2, x3, ...],
   [y0, y1, y2, y3, ...],
   [t0, t1, t2, t3, ...]
 ],
 ... // Additional
strokes
```

Figure 2.3: Sketch-specific representations. Representations from left to right: sparse matrix (black background with white lines), dense picture (white background with black lines), graph, stroke sequence. Both graph and stroke sequence representations are based on the key stroke points. In stroke sequence, ’x’ and ’y’ are the pixel coordinates, and ’t’ is the time in milliseconds since the first point. ’x’ and ’y’ are real-valued while ’t’ is an integer.

2.3 A Brief History of SBIR in the Deep Learning Era

In the past five years, the free-hand sketch community has developed rapidly as summarized by Figure 2.4 from the perspectives of: tasks, datasets, representations and supervision.

In 2015, Sketch-a-Net [222] was proposed as the first CNN engineered specifically for free-hand sketch. It gained note as the first to achieve a recognition rate surpassing humans and helped to popularize deep learning for sketch analysis.

In 2016, three fine-grained sketch-based image retrieval (FG-SBIR) datasets were released, i.e., QMUL Shoe and Chair [226], and Sketchy [160]. Combined with deep triplet ranking [46], these fine-grained cross- modal datasets motivated a wave of followup FG-SBIR and other fine-grained tasks.

In 2017, Google released a million-scale sketch dataset, i.e., Google QuickDraw, via the online game “QuickDraw”. QuickDraw contains over 50M sketches collected from players around the global world, making it a rich and diverse dataset. Furthermore, based on the QuickDraw dataset, [72]. proposed “SketchRNN”, a RNN-based deep Variational AutoEncoder (VAE) that can generate diverse sketches. This work motivated the community to go beyond considering sketches as static pictures to be processed by CNN; and inspired subsequent work to use stroke sequences as input and study temporal processing of sketches. In 2017, some sketch-based deep generative image models [161] began to appear in the top conferences in computer vision.

From 2018 to date, based on deep learning techniques, various novel methodologies – e.g., sketch hashing [213], sketch transformers [214]; and applications – e.g., sketch abstraction [151], sketch-based photo classifier generation [77], sketch perceptual grouping [107] have been proposed. See Figure 2.4 for a chronological summary.

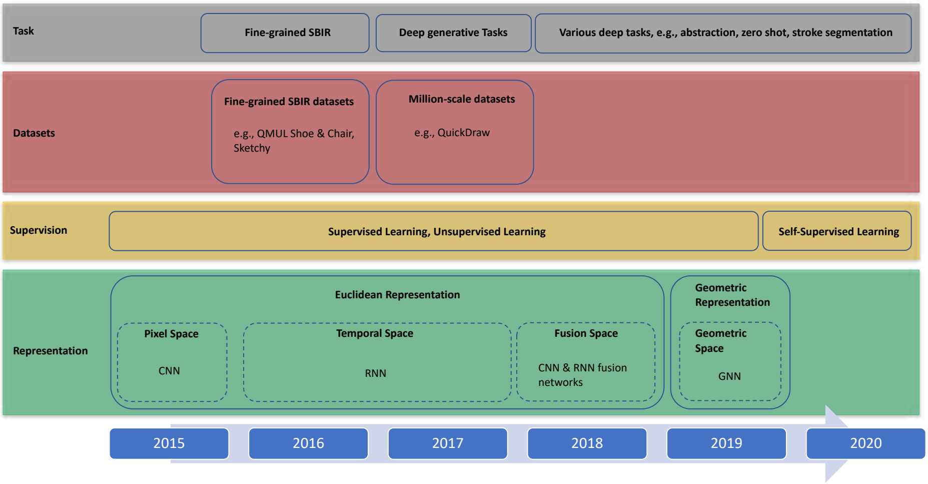


Figure 2.4: Milestones of deep learning based sketch research, from the perspectives of task, datasets, supervision, and representation

2.4 Sketch Datasets

Free-hand sketch datasets can be grouped in terms of: (i) single vs. multi-modal, and (ii) coarse vs. fine-grained. Single-modal datasets consist only of sketches and are typically used for recognition, sketch-sketch retrieval, grouping, segmentation and generation. Multi-modal datasets support cross-modal tasks by providing sketches paired with samples of other modalities such as natural photo, 3D shape, text, or video. These are mainly used for cross-modal retrieval/matching, or cross-modal generation/synthesis. Coarse grained datasets (e.g., TU-Berlin [45], QuickDraw [72]) are usually used for sketch recognition, sketch retrieval; while fine-grained datasets (e.g., QMUL Shoe [224]) provide fine-grained visual details and manual annotations.

More specifically, coarse-grained single-modal datasets [72, 45] support sketch recognition and retrieval; while coarse-grained multi-modal datasets (e.g., QuickDraw- Extended [40]) support category-level sketch-based image retrieval. Fine-grained single-modal datasets [144, 107] support perceptual grouping, segmentation, and parsing. Fined-grained multi-modal datasets (e.g., QMUL Shoe [224]) provide the instance-level pairing information to support retrieval. Table 2.2 summarizes representative sketch datasets of each type in terms of: modalities, size, number of categories, stroke information, annotation, etc. Note that SVG files are able to generate picture files such as JPEG and PNG, but not vice versa. We exclude some well-known but overly-small datasets such as [91]. Datasets such as [37] is also omitted due to not much similar dataset to compare against. Table 2.2 also mention about Sketch&UI [82] where 'UI' stands for user interface.

Table 2.2: Summary of the representative sketch datasets. "✓" denotes "yes/ available/ provided". Both 'grouping" and "segmentation" annotations refer to stroke-level. "K" and "M" mean "thousand" and "million", respectively. "Cat." means "category". Stroke "✓" denotes sketches provided as SVG files or coordinate arrays.

Single-Modal Datasets	Fine-Grained	Public	Sample Amount	Cat.	Stroke	Object/ Scene	Instance Pairing	Annotations	Remarks
Tu-Berlin [45]		✓	20K sketches	250	✓	o	-	class	
QuickDraw [72]		✓	50M+ sketches	345	✓	o	-	class	
QuickDraw-5-step [26]			38M+ sketches	345		o	-	class	
SPG [107]	✓	✓	20K sketches		✓	o	-	class, grouping	
SketchSeg-150K [144]	✓		150K sketches		✓	o	-	class, segmentation	57 semantic labels
SketchSeg-10K [194]	✓	✓	10K sketches			o	-	class, segmentation	24 semantic labels
SketchFix-160 [164]		✓	3904 sketches		✓	o	-	class, eye fixation	
Multi-Modal Datasets	Fine-Grained	Public	Modalities & Sample Amount	Cat.	Stroke	Object/ Scene	Instance Pairing	Annotations	Remarks
QMUL Shoe [224]	✓	✓	419 sketches, 419 photos	1		o	✓	pairing, triplet, attribute	21 binary attributes
QMUL Chair [224]	✓	✓	297 sketches, 297 photos	1		o	✓	pairing, triplet, attribute	15 binary attributes
QMUL Handbag [174]	✓	✓	568 sketches, 568 photos	1		o	✓	pairing	
Sketchy [160]	✓	✓	75K sketches, 12K photos	125	✓	o		class	12K objects
Sketch & UI [82]	✓		1998 sketches, 1998 photos	23		o	✓	class, pairing	UI
QuickDraw-Extended [40]		✓	330K sketches, 204K photos	110		o		class	
Sketch Transfer [99]			112.5K sketches, 90K CIFAR-10 photos	9		o		class	resolution of 32 × 32
TU-Berlin Extended [227]			20K sketches, 191K photos	250		o		class	
Sketch Flickr 15K [79]		✓	330 sketches, 15K photos	33		o		class	
Aerial-SI [87, 87]			400 sketches, 3.3K photos	10		o, s		class	aerial scene
HUST-SI [202]		✓	20K sketches, 31K photos	250	✓	o		class	
SBSR [52]		✓	1814 sketches, 1814 3D models	161		o		class	
SHREC'13 [104]	✓	✓	7200 sketches, 1258 3D models	90		o		class	
SHREC'14 [105]	✓	✓	12680 sketches, 8987 3D models	171		o		class	
PACS DG [116]		✓	9991 (sketches, photos, cartoons, paintings)	7		o		class	domain generalization
Flickr1M [208]			500 sketches, 1.3M photos	100		o		class	
Cross-Modal Places [21]		✓	16K sketches, 11K descriptions, 458K spatial texts, 12K clip arts, 1.5M photos	205		s		class	
DomainNet [140]		✓	0.6M (cliparts, infographs, paintings, QuickDraw sketches, real photos, professional pencil sketches)	345	✓	o		class	

2.5 Different tasks and categorisation of methodology

According to the data modalities involved, free-hand sketch related tasks can be divided into single- and multi-modal tasks, where single-modality sketch analysis techniques are often used as building blocks for multi-modal methods. This section will define the popular sketch analysis tasks and introduce the corresponding deep learning methods, providing a detailed taxonomy. Figure 2.5 provides a mind-map diagram of the existing free-hand sketch tasks. Single-modality tasks study sketches in isolation without other data modalities. Key deep learning-based applications in this area include recognition, retrieval/hashing, generation, grouping, segmentation, and abstraction. Just keeping in mind the theme of the thesis, we will just restrict our discussion to a few multi-modal tasks those are particularly important.

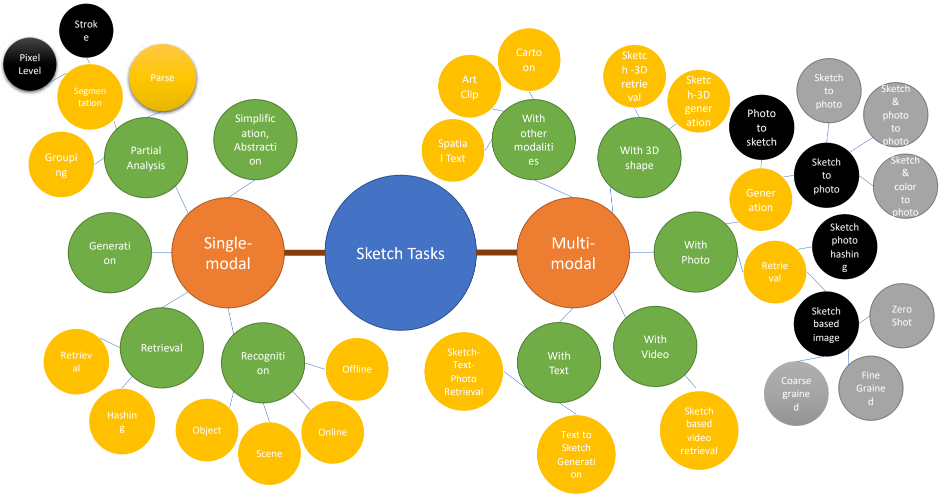


Figure 2.5: A mind map diagram of the sketch task taxonomy

2.5.1 Mutlti-Modal

Free-hand sketch has several cross-modal applications when paired with other data modalities. In this section we review sketch-related cross-modal topics including visual (e.g., natural photo, 3D shape, video) and textual domains. Nowadays, most visual retrieval approaches work under the “query-by-example” (QBE) [181] setting where users provide examples of the content that they seek. Compared with other query modalities (e.g., photo, video, text), sketch has several unique advantages: In some scenarios users do not know the name of the object that they seek, or find it hard to describe (such as fine-grained details of a fashion item) in order to query-by-text. Meanwhile, it may be difficult or impractical to provide photos or video examples of the object that they seek. Sketch-based image retrieval provides a query modality where users express their target object by rendering their mental image in sketch. It is particularly useful when searching at the fine-grained instance-level. Thus sketch can be used as a modality to retrieve natural photo, manga [122], 3D shape, video, etc.

Sketch based Image Retrieval

SBIR is also known as sketch-photo retrieval [79, 224, 215, 213]. SBIR is challenging for all the reasons that sketch-analysis in general is challenging (sparse and abstract input). It is particularly taxing because of the difficulty of comparing sparse line drawings with dense pixel representations, especially when the input could be a very abstract, or iconic (symbolic) representation that is hard to compare directly to accurate perspective projection photos.

Figure 2.5 includes a taxonomy for SBIR. From the perspective of evaluation criterion, SBIR can be divided into conventional/coarse-grained SBIR (i.e., category-level SBIR), mid-grained [16], and fine-grained SBIR (i.e., instance-level SBIR). FG-SBIR is essentially a kind of instance-level retrieval [231]. From the perspective of retrieval embedding space, SBIR can be further divided into common nearest-neighbor and fast hashing-based retrieval. From the perspective of supervision involved in training, SBIR can be divided into fully-supervised and zero-shot retrieval. We will go into the details of the topics that are in line with the thesis.

Category and Instance Level SBIR

In coarse-grained SBIR, given a target sketch as query, a ranking list is returned based on the similarity (e.g., Euclidean or Hamming distance). The retrieval is judged as correct, if the photo ranked at the top has the identical class label as the query. However, in fine-grained SBIR, the retrieval is judged as correct only when the returned photo is from the same instance pair as the query sketch. Based on SBIR ideas, several sketch-based commodity search engines have been implemented, e.g., sketch-based skirt image retrieval [94], fine-grained sketch-based shoe [224, 176], chair [224, 176], and handbag [176] retrieval systems.

Some previous SBIR works [8, 103] have used edge-maps (image contours) of photos as an approximation to corresponding sketch images in order to perform matching. Canny edge detector [18], Edge Boxes toolbox [236], and holistically nested edge detection (HED) [209] were usually used to extract the edges from natural photos. However, this kind of hand-designed process is now commonly replaced by end-to-end deep feature learning. Table 2.3 shows different type of feature used in SBIR tasks. The table lists the type of feature employed by each work. For the histograms type and deep features type, we also list the exact feature(s)/network architectures for comparison. For the other two types, we do not list the exact details, as the features in them are generally pixels or lines. Through the table, we can observe that it started from the pixel-based features and turned majorly into histogram features, with only a pair of works employing contour-based features. Then the trend has been to favor the deep features. For the deep features, current studies have generally focused on utilizing or combining existing advanced deep architectures.

Table 2.3: The summary of the types of features employed for SBIR

Year	Author	Pixel-based	Contour-based	Histograms	Deep features
1992	Hirata and Kato [75, 91]	✓			
1994	Faloutsos et al. [54, 129]	✓			
1997	Del Bimbo and Pala [33, 34, 35]	✓			
1997	Chans et al. [24]		✓		
2000	Rajendran and Chang [148]			Curvature-direction representation	
2005	Chalechale et al. [22]			APAI	
2010	Eitz et al. [49, 47]			HOG, Tensor, ARP and EHD	
2011	Eitz et al. [50]			SC, spark feature, HOG and SHOG	
2011	Hu et al. [80]			GF-HOG, SIFT, SSIM	
2011	Cao et al. [20]	✓			
2013	Hu and Collomosse [79, 78]			GF-HOG, HOG, SIFT, SSIM, SC and Tensor	
2014	Parui and Mittal [137]		✓		
2014	Li et al. [110]			HOG	
2016	Yu et al. [224]				Triplet network, Sketch-a-Net
2016	Sangkloy et al. [160]				Triplet network, GoogLeNet
2016	Song et al. [174]				Triplet network, Sketch-a-Net
2018	Dey et al. [38]				Unified CNN network
2018	Dey et al. [37]				Unified CNN network
No.	18	4	2	7	5

Deep Sketch-based image retrieval (SBIR) has been widely studied [27, 29, 38, 143, 160, 174, 175, 203, 223, 224] in recent years. Existing SBIR solutions generally aim to train a joint embedding space where sketch and photo can be compared using nearest neighbor techniques. Common embedding learning approaches include: (a) contrastive comparison based methods (implemented by pair-wise loss [191]), (b) ranking based methods [160, 224], (c) reinforcement learning based methods [10], (d) deep canonical correlation analysis (DCCA) [136] based methods [81], (e) cross-domain dictionary learning [213], etc. The most widely-studied methods are ranking-based, including triplet ranking [28, 223] and quadruplet ranking [169]. Table 2.4 shows different type of metric used in SBIR settings.

Table 2.4: The summary of similarity comparison metrics for different feature types

Feature type	Similarity metrics
Pixel-based	Accumulated pixel value differences
Contour-based	Accumulated scores of the match segment pairs
Histograms	City block,cosine, Chi-square and histogram intersection distances
Deep features	Euclidean distance

Ranking-Based SBIR

We next introduce the popular triplet and quadruplet ranking SBIR methods in detail. Given a sketch anchor $\mathbf{X}_n$ and its positive and negative photo retrieval candidates $\left(\mathbf{X}_{n,+}, \mathbf{X}_{n,-}\right)$, the goal of triplet ranking is

$$\mathscr{D}\left(\mathscr{F}\left(\mathbf{X}_n\right), \mathscr{F}\left(\mathbf{X}_{n,+}\right)\right) < \mathscr{D}\left(\mathscr{F}\left(\mathbf{X}_n\right), \mathscr{F}\left(\mathbf{X}_{n,-}\right)\right)$$

where $\mathscr{D}(.,.)$ is a distance metric (e.g., ℓ_2 distance). In common practice[224, 223], the positive sample is usually selected from the same class as the anchor. Specifically, the loss function of triplet ranking is typically

$$\mathscr{L}_{\text{triplet}} = \sum_{n=1}^{N} \max(0, \Delta + \left\| \mathscr{F}_\Theta(\mathbf{X}_n) - \mathscr{F}_\Theta(\mathbf{X}_{n,+}) \right\|_2^2 \\ - \left\| \mathscr{F}_\Theta(\mathbf{X}_n) - \mathscr{F}_\Theta(\mathbf{X}_{n,-}) \right\|_2^2 \Big)$$

where Δ is the margin to guarantee the minimum distance between the embedding pairs of $\{\mathscr{F}(\mathbf{X}_n), \mathscr{F}(\mathbf{X}_{n,+})\}$ and $\{\mathscr{F}(\mathbf{X}_n), \mathscr{F}(\mathbf{X}_{n,-})\}$

For quadruplet ranking [169], the input atom is a quadruplet of anchor $\mathbf{X}_n$, positive candidate $\mathbf{X}_{n,+}$, negative candidate $\mathbf{X}_{n,-}$ from the class of anchor, negative candidate $\mathbf{X}_{n,--}$ from a different class to anchor. The goal of quadruplet ranking is to ensure

$$\mathscr{D}\left(\mathscr{F}(\mathbf{X}_n), \mathscr{F}(\mathbf{X}_{n,+})\right) < \mathscr{D}\left(\mathscr{F}(\mathbf{X}_n), \mathscr{F}(\mathbf{X}_{n,-})\right) \\ < \mathscr{D}\left(\mathscr{F}(\mathbf{X}_n), \mathscr{F}(\mathbf{X}_{n,--})\right)$$

Based on this, quadruplet ranking is essentially multi- task or multiple triplet ranking by constructing two extra triplet relationships, in order to encode more semantic information into the embedding space. In particular, Seddati et. al. [169] constructed three triplets from each quadruplet, including $triplet_a = \{\mathbf{X}_n, \mathbf{X}_{n,+}, \mathbf{X}_{n,-}\}$, $triplet_b = \{\mathbf{X}_n, \mathbf{X}_{n,+}, \mathbf{X}_{n,--}\}$, $triplet_c = \{\mathbf{X}_n, \mathbf{X}_{n,-}, \mathbf{X}_{n,--}\}$ Therefore, the quadruplet ranking loss is defined as

$$\mathscr{L}_{quadruplet} = \mathscr{L}_{triplet_a} + \lambda_b \mathscr{L}_{triplet_b} + \lambda_c \mathscr{L}_{triplet_c}$$

where λ_b, λ_c, are the weights.

In ranking-based SBIR, the anchor is usually from the sketch domain, and other samples are photos. Both triplet and quadruplet ranking can be used for either category level or instance level SBIR tasks.

The essential principle of triplet loss is using local partial orderings to establish a global ordered relationship in the embedding space, so that triplet ranking can be Topological Sorting. The triplet annotations work as a partially ordered set. Compared with other loss functions, the main advantages of triplet loss are:

1. It helps to involve more local partial orderings and annotations to learn more fine-grained embedding space.

2. Given a limited number of N training samples, their triplet orderings have C_N^3 combinations, producing significant annotation augmentation. This is beneficial for training deeper net- works on smaller sketch datasets. It should be noted that the performance of triplet loss is heavily dependent on

   i) the choice of margin parameter and

   ii) the triplet construction strategy.

Note also that ranking-based SBIR models can be improved by multi-task training along with classification [160, 15, 112]. Furthermore, rather than purely discriminative training, SBIR can also be tackled by generating one modality from the other [71], e.g., using conditional GAN [147]; or using generative losses to regularize discriminative training [135]. SBIR training can also be combined with post-processing re-ranking [8, 9, 81] to refine the initial learned embedding spaces.

Network Architectures in SBIR

SBIR models generally need two or more branches to process sketches and photos for comparison using the metrics introduced above. As, shown in Figure 2.6, Both triplet and quadruplet ranking models can use backbone networks that are:

1. Siamese networks [224] use full weight-parameter sharing across branches.
2. Semi-heterogeneous network [14] use partial weight sharing across the branches. Typically early layers are modality specific, and weight-tied layers are at deeper layers.
3. In heterogeneous networks [102], the sketch branch uses independent parameters to the photo branches.

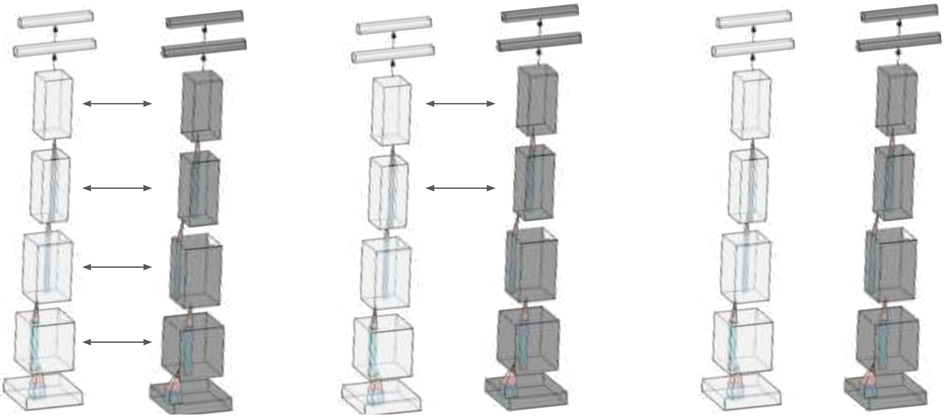


Figure 2.6: Different weight sharing manners (left: Siamese, middle: semi- heterogeneous, right: heterogeneous) for CNN-based cross-modal net-works. The hollow and shaded networks denote the branches for sketches and photos, respectively. The double sided arrows indicate sharing of weights

The trade-offs underlying these architectures are that sharing weights enables more data (both sketch and photo) to be used to estimate parameters, reducing over-fitting. But separating weights enables the sketch/photo branches to adapt more specifically to their respective domains. Weight sharing considerations are discussed in more detail in [14].

In order to achieve faster sketch-based image retrieval, recent research has studied optimizing the feature coding (e.g., sketch-image hashing [116, 228]), and the feature map (e.g., Asymmetric Feature Maps [184]). In particular, sketch-image hashing (or hashing SBIR) has gained attention. Liu et al. [116], propose the first deep hashing model for SBIR, which is a classic deep hashing pipeline including:

1. Feature extractor network,
2. Hashing layer with binary constraints, and
3. Hashing loss which is sometimes alternatively optimized in two step along with the extractor backbone.

The last hashing loss pipeline has been widely studied in photo oriented deep hashing [113, 193] where the hashing layer is typically fully-connected with sigmoid or tanh activation, and a discrete binary constraint. The loss functions of deep hashing models are often non-differentiable, due to the discrete binary constraints. Thus common practice is that the feature extractor backbone and hashing layer are alternatively optimized in two separate steps by fixing one and optimizing the other. Existing SBIR hashing models work on the SBIR bench- marks, i.e., Sketchy [160] (75K sketches) and TU-Berlin Extended [227] (20K). The scale of these benchmarks is not yet large enough to thoroughly test hashing SBIR methods.

Current issues in SBIR include: Self-supervised pre-training for SBIR [136], optimizing SBIR for early retrieval using partially drawn sketches, for example using reinforcement learning [10]; investigating whether costly sketch-photo annotation pairs can be replaced with edge-maps [146] and cross-category generalization of SBIR.

2.5.2 Zero-Shot SBIR

Many existing SBIR works assume that categories to be queried are included in the training set. In recent years, motivated by the zero-shot validation criterion for supervised photo retrieval [158], zero-shot sketch-based image retrieval (ZS-SBIR) has also been studied [40, 42, 43, 106, 116, 120, 132, 133, 134, 183, 187, 216, 220, 232]. Similar to natural photo zero-shot learning/recognition [59, 201], ZS-SBIR systems aim to enable query and retrieval of categories that are from unseen categories. i.e., categories that have not been involved in training stage. This is important in practice, e.g., for an e-commerce application of SBIR, where new products should ideally be enrolled in the search engine without requiring re-training.

ZS-SBIR systems can follow conventional zero-shot learning methods [58, 59, 201] in exploiting auxiliary knowledge such as word vectors [124], attributes [100], or class hierarchy to define the model for the unseen class. However, directly synthesizing a retrieval model for novel classes with auxiliary knowledge leads to the same challenges of ZSL (cross-category domain-shift [58]; inconvenient need to specify nameable categories at testing time [58]). Meanwhile it would entail new challenges specific to SBIR: (i) Knowledge transfer needs to occur across both sketch and photo views. (ii) Some kinds of auxiliary knowledge may not make sense for sketch (e.g., banana-is-yellow may be visible in photo but not sketch). Meanwhile, auxiliary knowledge transfer is not strictly necessary for retrieval in the way that it is for category recognition. Therefore many ZS-SBIR methods tackle the problem in a domain generalization [116] manner. That is, training a robust matching network on the training categories and then applying it directly to unseen testing categories.

Thus common approaches are to train ranking [40, 133] or generative [42, 219] models for retrieval, which are enhanced and mode robust by constraints such as domain-alignment losses [40, 42, 133] and auxiliary semantic knowledge reconstruction [40, 42]. In these cases the auxiliary semantic knowledge is only used to constrain representation learning at test time and is not required during training time as for conventional ZSL – thus maintaining the vision that SBIR should only depend on ability to depict and not to verbally describe. Current directions include extending SBIR to the generalized zero-shot setting, where testing categories are a mix of training and unseen categories [42, 133]; extending sketch-photo hashing to the zero-shot setting [170]; and training SBIR without paired samples [43].

2.6 Generation

Sketch generation has grown rapidly in recent years as deep learning-based approaches easily outperform earlier classic sketch generators [128, 111]. Sketch generation has several practical applications, e.g., synthesizing novel pictures, assisting artist design, and finishing incomplete sketches. It can be addressed using various deep learning tools, e.g., recurrent [166] [108], variational autoencoder (VAE) [72, 85, 86], Generative Adversarial Network (GAN) [186], VAE- GAN [186], and reinforcement learning (RL) [186, 233].

The seminal model SketchRNN [72] is a sequence-to- sequence VAE for conditional and unconditional generation of vector sketches. Its encoder and decoder are implemented by bidirectional RNN [168] and unidirectional RNN, respectively. As stated earlier, free-hand sketches can be represented as a sequence of key points defining strokes. The main idea of SketchRNN is to simulate human sketching by sequential generation of these key points in terms of location and pen up/down status.

The VAE encoder of SketchRNN takes vector sketches as input, and encodes it as a vector h, which is the RNN's last hidden state. This vector will be further encoded as two parameters μ and σ to model a Gaussian distribution N (μ, σ), from which a latent vector z will be sampled. Then, the Long Short Term Memory (LSTM) based VAE decoder will generate the coordinates and pen states of the key stroke points, conditional on z. In particular, the coordinate and state for each key point is sampled from a Gaussian mixture model (GMM), and also used as input for the next decoder step. To improve SketchRNN to deal with multi-class generation, Cao et. al. [19] propose a generative model named as "AI-Sketcher", which is also a VAE based network.

Another line of work within sketch generation uses differentiable rendering [232] or reinforcement learning [186, 233, 60, 123] to train policies that draw sketches iteratively according to different criteria such as adversarial training against human sketches [60]. This line of work often considers factors not addressed by SketchRNN such as brush style and color. By considering an interpretable latent representation of sketches, such methods can also potentially be used to de-render sketches into programs or symbols [53, 44].

Recently, there are several trends in sketch generation, notably:

1. Fine-grained sketch generation [86].
2. A novel evaluation metric "Ske-score" [186], aims to provide a better metric to quantify the goodness of generated vector sketches.
3. Transformer-based architectures [152, 207] are being applied to sketch generation.

2.7 Applications of SBIR

SBIR is largely at an early stage of development as most SBIR systems are for demonstration only. In this section we review several popular SBIR applications of which some are also in the field of CBIR.

QBIC [56] supports both blob-based and line-art SBIR alongside CBIR. A sketch query can also be composed using a blocky colour layout diagram. For scalability, the database images are clustered with each cluster having a representative descriptor. A query is first matched against the representatives before a subsequent exhausted search is performed on the top-ranked clusters.

MindFinder [190] and **Sketch2Photo** [25] both offers visual search with annotated sketches. The applications uniquely allow search of multiple objects with spatial constraints in a combined QKW and QVE paradigm. The user-interface is friendly allowing fast query construction although text search is not conducted directly. A simplified version of MindFinder is SketchMatch, a shape-only SBIR game available on Windows Apps Store.

Gboard [1] and **Emoji Recognizer** developed by Google for emoji search on mobile phone keyboard and Android Wear respectively. These applications support SBIR on a small database of thousands of emoji.

DeTexify is an publicly available sketch based symbol search tool developed to enable users of the LaTeX authoring environment to quickly locate markup command codes for mathematical symbols.

Chapter 3

Learning Cross-Modal Deep Embeddings for Multi-Object Image Retrieval using Text and Sketch

I have become intrigued with the combining of seemingly unrelated ideas or images, or the drawing upon the many, sometimes dissimilar, meanings a word might have.
– by John Barton

In this chapter we introduce a cross-modal image retrieval system that allows both text and sketch as input modalities for the query. A cross-modal deep network architecture is formulated to jointly model the sketch and text input modalities as well as the image output modality, learning a common embedding between text and images, and between sketches and images. In addition, an attention model is used to selectively focus the attention on the different objects of the image, allowing for retrieval with multiple objects in the query. Experiments show that the proposed method performs the best in both single and multiple object image retrieval in standard datasets.

3.1 Introduction

With the advent of touch screen and pen input devices, sketches have emerged as an alternative mode to provide the query that can deal with the limitations of text and images. Sketches are a natural way to conceptualize visual objects in terms of simplified shapes and their pose, however they have a few constraints. Sketching needs some idea and ability in drawing shapes that can make some users uncomfortable with this modality. Thus a SBIR framework can not replace conventional text based retrieval which has its own benefits (e.g. utilization of keyboard versus stylus). Thus both modalities can complement each other. As most existing systems allow users to query by image (CBIR) or by text (TBIR). Though these two alternatives provide an effective way to interact with a retrieval system by providing the query as an example image or text, they also pose some constraints in situations where either an example image is not available or text is not sufficient to describe the query. In this type of scenarios sketches arise as an alternative mode to provide the query that can handle the limitations of text and image. As sketches can efficiently and precisely express the shape, pose and fine-grained details of the target images, they are becoming popular as an alternative form of query.

Though in many scenarios sketches are more intuitive to express the query they still present some limitations. Drawing a sketch can be tedious for some users not skilled in drawing, and a sketch based interface may need special hardware (e.g. stylus) which might not be always available. Thus a SBIR system can not entirely replace traditional text based retrieval which has its own convenience (e.g. use of keyboard vs stylus), but can effectively supplement and/or complement the traditional text based querying in many cases. Thus, a multimodal image retrieval system can be envisioned, where the query can be either a sketch or a text. The popular use of smartphones with touch screen interfaces where people stores many personal pictures will bring innovative search services based in this multimodal input. Although recently many approaches [155, 160, 143] for sketch based image retrieval have been proposed, none of them permit to use text as an additional or complementary input modality. On the other hand traditional text based retrieval systems only permit to use text as their query modality.

Another significant limitation of most existing image retrieval pipelines is that they can only deal with scenarios where only one salient object is significant (see Fig. 3.1 to better understand the limitations of current methods). To the best of our knowledge none of the sketch based image retrieval methods can deal multi-object scenarios, however few methods [70] based on textual queries are able to retrieve relevant images containing multiple objects by learning a common subspace between text description and image. Nevertheless, these methods need a detailed description (caption) about the images to be retrieved, sometime which is unfeasible to provide. Additionally these methods rely on co-learning of text and images in a semantic space, and often limited to a closed dictionary. Hence, we propose a unified image retrieval method, which can take both sketch or text as query. The method obtains different representations for text, sketch and image and then learns a common space where it is more meaningful to compute distances between text and images and between sketches and images. Additionally, the method includes an attention mechanism that permits to focus retrieval on those parts of the image relevant to the query. In this way, our approach can also perform multi-object image retrieval.

In this chapter we describe the first contribution of this thesis that can be summarised as follows:

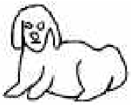
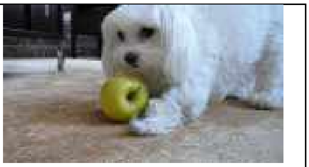
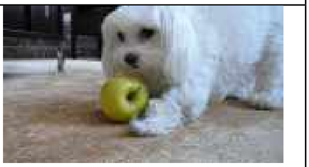
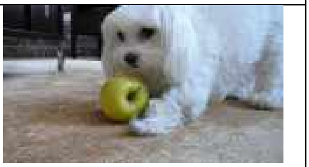
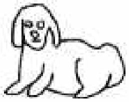
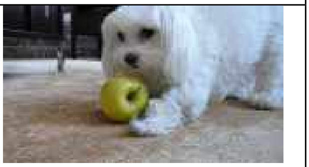
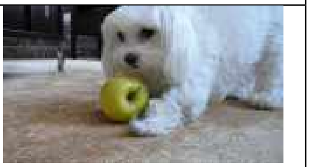
Modalities	Input	Retrievals
Sketch		
Text	“apple”	
Description	“Dog with apple”	
Sketch + Sketch	+ 	
Text + Text	“dog”+“apple”	

Figure 3.1: Input modality vs retrieval results: examples shown in the last two rows are addressed by our proposed method.

- A unified framework for cross-modal image retrieval that can perform both sketch to image or text to image retrieval based on learning a common embedding between text and images and between sketches and images.

- A retrieval framework for images having multiple salient objects that leverages an attention model to select the subset of image features relevant to each query object.

The rest of the chapter is organized as follows: Sect. 3.3 describes our proposed cross-modal and multi-object image retrieval framework with all details on text, sketch and image models. In Sect. 3.4, we describe the datasets and the experimental protocols we have used to show the effectiveness of the method and present the results of the experiments. Finally, Sect. 3.5 concludes the paper and some future directions of the present work are mentioned.

3.2 Related Work

As we are proposing a retrieval system where input queries can be of different modalities, our work is related to multimodal retrieval approaches like [198]. However, none of the existing works in multimodal retrieval actually propose to combine text and sketch, but there are several image retrieval systems for each of these two modalities, which are in a way related to our work. Thus, in this section we will review those works related to multimodal retrieval of images focusing on sketch based image retrieval and text based image retrieval approaches.

Sketch Based Image Retrieval (SBIR) is challenging because free hand sketches drawn by humans have no references but focus only on the salient object structures. In comparison to natural images, sketches are usually distorted. In recent years some studies are made to bridge the domain gap between sketches and natural images, in order to deal with this problem. These methods can be grouped into two categories namely hand-crafted methods and cross domain deep learning-based methods. Most of the hand-crafted SBIR methods first generate an approximate sketch by extracting edge or contour maps from the natural images. Then hand-crafted features (e.g. SIFT [119], HOG [30], gradient field HOG [79], histogram of edge local orientations (HELO) [155] and Learned Key Shapes(LKS) [156]) are extracted from both the sketches and edge maps of natural images. Sometimes these features are further clustered using 'Bag-of-Words'(BoW) methods to generate an embedding for SBIR. One of the major limitations for such methods is the difficulty to match the edge maps to non-aligned sketches with large variations and ambiguity. To address the domain shift issue, convolutional neural networks (CNNs) methods [96] have recently been used to learn domain transformable features from sketches and images with an end-to-end framework [160, 143, 224]. Both category [48, 155, 156] and fine grained SBIR [160, 224, 108] tasks achieve higher performance with deep methods which better handles the domain gap. All the current deep SBIR methods tend to perform well only in a single object scenario with a simple contour shape on a clean background.

On the other hand few works exist which jointly leverage images and natural text for different computer vision tasks. Zero shot learning [13], language generation [5], multimedia retrieval [188], image captioning [55] and Visual Question Answering [4] build a joint space embedding for textual and visual cues to compare both modalities directly in that space. The first category of methods are based on Canonical Correlation Analysis (CCA) for obtaining a joint embedding [73]. There are recent methods that also use deep CCA for such embedding [93]. Alternatively to CCA there are other methods that learn a joint space embedding using ranking loss. A linear transformation of visual and textual features with a single-directional ranking loss is presented in WSABIE [206] and DeViSE [57]. Bidirectional ranking loss [90] with possible constraint is seen in [199]. Cross-modal queries are often done in these joint image and text embedding i.e. retrieve image with textual queries and vice-versa [199]. In many of these works learning the joint embedding is by itself, the final objective. In contrast to these works we try to leverage both the joint space embedding and a sequential attention model to retrieve images having multiple objects. SBIR suffers the major drawback of varying drawing skills amongst users. Certain visual characteristics can be cumbersome to sketch, yet straightforward to describe in text. We hypothesize that these two input modalities can be modelled jointly with respect to the third for a successful retrieval system as semantically they represent the same. In experiments we will show that using a neural embedding we can jointly learn embedding spaces, where image and the query (sketch and/or text) can be represented as a vector, thus a simple nearest neighbour can be used for retrieval.

From the methodological point of view in order to learn different representation for different objects(in image), we exploit the recent advances in deep learning and, in particular, attention mechanisms [7] to select features from salient regions. Attention models allow to select a subset of features by an elegant weighting mechanism. Soft attention has been widely used in machine translation [7], image captioning [212]. Recentl, attention has also been used for multi object detection in [204]. In this work, we rely on an attention module to do a soft detection of different objects by giving more weights to the salient regions of the image corresponding to the query object. Thus, our attention module is responsible for doing an object detection in place.

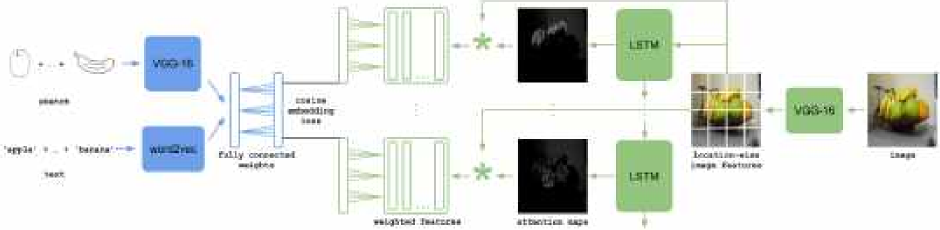


Figure 3.2: Overall architecture of our framework

3.3 Cross-modal and multi-object retrieval framework

In this section, we introduce our proposed framework (see Fig. 3.2), which is a common neural network structure that allows to query image databases with sketch as well as text inputs. For that, on one side, we have separately designed sketch and text models that respectively permit to obtain a feature representation of sketches and text. On the other hand, we use an image model that processes input images and outputs a set of image features weighted by the attention model. Finally, the network learns a common embedding between text/sketch features and image features. Below we provide the details of each part of the framework.

3.3.1 Sketch representation

For the sketch representation, we adopt a modified version of the VGG-16 network [171]. We replace the input layer to accept a single channel sketch image and the last fully connected layer to produce class wise probability for 125 classes. Henceforth, we retrain the entire network as a classification framework on the sketch images from the Sketchy dataset [160] having 125 sketch classes. Once trained we remove the last fully connected (FC) and softmax layer, for obtaining the sketch representation.

3.3.2 Text representation

For word representation, we have used the standard word2vec [124] representation, which is pre-trained on the set of words from the English Wikipedia[1]. This word representation produces a feature vector of 1000 dimensions.

3.3.3 Image representation

The image representation relies on the VGG-16 network [171] pre-trained on ImageNet. However, for obtaining the image representation the top FC layers are removed, and we take the output of the last convolutional layer. In this way, we extract a mid-level M-dimensional visual feature representation on a grid of spatial locations. Specifically, we use $P = \{p_l | p_l \in \mathbb{R}^M; l = 1, \ldots, L\}$ to represent the spatial CNN features at each of the L grid locations. Given a query with n objects (either sketches or text descriptions), we compute n different sequentially dependent attention maps for a single image utilizing an LSTM. Considering, λ_{i_l}, for $l = 1, 2, \ldots, L$ to be the weights on the L spatial grids for the ith query, we impose the following condition for obtaining a statistically meaningful feature description for each image and each query object.

$$\lambda_{i_l} \propto \texttt{exp}(F(p_{i_l})) \text{ s.t. } \sum_{l=1}^{L} \lambda_{i_l} = 1 \tag{3.1}$$

where F is the neural network model used for obtaining the weights on the spatial grids of the image. Practically, F is designed with an LSTM. Hence the final image representation for ith query results in as:

$$\mathbf{f}_i = \sum_{l=1}^{L} \lambda_{i_l} p_{i_l} \tag{3.2}$$

Each image feature $\mathbf{f}_i$ is the weighted average of image features over all the spatial locations $l = \{1, 2, \ldots, L\}$. In practice, the convolutional layers produce image features of $7 \times 7 \times 512$ dimensions, which after incorporating the attention weights results in 512 dimension feature vector delineating an image for a particular query object. Our LSTM

[1] `https://www.wikipedia.org/`

based attention map generator is a soft attention mechanism, that remembers the attended regions through hidden vector which negates the possibility of attention at the same object multiple times. The LSTM takes the CNN-based features as well as the hidden representation as input to generate the attention maps at each time step.

3.3.4 Joint neural embedding

Given the query (either text or sketch) and image representation as above, the goal is to project the respective representations into a common space. For doing so, in each case of sketch and text representation, we employ a non-linear transformation containing two fully connected layers. In case of image representation, it is done by augmenting a non-linear transformation consisting a single MLP layer on the weighted set of image features. We adjust the sizes of these respective non-linear transformations accordingly, to produce a 512 dimensional feature vector as the final representation, for each modality. The goal is to learn mappings of different modalities to make the embedding "close" for a given same query-image pair, and "afar" for different query-image pair. The training of this system is done by including the cosine embedding loss as follows:

$$\mathscr{L}_{\cos}((\mathbf{q},\mathbf{f}),y) = \begin{cases} 1-\cos(\mathbf{q},\mathbf{f}) & \text{if } y=1 \\ \max(0,\cos(\mathbf{q},\mathbf{f})-m) & \text{if } y=-1 \end{cases} \tag{3.3}$$

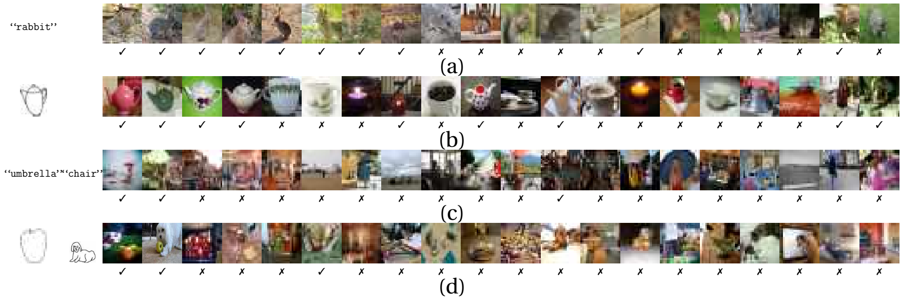
where **q** and **f** are respectively the learned query (either text or sketch) and image representation, cos denotes normalized cosine similarity and m is the margin. For training through this paradigm, we generate positive ($y = 1$), as well as, negative examples ($y = -1$), which respectively correspond to the query-image pairs belonging to the same and different classes.

3.3.5 Multiple objects queries

Our framework permits querying by multiple objects represented with the same modality, which is particularly useful for retrieving images containing multiple objects. In this case, the loss is computed in a cumulative manner over all the query objects:

$$\sum_{i=1}^{n} \mathscr{L}_{\cos}((\mathbf{q}_i,\mathbf{f}_i),y) \tag{3.4}$$

where n is the number of queries and **q** is the representation of any query object. Although, our method supports querying with multiple objects, in practice, we consider querying at most with 2 different objects. This is mainly because of the unavailability of the appropriate datasets needed for training and retrieval. While querying with multiple objects the sum of distances between the queries and the image is considered for ranking the retrievals.



(a)

(b)

(c)

(d)

Figure 3.3: Qualitative results obtained by our proposed method while querying the database with a combination of texts and sketches: (a) text query ‘‘`rabbit`’’, (b) sketch query *teapot*, (c) combination of text queries ‘‘`umbrella`’’ and ‘‘`chair`’’ and (d) combination of sketch queries *apple* and *dog*. (Best viewed in pdf)

3.4 Experimental results

3.4.1 Datasets

Sketchy

The Sketchy dataset [160] is a large collection of sketch-photo pairs. The dataset consists of images belonging to 125 different classes, each having 100 photos. After having these total 125 × 100 = 12500 images, crowed workers are employed for sketching the objects that appear in these 12500 images, which resulted in 75471 sketches. The Sketchy database also gives a fine grained correspondence between particular photos and sketches. Furthermore, the dataset readily contains various data augmentations very useful for deep learning based methods.

COCO

Originally the COCO dataset [114] is a large scale object detection, segmentation, and captioning dataset. We use the COCO dataset for constructing a database of images containing multiple objects. We use the class names of the Sketchy dataset and take all possible combinations by taking two class names. Afterwards, we download the images belonging to these combined classes, and use them for training and retrieval. Few combined classes having very less number (less than 10) images are eliminated, leaving 365 number of combined classes for the experiment.

3.4.2 Experimental protocol

Single object

For single object query, we use the Sketchy dataset [160].

Sketch-based image retrieval During the training step, the positive examples are fabricated considering the fine grained information. This follows that for an image I, we consider all the corresponding fine grained sketches drawn by different sketchers. Let n_s be the total number of sketches that correspond to the image I according to the fine grained information. Therefore, we construct n_s different positive training instances using the same image I. The negative training examples for all the n_s sketches are created by randomly choosing images from the classes other than that of I. In this way, we randomly select 80% of the images from each class for training and the rest for testing. However, during the test phase, we obtain the sketch and image representation from the respective models in independent manner, and the retrieval is done according to the distances between them. Here it is to be noted that the fine grained information is not considered for ranking the retrievals.

Text-based image retrieval In case of text-based image retrieval as well, we randomly select 80% of the images from each class for training and the rest for testing. In this case, the positive training instances are created by considering the class label, the image, which creates training examples equal to the number of training images. Equal number of negative instances are created by randomly selecting images belonging to the class different from the texts.

Multiple objects

For multiple objects, we consider the database derived by us from the COCO dataset [114].

Sketch-based image retrieval As we mentioned before, in practice, we consider only 2 different objects, which is due to the unavailability of appropriate dataset for the particular task. In this case as well, 80% images of each combined class are selected for training set, and the rest of the images from each class are kept for the testing phase. To comply with diverse sketch instance, style and appearance, an image I is reconsidered n_m times. These n_m instances of the same image is framed with n_m different combinations of sketches that belong to the individual classes creating the combined class. The negative examples are created by associating each combined sketch query an image that does not belong to the same combined class.

Text-based image retrieval Creating training pairs for text-based image retrieval is relatively straight forward. As in the previous case, 80% images from each combined class are selected for training set, and the rest of the images from each class are kept for the testing purpose. For creating the positive training examples each image that belongs to a combined class is associated with the text labels of the individual classes. The negative examples are created by bracketing each combined text query an image that belongs to a different combined class.

For training our proposed neural network, we have used the Adam [92] optimization algorithm with a learning rate of 0.01 and a learning rate decay of 0.6. All the learnable parameters used in the model are initialized from a standard normal distribution $\mathcal{N}(0,1)$. The results shown in Table 3.1 are produced after training the respective text and sketch models for 50 epochs.

Table 3.1: Results obtained by our proposed method and comparison with state-of-the-art methods. (mAP)

Methods	Single Object		Multiple (two) Objects	
	Text	Sketch	Text	Sketch
S-HELO [155]	–	16.10	–	3.07
Sketch-a-Net [226]	–	20.80	–	3.39
GN Triplet [160]	–	65.18	–	4.23
Proposed method	**76.28**	**68.81**	**18.65**	**12.06**

3.4.3 Discussions and comparisons

In Table 3.1, we have presented the results obtained by our proposed framework, and compared them with three state-of-the-art methods: S-HELO [155], Sketch-a-Net [226], and GN (Google Net) Triplet [160]. All these three methods proposed some kind of sketch-based image retrieval strategy either based on handcrafted or learning based feature representation. We have used either their technique or trained model to extract the final representation of the sketches and the images from our test set, and have employed them for retrieving images based on a sketch query. This step produces a mean average precision (mAP) [117] score for each of these methods, which are shown in Table 3.1. In case of querying by multiple sketches, we first obtain individual representation of each sketch query, used those representations to compute multiple (equal to the number of sketches) distance matrices, and take the average of them to calculate the final retrieval list. As these methods do not allow querying by text, we do not use them for text based image retrieval procedure. Although, in literature, there are some methods that allow querying by caption [55] they use a detailed text description of the query which is not our case. This is why we have not compared our method with any other text-based image retrieval method. Here, it is worth reminding that our method can retrieve images based on the combination of words describing the principal objects that appear in the images.

From the results shown in Table 3.1, it is evident that in case of single object images, our proposed sketch-based image retrieval method has performed the best. In this case, GN Triplet has performed quite closely to us. In case of multiple objects images, all three state-of-the-art methods have performed quite poorly. Although used by us for retrieving images based on multiple queries, these state-of-the-art methods had not been designed for doing this specific task. Therefore, averaging the distances over multiple queries might have lost substantial information, which can explain these poor results. We have observed that our text-based image retrieval system performed considerably better than the sketch-based system, both in case of single and multiple object scenarios. This is probably because the spaces for sketch and image are apart than the ones from text and image.

In Fig. 3.3(a) and Fig. 3.3(b), we present two qualitative results respectively by querying with a text and a sketch input for single object. From the retrieved images shown in the figures, it is clear that the quality of the retrieval is quite satisfactory, as the first few images in each case belong to the same class as the query. Furthermore, the wrong ones have the major amount of context similar to the correct retrievals. For example, in Fig. 3.3(a), together with "rabbit" some images of "squirrel" also appear because they share substantial amount of context like grass, soil etc. This phenomena also appeared to be true in case of sketch based retrieval (see Fig. 3.3(b)). The qualitative results of retrieving images based on multiple text and sketch queries are respectively shown in Fig. 3.3(c) and Fig. 3.3(d). In these cases as well, the first few retrieved images are mostly true images. However, the number of true images retrieved are much less than the single object case. The majority of the false retrieved images contain objects belonging to one of the queried class, which is quite justified.

3.5 Conclusions and future work

In this chapter, we have proposed a common neural network model for sketch as well as text based image retrieval. One of the most important advantage of our framework is that it allows one to retrieve images queried by multiple objects of the same modality. We have designed an image attention mechanism based on LSTM that allows to put attention on the specific zones of the images depending on the inter related objects which usually co-occur in nature. This has been learned by our model from the images in the training set. We have tested our proposed framework on the challenging Sketchy dataset for single object retrieval and on a collection of images from the COCO dataset for multiple object retrieval. Furthermore, we have compared our experimental results with three state-of-the-art methods. We have found that our method has performed satisfactorily better than the considered state-of-the-art methods on all the two datasets with some cases of failure with justifiable reasons.

One of the future directions of this work will obviously focus on improving the retrieval performance on both type of datasets (single or multiple objects). For this purpose, we plan to investigate on more efficient training strategies. Furthermore, our framework can potentially allow to query by multiple modalities at the same time. We will also explore this possibility which will allow the users to query a database in a more efficient and effortless manner. In this regard, we should make both the queries and the loss calculation more efficient to accommodate both multi-modal combined query search and faster retrieval strategies

Chapter 4

Aligning Salient Objects to Queries: A Multi-modal and Multi-object Image Retrieval Framework

Everything should be made as simple as possible,
but not one bit simpler.
– by Albert Einstein

In this chapter we propose an approach for multi-modal image retrieval in multi-labelled images. A multi-modal deep network architecture is formulated to jointly model sketches and text as input query modalities into a common embedding space, which is then further aligned with the image feature space. Our architecture also relies on a salient object detection through a supervised LSTM-based visual attention model learned from convolutional features. Both the alignment between the queries and the image and the supervision of the attention on the images are obtained by generalizing the Hungarian Algorithm using different loss functions. This permits encoding the object-based features and its alignment with the query irrespective of the availability of the co-occurrence of different objects in the training set. We validate the performance of our approach on standard single/multi-object datasets, showing state-of-the art performance in every dataset.

4.1 Introduction

As mentioned in the previous chapter, current trends explore TBIR systems that could bridge the semantic gap with queries based on natural language descriptions of the image content. Despite its wider expressiveness there can still be circumstances where text cannot be adequate to portray the query and it can be difficult to establish the link between text and image contents.

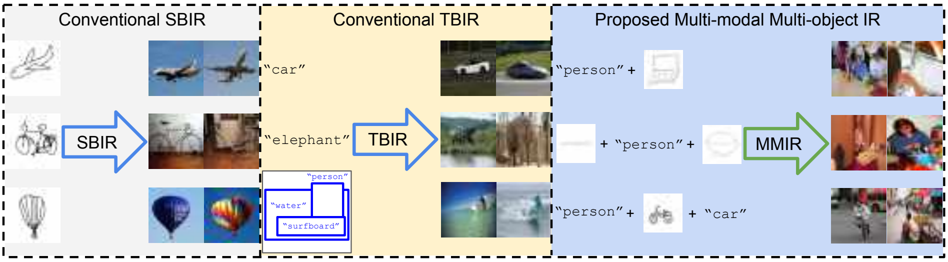


Figure 4.1: In conventional SBIR and TBIR, during the training phase, respectively a sketch and a text representation is mapped to the image representation of corresponding class. Querying images with multiple labels has been explored within the TBIR domain [121] (as shown in the last row of Conventional TBIR column). Querying images with multiple objects using multi-modal queries provides convenience in searching, but it is an extremely challenging task and has not been addressed yet.

Although numerous methodologies for SBIR have been proposed, none of them allow to utilize text as an extra or complementary query modality. Thus, the first motivation for the work presented in this chapter is to propose a multi-modal image retrieval approach where the query can be either sketch or text or both. To make the different modalities compatible, a common semantic embedding space is defined.

Another noteworthy constraint of most existing image retrieval pipelines is that they can only manage situations where just a single salient object is significant – see Fig. 4.1. This motivates the second challenge of the present work, i.e. allowing to express queries that can refer to multiple objects. In this way, the proposed model provides more expressiveness to the search language since users can construct queries consisting of different concepts that are aligned to the salient objects of the target images. To the best of our understanding, none of the SBIR techniques can deal with multi-object scenarios. Although some of the strategies [70] based on textual queries can recover relevant images containing multiple objects.

Consequently, we propose a unified multi-modal and multi-object image retrieval (MMIR) framework, which permits to retrieve images containing multiple objects, expressing the query using text, sketches or a combination of both. The framework is based on a deep network architecture in which multi-modality is addressed by projecting word2vec representations of sketches and text into a common semantic space aligned with the image feature space. To deal with multi-object search we integrate an

LSTM-based visual attention model that learns to discover relevant zones of the image. To match the set of attention glimpses with the set of queries we propose a Hungarian loss that finds the best correspondence between both sets. The Hungarian loss is also used to introduce supervision while training the visual attention model by guiding the result of the attention towards object bounding boxes.

The main contributions of this work are: (1) The proposal of a common semantic space among text and sketches, obtained through word2vec representation of both input modalities, and aligned with the image feature space; (2) A visual attention model that automatically detects salient objects from an image, that is trained in a supervised way in order to minimize the assignment cost between attention output and object bounding boxes.

The rest of the chapter is organized as follows: in Sect. 4.2, we review the relevant state-of-the-art. Sect. 4.3 describes in detail our proposed cross-modal/multi-object image retrieval framework. In Sect. 4.4, we describe the experimental framework and present the results of the experiments. Finally, Sect. 4.5 draws the conclusions and outlines the future directions.

4.2 Related Work

In this section we review image retrieval using text – text based image retrieval (TBIR) – and sketches – sketch based image retrieval (SBIR) keeping the chapter in mind. Subsequently, as our method detects object as part of multi-object image retrieval by means of an attention procedure, we also discuss about the state of the art on attention models.

Advances in feature learning have recently provided effective feature representations for different modalities such as text [57, 126], images [69, 139], and hand-drawn sketches [224, 160] which have been shown to greatly improve the retrieval performance. A common approach when dealing with multi-modal data is to learn a joint embedding to map all modalities into a common latent space [150]. However, in multi-modal image retrieval [198], the complementary use of text and sketch to express the queries has not been much explored [39].

TBIR, dating back to the late 1970s, has evolved from just a keyword-based task to a more challenging task based on natural language descriptions (e.g., sentences and paragraphs) [95]. Queries in the form of sentences rather than keywords refer not only to object categorical information but also interactions, such as spatial relationships between objects [101, 210]. In our work we keep text in the simple form of keywords, but we permit to express objects relationships as combination of several text or sketch based queries. Recently, projecting text into the word2vec space has been shown to achieve a high level of accuracy in TBIR [67]. Thus, we also rely on word2vec to represent text queries, but we also extend this idea to obtain a semantic embedding of sketches into the word2vec space.

SBIR is one of the alternative ways of searching and overcoming the limitations imposed by TBIR systems. Apart from images, sketches have been successfully used for 3D shape retrieval purpose as well [234, 209]. Since sketch is more close to the semantic representation, it tends to help retrieving target results in the user mind from a semantic perspective [208]. Here, the main challenge is to bridge the domain gap between sketches and natural images. In literature, methods addressing this issue can be grouped into two categories: (1) hand-crafted methods, (2) deep learning-based methods. The hand-crafted features (e.g. SIFT [119]), gradient field HOG (GF-HOG [79]) are extracted from both the sketches and edge maps of natural images and further clustered using 'Bag-of-Words' (BoW), histogram of edge local orientations (S-HELO) [155] or Learned Key Shapes (LKS) [157]).

One of the major limitations for such methods is the difficulty to match the edge maps to non-aligned sketches with large variations and ambiguity. To address the domain shift issue, convolutional neural networks (CNNs) methods [96] have recently been used to learn domain-transformable features from sketches and images with an end-to-end framework [160, 224]. In our work, we address these semantic gap by directly projecting sketches to a semantic space using word2vec, that is further aligned with the image space. The current deep SBIR methods tend to perform well only in a single object scenario with a simple contour shape on a clean background [116, 160, 224]. Recently, there have been attempts to apply deep learning to multi-label image recognition task [216, 205, 197]. Razavian *et al.* [149] applies off-the-shelf features extracted from a deep network pre-trained on ImageNet [154] for multi-label image classification. Wang *et al.* [197] utilize RNNs to learn a joint image-label embedding to characterize the semantic label dependency as well as the image-label relevance. Some works exploit object proposals to only focus on the informative regions, which effectively eliminate the influences of the non-object areas and thus demonstrate significant improvement in multi-label image recognition task [205, 216]. More specifically, Wei *et al.* [205] propose a Hypotheses-CNN-Pooling framework to aggregate the label scores of each specific object hypotheses to achieve the final multi-label predictions. Yang *et al.* [216], formulate the multi-label image recognition problem as a multi-class multi-instance learning problem to incorporate local information and enhance the discriminative ability of the features by encoding the label view information. As an alternative we propose to use an attention model to discover image regions relevant to each of the objects. Visual attention model has gained a lot of interest recently. In image captioning [211] visual attention assists the generation of descriptive captions. [220] targeted attention on a set of concepts extracted from the image to generate captions. In visual question answering [172, 235] several models have been proposed which attend to image regions or questions when generating an answer. Concurrently, [31] analyzed the consistency between human and deep network attention in visual question answering. Our goal differs in that we are interested in how attention on salient objects can be aligned with the queried object. We use the attention correction proposed in [115] to create a supervised attention for salient object detection. A Hungarian Loss [177] function is proposed to match the salient objects with the queries projected to a semantic embedding space.

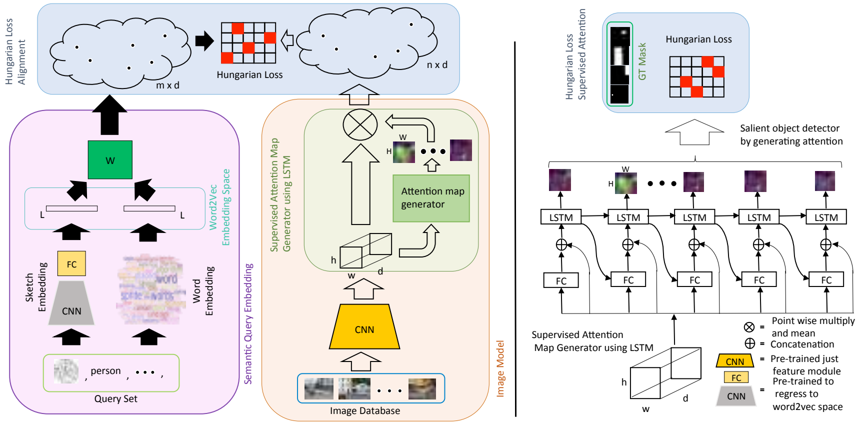

Hungarian Loss Alignment
m x d
Hungarian Loss
n x d
W
L
L
Word2Vec Embedding Space
Sketch Embedding
FC
CNN
Word Embedding
person , • • • ,
Query Set
Semantic Query Embedding
Supervised Attention Map Generator using LSTM
W
H
Attention map generator
h
w
d
CNN
Image Database
Image Model
Hungarian Loss Supervised Attention
GT Mask
Hungarian Loss
Salient object detector by generating attention
W
H
LSTM
LSTM
LSTM
LSTM
LSTM
FC
FC
FC
FC
FC
Supervised Attention Map Generator using LSTM
h
w
d
⊗ = Point wise multiply and mean
⊕ = Concatenation
CNN = Pre-trained just feature module
FC CNN = Pre-trained to regress to word2vec space

Figure 4.2: Architecture of the proposed unified MMIR framework. The semantic query network (in purple) and the image model (in orange) are aligned using the Hungarian loss (in blue). Right side of the figure elaborates supervised attention map generator (in green). The Hungarian loss for attention (in blue) computes the assignments in test time and computes the assignments and loss during training.

4.3 Multi-modal and Multi-object Image Retrieval

In this section we describe the proposed methodology (see the architecture in Fig. 4.2). The query (text and sketch) is embedded into a common semantic space. For each image in the image database, an LSTM based attention map generator finds several attention maps (one at every time step of LSTM) based on a CNN feature map and the previous attention map. These attention maps are trained in a supervised way and can be thought of as the relative importance of the different areas of the image in order to get the feature representation of the different salient objects in the image. For every attention map, the CNN features are weighted and averaged to get the final feature representation at every step. Thus, we obtain a set of features corresponding to different salient objects in the image. On the other hand, we allow for multiple queries each of them embedded in the semantic space. Consequently we have a set of query features and a set of image features that have to be matched and aligned. To compute a distance between these two sets, we use a Hungarian loss that gives the minimum cost assignment between them. Cosine distance between query features and object features is used to compute the individual cost between each pair query/object. In the following sections we introduce the details of each of the components of this global architecture.

4.3.1 Semantic Query Embedding

For successful retrieval of images given text or sketch as query, a proper embedding space must be defined, where both text and sketch can be directly compared to the image with a distance measure. In the case of sketches this has been achieved by learning a global feature using triplet loss [160] or similarity loss [143]. For text, either one hot vector encoding based on a fixed vocabulary or some semantic embedding mapping words into a continuous vector space can be used. In our case, we chose to use a semantic embedding as it gives the opportunity to use generic words as query and the model is not restricted to a certain vocabulary of words. For the combination of sketch and text, in [39] two different subspaces were proposed, one between text and image another between sketch and image. However, end to end training of different subspaces is difficult and unstable.

We argue that a better option is to find a subspace where all three modalities can be compared. Therefore, we propose a semantic space to embed the queries, capturing the common properties of text and sketch. In particular, we use word2vec [124] representation to obtain this semantic space. For the text we use directly the word2vec embedding of the words. Sketch images are regressed to the word2vec space by using a CNN based regressor. In the following we provide the details of each embedding.

For word encoding, we have used the standard word2vec [124] representation, which is pre-trained on the set of words from the English Wikipedia[1]. This word representation produces a feature vector of 1000 dimensions.

For the regression of sketch images into semantic space, we adopt a modified version of the VGG-16 network [171]. We modified the top fully connected layer module to accommodate the output vector dimension to that of word2vec. The entire network was then trained as a regression framework with cosine embedding loss to project the sketches in a space parallel to word2vec. For doing so, we used the sketch images from the Sketchy dataset [160] to produce the corresponding word2vec representation of the class name. Once trained, we used the network for obtaining the sketch representation mapped into the word2vec space.

4.3.2 Supervised Attention for Image Representation

With the goal of retrieving images relevant to a set of query objects, our aim is to extract a set of feature vectors representing salient objects from a particular image. Another possible alternative would be to encode the image into a global signature and aggregate multiple queries into a single representation making retrieval a nearest neighbour problem in this feature space.

However we argue that this approach has severe limitations. Firstly, to find a suitable global image representation that can encode the presence of multiple objects we should train with a well curated dataset containing possible combination of different

[1] `https://www.wikipedia.org/`

objects. Such a dataset would be huge and training would be complex and take a lot of time. Secondly, although aggregation of queries could be solved in various ways, there is not an easy way to encode their relative position, which can be crucial to match with that of image. In [121] this problem is dealt with by using a spatial query box and then learning the relative position by using a conventional CNN. However, the position has to be provided by user which can limit the usability of the query interface. Another possibility would be to use a state-of-the-art object detector to detect object and compute the corresponding feature vectors for each object. However, most object detection pipelines assume rectangular objects and do not consider the image context around them. In addition, such an approach would eventually make the pipeline hard to train end-to-end due to different loss functions.

With all this in mind, we propose a different alternative by using a more flexible method based on an LSTM together with an attention model to detect multiple objects (in the form of mask) one at every step. The input to the attention model is a set of features extracted with a conventional CNN corresponding to a spatial grid in which every point represents an area in the image (through its receptive field). In a nutshell, our LSTM based object detector is trained to output one attention map at every step which depends on the previous attention map and the CNN features. The LSTM remembers the attended regions through the hidden vector, which prohibits it to attend the same object multiple times.

The soft attention mechanism was first used in computer vision by [211] where the attention model is not supervised in a sense that there is no loss calculated directly on the attention weights. Thus, the attention mechanism is free to attend anywhere, being only guided by the final output, which is calculated from the result of the attended features. However this model can be extended by applying a direct supervision i.e by directing the model "where to attend" [115]. In this work we used a similar framework but we did the following changes. Firstly, we changed the cross entropy loss between the target and the generated attention map and the softmax over the grid locations by a sigmoid cross entropy loss and a sigmoid activation function to generate a binary map over the grid. This follows from the hypothesis that that every grid location is independent and can be a potential region to attend. Secondly, in our case the order in which the different targets (corresponding to objects) are attended is not relevant. Thus, we need a way to match the un-ordered set of targets with the unordered set of generated attention maps. We solve this by finding minimum cost between the two sets by using a generalized Hungarian loss. We outline the details of the formulation for this matching later.

More formally, the attention model computes n different attention maps and corresponding image representations for a single image using an LSTM network. The input to the LSTM is a set of features extracted from a pre-trained CNN-based feature extractor resulting in L vectors, $\{\mathbf{a}_1, ..., \mathbf{a}_L\}, \mathbf{a}_i \in \mathbb{R}^d$ that correspond to L spatial locations of an image. The LSTM generates n attention maps, $\{\alpha_t \in [0,1]^L, t = 1, \ldots, n\}$ and their corresponding image representations: $\{\mathbf{x}_t \in \mathbb{R}^d, t = 1, \ldots, n\}$. Our LSTM model can be visualized in Fig. 4.2. At every step, the input the LSTM takes as intput the hidden representation $\hat{h}$, the local features given by $\{\mathbf{a}_1,...,\mathbf{a}_L\}$ and the attention map generated at previous time step. To generate the attention map the hidden vector is passed through an MLP layer followed by sigmoid activation, which provides attention maps that can be interpreted as the importance of every spatial location for the detection/retrieval of a certain object. Considering λ_{i_l}, for $l = 1,2,\ldots,L$ to be the weights on the L spatial grids for the i^{th} step, the final image representation for i^{th} attention map is the weighted average of image features over all spatial locations, $\mathbf{x}_i = \sum_{l=1}^{L} \lambda_{i_l} \mathbf{a}_l$

4.3.3 Hungarian Loss

In our framework we have to match two unordered sets of elements keeping two constraints. On one hand, we have m queries (represented as points in the regressed word2vec space) and n (where $n > m$) different image representations of every single image corresponding to the different attention computed through the LSTM. For retrieval we have to find the best matching between these two sets. On the other hand, in order to train in a supervised way the attention model we have to align the set of bounding boxes of the salient objects with the result of the n steps of the LSTM.

We have solved both problems using the same framework formulating them as a bi-partite graph matching problem, and using a variation of the Hungarian loss introduced in [177]. In this way, we compute the loss as the minimum cost assignment between every pair of elements in both sets. Given a cost matrix between every pair of elements, the computation of the minimum cost assignment is done by the Hungarian algorithm [127] in polynomial time.

We use two different cost functions for each of the two problems. The cost between the query features q and the computed image features x is given by the cosine dissimilarity between the query and the feature.

$$C_{\text{sim}}(\mathbf{q}_i,\mathbf{x}_j) = 1 - \cos(\mathbf{q}_i,\mathbf{x}_j) \tag{4.1}$$

For supervising the attention model, we have used the binary cross entropy as a cost function between each of the ground truth masks β and each of the generated attention maps λ.

$$C_{\text{attn}}(\lambda_i,\beta_j) = -(\lambda_i \log(\beta_j) + (1-\lambda_i)\log(1-\beta_j)) \tag{4.2}$$

4.4 Experimental Results

4.4.1 Implementation Details

We have implemented our method on PyTorch framework. For all experiments, the image features are extracted using the feature module of the pre-trained VGG-16 network model. This feature representation is particularly appropriate for our task as it can

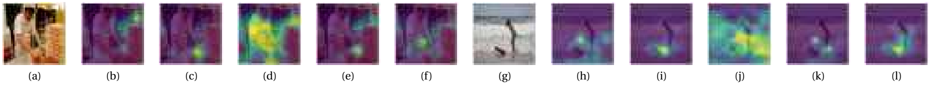


Figure 4.3: Images of (a) person with pizza and (g) person with surfboard; (b)-(f) and (h)-(l) are respective n ($n = 5$) attention maps for image (a) and (b) obtained by our LSTM-based image model.

effectively capture high-level semantic information from the images and at the same time it naturally retains most spatial information. For sketches, we used the same VGG-16 model but replacing the last layer in the classifier module with a specific layer to accommodate the regression of the features extracted from the sketch to the word2vec space. The output is a feature vector of 1000 dimensions, which is then mapped to a common joint neural embedding space. Its jointly trained by freezing the feature extraction part on both the query and the image side. The salient object detector based on supervised attention model was implemented using the mask generated from the bounding boxes of the objects in MS-COCO images. In case of the Sketchy database, we used the bounding box of the sketches in the image mask. This was possible because the dataset was designed for fine grained SBIR. We first compare the proposed method with several previous SBIR methods for single object, including hand-crafted features: GF-HOG [78], S-HELO [155], LSK [157]; and deep learning based: Siamese CNN [143], Sketch-a-Net (SaN) [224], GN Triplet [160], 3D shape [192], DSH [116] as shown in the Table 4.2. For all the methods mentioned above we follow the same protocol and evaluation metrics as in [116]. In the case of [116] we used the trained model for 128-bits provided in the author's github repository.

4.4.2 Datasets

Sketchy Dataset [160]: This is a large collection of sketch-photo pairs. The dataset consists of images belonging to 125 different classes, each having 100 images. After having these total $125 \times 100 = 12,500$ images, crowed workers were employed for sketching the objects that appear in these 12,500 images, which resulted in 75,471 sketches.

TU-Berlin Dataset [51]: The TU-Berlin dataset contains 250 categories with a total of 20,000 sketches. We also utilize the extended set provided in [116], with natural images corresponding to the sketch classes with a total size of 204,489.

MS-COCO Dataset [114]: Originally it is a large scale object detection, segmentation, and captioning dataset. We use the MS.COCO dataset for constructing a database of images containing multiple objects. As the label number for each image also varies considerably, rendering MS-COCO is even more challenging. We use the class names

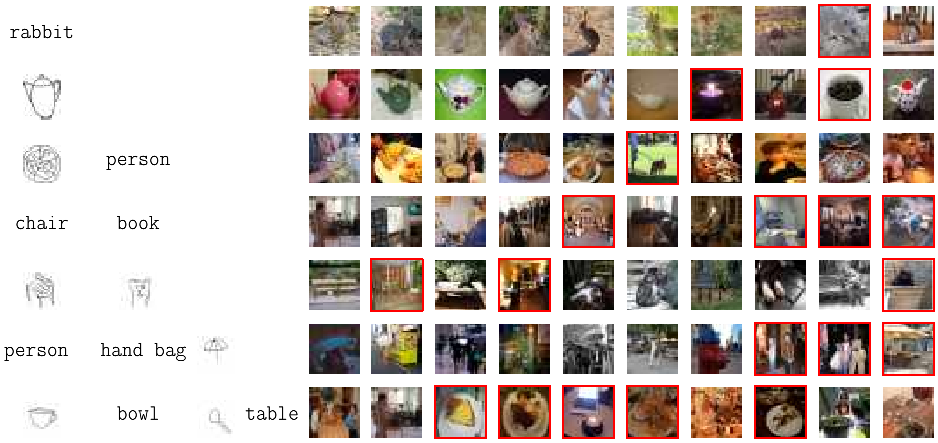


Figure 4.4: Qualitative results obtained by our proposed method: eight example queries consisting texts as well as sketches with their top-10 retrieval results on the Sketchy, TU-Berlin and MS-COCO dataset. Red boxes indicates false positives. (Best viewed in pdf)

of the Sketchy dataset and take all possible combinations by taking two, three, four, five class names. Afterwards, we download the images belonging to these combined classes, and use them for training and retrieval. Few combined classes having less than 10 images are eliminated, leaving 125 number of combined classes for the experiment with at least 900 images.

4.4.3 Ablation Analysis

In this subsection, we perform some experiments to carefully analyze the contribution of the critical components of our proposed model. For this study we used single object retrieval from the Sketchy [160] using sketch as a query modality. In an effort to evaluate each of our contributions, we trained every variation of the system with exactly the same training data.

In the first row of Table 4.1 we show the results that we obtain if we replace the VGG-16 network that regresses the word2vec representation of sketches by another VGG-16 network that just outputs a one-hot encoding representation of the word that represents the class of the sketch. The sketch features obtained from the one-hot encoding produce reasonable retrieval performances, but the figures are still clearly lower than our original design in the last row of the table (12% less in mAP). We speculate this is because hidden representations and knowledge within the trained neural network is able to store more information by correlating the data points that are semantically close.

We also introduce two other variations by replacing the supervised attention mod-

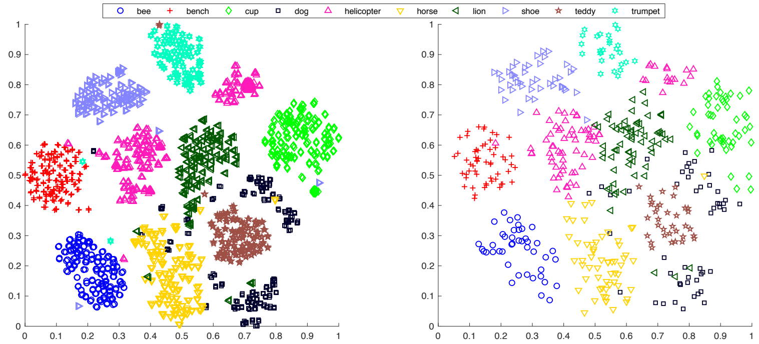


Figure 4.5: *t-SNE* visualization of our image and sketch embeddings of 10 representative categories from the Sketchy dataset. After embedding the images as well as the sketches to the common space by our model, natural images and sketch queries from most of the categories are almost scattered into the same clusters. (Best viewed in pdf)

ule with a global average pooling of image features (second row of Table 4.1) and a standard attention model without mask level supervision (third row) in order to show the impact of the supervised attention model in detecting salient objects. The comparison reveals the fact that supervised attended regions have a relevant impact on the results and therefore, are capable of discovering the discriminating regions, which facilitates the task of multi-label image classification. The global pooling basically provides no segregated object information and fails to generalize knowledge from the seen objects together to the unseen ones. The general attention is also suboptimal in exploring the object features individually.

Description	**Sketchy**
Regressed vector → one hot vector	68.80
Supervised attention → global pooling	72.60
Supervised attention → general attention	75.40
Full model	**80.90**

Table 4.1: Ablation analysis

Finally, in Fig. 4.6(d), we elucidate t-SNE [185] visualization of the image features and sketch features corresponding to 10 different categories. We can see that the distributions of the features in both domains are very similar reflecting that the network is capable to align features among both modalities.

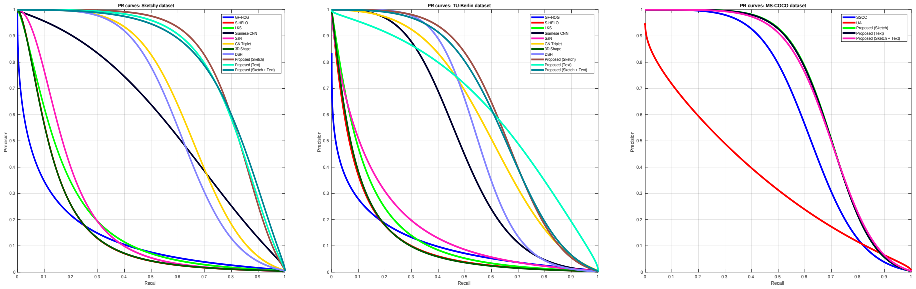
Methods	Sketchy		TU-Berlin	
	mAP	**Precision @ 200**	**mAP**	**Precision @ 200**
GF-HOG [78]	13.5	16.7	11.4	15.8
S-HELO [155]	16.1	18.1	12.1	15.3
LKS [157]	19.3	23.1	15.1	17.2
Siamese CNN [143]	58.7	74.5	48.9	62.3
SaN [224]	20.8	29.2	17.8	18.2
GN Triplet [160]	65.1	79.7	59.7	78.2
3D shape [192]	16.1	18.1	12.3	14.7
DSH [116]	62.0	69.4	55.6	74.3
Proposed (Sketch)	**80.9**	**88.6**	**65.3**	**79.6**
Proposed (Text)	**80.2**	**88.1**	**64.1**	**72.7**
Proposed (Sketch + Text)	**80.3**	**88.2**	**64.5**	**73.5**

Table 4.2: Image retrieval performance

4.4.4 Results and Discussions

Single Object: In Table 4.2, we report the comparison of mAP and precision@200 over all SBIR methods on two datasets. Generally, deep learning-based methods can achieve much better performance than handcrafted methods and the results on Sketchy are higher than those on TU-Berlin since the data in Sketchy is relatively simpler with fewer categories. The corresponding precision-recall (P-R) curves for both the datasets are illustrated in Fig. 4.6(a) and (b). Our method leads with significant improvements over the best-performing comparison methods on the two datasets, respectively in both mAP and precision@200. We argue that this is because our architecture is specifically designed to handle visual alignment of the query to the images using the Hungarian Loss.

Multiple Objects: For retrieving images with multiple objects, we have considered the MS-COCO dataset as mentioned above. Two existing methods are considered for comparison: SSCC [121] and UA [39]. However, it is worth mentioning that none of the above two methods works with multi-modal queries; as they allow retrieving images with multiple queries, we slightly modified these methods to accept multi-modal queries. The multi-modal multi-object image retrieval mean average precision (mAP@all) of our proposed method and the two baselines are reported in Table 4.3 and the corresponding precision-recall (P-R) curves are shown in Fig. 4.6(c). The performance margins between our proposed method and the selected state-of-the-art methods are significant, suggesting the existing cross-modal image retrieval methods fail to handle the multi-modal multi-object image retrieval task. SSCC [121] attains relatively better results. A possible reason for this is allowance of multiple queries and a relatively simple model for query processing. However, this method is designed only for text



(a) (b) (c)

Figure 4.6: Precision-recall curves obtained by different methods on (a) Sketchy, (b) TU-Berlin, (c) MS-COCO datasets. (Best viewed in pdf)

Methods	**MS-COCO**	
	mAP	**Precision @ 200**
SSCC [121]	62.3	66.7
UA [39]	35.4	41.3
Proposed (Sketch)	69.7	75.3
Proposed (Text)	69.6	75.1
Proposed (Sketch + Text)	**69.3**	**74.9**

Table 4.3: Multiple objects

modality and also deals with semantic constraints, which can be a reason of the worse performance than our proposed system. Some qualitative results of retrieving images using multi-modal and multi-object queries are Fig. 4.4. It can be seen that our proposed method is able to produce acceptable retrieval results. Albeit some false alarms are produced, they mostly have some visual similarity with the actual retrieval.

4.5 Conclusions and Future Work

In this chapter, we have proposed a common neural network model for sketch as well as text based image retrieval. One of the most important advantages of our framework is that it allows to retrieve images queried by terms of multiple modalities (text and sketch). We have designed an image attention mechanism based on LSTM that allows to put attention on the specific zones of the images depending on the inter related objects which usually co-occur in nature. This has been learned by our model from the images in the training set. We have tested our proposed framework on the challenging Sketchy dataset for single object retrieval and on a collection of images from the COCO dataset for multiple object retrieval. Furthermore, we have compared our experimental results with three state-of-the-art methods. We have found that our method performs satisfactorily better than the considered state-of-the-art methods on all the two datasets with some cases of failure with justifiable reasons. For the future we plan to investigate on more efficient training strategies, as few shot learning or zero shot learning approaches that learn from a small amount of training data in human-centered scenarios that allow users to search in their own databases in a more efficient and effortless manner. We also noticed while doing this work it not possible to train a model with all the object classes possible in real life scenario. Different training strategies are very important to be explored in order to figure out the best possible way to mimic human ways of depicting objects, combining them, and coming up with new concepts.

Chapter 5

Doodle to Search: Practical Zero-Shot Sketch-based Image Retrieval

Draw as if the object being drawn has never existed - because it hasn't.
– Scale & the Incas (2018), by Andrew Hamilton

In this chapter, we investigate the problem of zero-shot sketch-based image retrieval (ZS-SBIR), where human sketches are used as queries to conduct retrieval of photos from unseen categories. We importantly advance prior arts by proposing a novel ZS-SBIR scenario that represents a firm step forward in its practical application. The new setting uniquely recognizes two important yet often neglected challenges of practical ZS-SBIR, (i) the large domain gap between amateur sketch and photo, and (ii) the necessity for moving towards large-scale retrieval. We first contribute to the community a novel ZS-SBIR dataset, QuickDraw-Extended, that consists of 330,000 *sketches and* 204,000 *photos spanning across 110 categories. Highly abstract amateur human sketches are purposefully sourced to maximize the domain gap, instead of ones included in existing datasets that can often be semi-photorealistic. We then formulate a ZS-SBIR framework to jointly model sketches and photos into a common embedding space. A novel strategy to mine the mutual information among domains is specifically engineered to alleviate the domain gap. External semantic knowledge is further embedded to aid semantic transfer. We show that, rather surprisingly, retrieval performance significantly outperforms that of state-of-the-art on existing datasets that can already be achieved using a reduced version of our model. We further demonstrate the superior performance of our full model by comparing with a number of alternatives on the newly proposed dataset.*

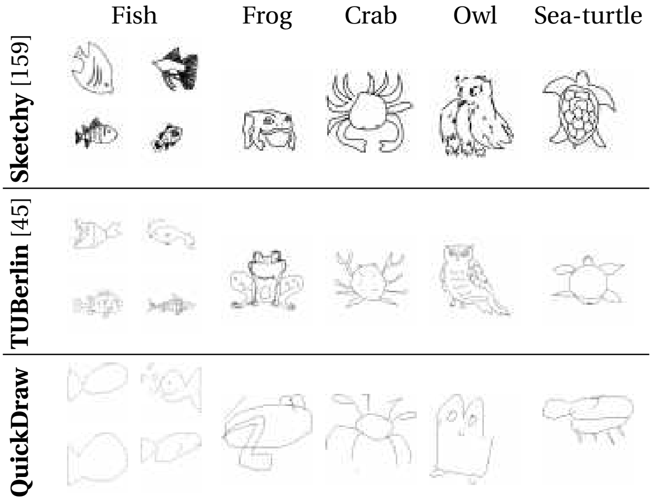


Figure 5.1: Qualitative comparison of sketch datasets, columns show examples belonging the same class. *Sketchy*, *TUBerlin* and *QuickDraw* datasets orderly contain sketches with increasing level of abstraction. It is worth noting that despite being the most abstract dataset, *QuickDraw* sketches can still be reliably recognised.

5.1 Introduction

In the context of retrieval, sketch modality has shown great promise thanks to the pervasive nature of touchscreen devices. Consequently, research on sketch-based image retrieval (SBIR) has flourished, with many great examples addressing various aspects of the retrieval process: fine-grained matching [224, 176, 143], large-scale hashing [116, 109], cross-modal attention [39, 176] to name a few.

However, a common bottleneck identified by almost all sketch researches is that of data scarcity. Different to photos that can be effortlessly crawled for free, sketches have to be drawn one by one by a human being. As a result, existing SBIR datasets suffer in both volume and variety, leaving only less than a thousand of sketches per category, with a maximum number of classes limited to a few hundreds. This largely motivated the problem of zero-shot SBIR (ZS-SBIR), where one wishes to conduct SBIR on object categories without having the training data. ZS-SBIR is increasingly being regarded as an important component in unlocking the practical application of SBIR, since million-scale datasets that have been used to train commercial photo-only systems [36] might not be feasible.

The problem of ZS-SBIR is extremely challenging. It shares all challenges laid out in conventional SBIR: (i) large domain gap between sketch and image, and (ii) high degree of abstraction found in human sketches as a result of variant drawing skills and visual interpretations. Additionally, it also needs the semantic transference from the seen to unseen categories for the purpose of zero-shot learning. Over and above all, in this paper, we are interested in moving towards the practical adaptation of ZS-SBIR technology. For that, a more appropriate dataset that best capture all these challenges is required.

Therefore, our first contribution in this chapter is a new dataset to simulate the real application scenario of ZS-SBIR, which should satisfy the following requirements. First, the dataset needs to mimic the real-world abstraction gap between sketch and photo. Such amateur sketches are very different from the ones currently studied by existing datasets, which are either too photo-realistic [50] or produced by recollection of a reference images [159] (Figure 5.1 offers a comparative example). Second, in order to learn a reliable cross-domain embedding between amateur sketch and photo, the dataset much faithfully capture of a full variety of sketch samples from users having various drawing skills. Our proposed dataset, *QuickDraw-Extended*, contains 330,000 sketches and 204,000 photos in total spanning across 110 categories. In particular, it includes 3,000 amateur sketches per category carefully sourced from the recently released Google Quickdraw dataset [88] – six times more than the next largest. It also has a search space stretching to 166million total comparisons in the test set, compared to *Sketchy-Extended* and *TUBerlin-Extended* with just 10 million and 1.9 million, respectively.

This dataset and the real-world scenario it mimics, essentially make the ZS-SBIR task more difficult. This leads to our second contribution which is a novel cross-domain zero-shot embedding model that addresses all challenges posed by this new setting. Our base network is a visually-attended triplet ranking model that is commonly known in the SBIR community to produce state-of-the-art retrieval performances [224, 176]. To our surprise, just by adopting such a triplet formulation, we can already achieve retrieval performances drastically better than that of the previously reported ZS-SBIR results on commonly used datasets. We attribute this phenomena to previous datasets being too simplistic in terms of the cross-domain abstraction gap and the diversity of sketch samples. This further justifies the necessity of a new practical dataset like ours. We then propose two novel techniques to help learn a better cross-domain transfer model. First, a domain disentanglement strategy is designed to bridge the gap between the domains by forcing the network to learn a domain-agnostic embedding, where a *Gradient Reversal Layer* (GRL) [61] encourages the encoder to extract mutual information from sketches and photos. Second, a novel semantic loss to ensure that semantic information is preserved in the obtained embedding. By applying a GRL only to the negative samples at the input of the semantic decoder helps the encoder network to separate the semantic information of similar classes.

Extensive experiments are first carried out on the two commonly used ZS-SBIR datasets, TUBerlin-Extended [45] and Sketchy-Extended [159]. The results show that the even a reduced version of our model can outperform current state-of-the-arts by a significant margin. The superior performance of the proposed method is further validated on our own dataset, with ablative studies to draw insights towards each of the proposed system components.

The rest of the chapter is organized as follows: in Sect. 5.2, we review the relevant state-of-the-art. Sect. 5.3 describes the different dataset available and the how-to its important to have a new dataset to benchmark the ZS-SBIR problem. Sect. 5.4 describes in detail our proposed zero-shot image retrieval framework. In Sect. 5.5, we describe the experimental framework and present the results of the experiments and provide necessary insights to pinpoint the main difficulties in solving this problem. Finally, Sect. 5.8 draws the conclusions and outlines the future directions.

5.2 Related Work

5.2.1 SBIR Datasets

One of the key barriers towards large-scale SBIR research is the lack of appropriate benchmarks. The Sketchy dataset [159] is the most used one for this purpose, which contains 75,471 hand-drawn sketches of 12,500 object photos belonging to 125 different categories. Later, Liu *et al.* [116] collected 60,502 natural images from ImageNet [36] in order to fit the task of large-scale SBIR. This dataset having contained highly detailed or less abstract sketches, models trained on Sketchy have high chance of getting collapsed in real-life scenario. Two more fine-grained SBIR datasets with paired sketches and images are *shoe* and *chair* datasets which were proposed in [224]. The *shoe* dataset contains altogether 6648 sketches and 2000 photos, whereas, the *chair* dataset altogether contains 297 sketches and photos. However, being fine-grained pairs these two datasets also have similar disadvantages as the Sketchy dataset. TU-Berlin [45] being the other popular dataset originally contains 250 classes of hand-drawn sketches, where each class roughly contains 80 instances. It was extended with real images by [227] for SBIR purposes. This dataset has a lot of confusion regarding the class hierarchy, for an example, `swan`, `seagull`, `pigeon`, `parrot`, `duck`, `penguin`, `owl` have substantial visual similarity and commonality with `standing bird` and `flying bird` which are another separate categories of the TU-Berlin dataset. To obliterate, these difficulties faced by the SBIR work, in this paper, we introduce QuickDraw-Extended dataset, where we take the sketch classes of the Google QuickDraw dataset [88] and provide the corresponding set of images to facilitate the training of large-scale SBIR system.

5.2.2 SBIR

The main challenge that most of the SBIR tasks address is bridging the domain gap between sketch and natural image. In literature, these existing methods can be roughly grouped into two categories: *hand crafted* and *cross-modal deep learning* methods. The hand-crafted techniques mostly work with Bag-of-Words representations of sketch and edge map of natural image on top of some off-the-shelf features, such as, SIFT [119], Gradient Field HOG [79], Histogram of Edge Local Orientations [155] or Learned Key Shapes [157]) etc. This domain shift issue is further addressed by cross-domain deep learning-based methods [159, 224], where they have used classical ranking losses, such as, contrastive loss, triplet loss [195] or more elegant HOLEF loss [176] within a siamese like network. Based on the problem at hand, two separated tasks have been identified: (1) *Fine-grained SBIR* (FG-SBIR) aims to capture fine-grained similarities of sketch and photo [110, 159, 224] and (2) *Coarse-grained SBIR* (CG-SBIR) performs a instance level search across multiple object categories [227, 79, 84, 192, 227], which has received a lot of attention due to its importance. Realising the need of large-scale SBIR, some researchers have proposed a variant of cross-modal hashing framework for the same [116, 229], which also showed promising results in SBIR scenario. In contrast, our proposed model overcomes this domain gap by mining the modality agnostic features using a domain loss along with a GRL.

5.2.3 Zero-Shot Sketch-based Image Retrieval (ZS-SBIR)

Early works on zero-shot learning (ZSL) were mostly focused on attribute based recognition [100], which is later augmented by another major line that focus on learning a joint embedding space for image feature representation and class semantic descriptor [23, 215, 89, 218, 118]. Depending on the selection of joint embedding space and type of projection function utilised between the visual to semantic space, existing models can be divided into three groups: (i) projected from visual feature space to semantic space [100, 131], (ii) projected from semantic space to the visual feature space [23], and (iii) an intermediate space that both are simultaneously projected to [230]. In contrast to these existing works, our model can be seen as a combination of the first and second groups, where the embedding is on the visual feature space, but asked to additionally recover its embodied semantics with a decoder.

Although SBIR and ZSL have been extensively studied among the research community, very few works have studied their combination. Shen *et al.* [170] propose a multi-modal network to mitigate the sketch-image heterogeneity and enhance semantic relations. Yelamarthi *et al.* [219] resort to a deep conditional generative model, where a sketch is taken as input and learned to generate its photo features by stochastically filling the missing information. The main motivation behind ZS-SBIR lies with sketches being costly and labour-intensive to source – sketches need to be individually drawn by hand, other than crawled for free from the internet. To enable rapid deployment on categories where training sketches are not readily available, it is important to leverage on existing sketch data from other categories. The key difference between ZS-SBIR and other ZS tasks, which is also the main difficulty of the problem, lies with the additional modality gap between sketch and photo.

5.3 QuickDraw-Extended Dataset

Existing datasets do not cover all the challenges derived from a ZS-SBIR system. Therefore, we propose a new dataset named *QuickDraw-Extended Dataset* that is specially designed for this task. First we review the existing datasets in the literature used for ZS-SBIR and motivate the purpose of the new dataset.

Thus, we provide a large-scale ZS-SBIR dataset that overcomes the main problems of the existing ones.

Existing datasets were not originally designed for a ZS-SBIR scenario, but they have been adapted by a redefining the partitions setup. In addition, the main limitations that we overcome with the new dataset are (i) the large domain gap between amateur sketch and photo, and (ii) the necessity for moving towards large-scale retrieval.

5.3.1 Sketchy-Extended Dataset

Originally created as a fine-grained association between sketches to particular photos for fine-grained retrieval [159]. This dataset has been adapted to the task of ZS-SBIR. On one hand, Shen *et al.* [170] proposed to set aside 25 random classes as a test set whereas the training is performed in the rest 100 classes. On the other hand, Yelamarthi *et al.* [219] proposed a different partition of 104 train classes and 21 test classes in order to make sure that test is not present in the 1,000 classes of ImageNet.

Its main limitation for the task of ZS-SBIR is its fine-grained nature, i.e., each sketch has a corresponding photo that was used as reference at drawing time. Thus, participants tended to draw the objects in a realistic fashion, producing sketches resembling that of a true edge-map very well. This essentially narrows the cross-domain gap between sketch and photo.

5.3.2 TUBerlin-Extended Dataset

It is a dataset [45] that was created for sketch classification and recognition benchmarking. In this case, drawers were asked to draw the sketches giving them only the name of the class. This allows a semantic connection among sketches and avoids possible biases. However, the number of sketches is scarce, considering the variability among the observations of a concept in the real world. Also, some of the design decisions on the selection of object categories prevent it to be adequate for our zero-shot setting: (i) classes are defined both in terms of a concept and an attribute (e.g., `seagull`, `flying-bird`); (ii) different WordNet levels are used, *i.e.* there are classes that are semantically included in others (e.g., `mug`, `beer-mug`).

Table 5.1: Dataset comparison in terms of their size. Partition is presented in terms of number of classes used for each set, moreover, # Comparisons stands for the number of comparisons sketch-image performed in test.

	Sketchy [159]	**TUBerlin** [45]	**QuickDraw**
Partition (tr+va, te)	(104, 21)	(220, 30)	(80, 30)
# Sketch/class	500	80	3,000
# Image/class	600-700	~ 764[a]	~ 1,854
# Comparisons	~ 10Mill.	~ 1.9Mill.	~ 166Mill.

[a] Extremely imbalanced

5.3.3 The Dataset

Taking into account the limitations of the previously described datasets in a ZS-SBIR scenario, we contribute to the community a novel large-scale dataset, *QuickDraw-Extended*. We identified the following challenges of a practical ZS-SBIR, (i) the large domain gap between amateur sketch and photo, and (ii) the necessity for moving towards large-scale retrieval. According to this, the new dataset must fulfil the following aspects: (i) to not have a direct one-to-one correspondence between sketches and images, *i.e.* sketches can be rough conceptual abstractions of images produced in an amateur drawing style; (ii) to avoid ambiguities and overlapping classes; (iii) large intra-class variability provided by the high abstraction level of different drawers.

In order to accomplish these objectives, we took advantage of the Google Quick, Draw! [88] data which is a huge collection of drawings (50 millions) belonging to 345 categories obtained from the *Quick, Draw!*[3] game. In this game, the user is asked to draw a sketch of a given category while the computer tries to classify them. The way sketches are collected provides the dataset a large variability, derived from human abstraction. Moreover, it addresses the large domain gap between non-expert drawers and photos that is not considered in previous benchmarks. Hence, we propose to make use of a subset of sketches to construct a novel dataset for large-scale ZS-SBIR containing 110 categories (80 for training and 30 for testing). Classes such as `circle` of `zigzag` are directly discarded because they can not be used in an appropriate SBIR. As a retrieval gallery, we provide images extracted from *Flickr* tagged with the corresponding label. Manual filtering is performed to remove outliers. Moreover, following the idea introduced in [219] for the *Sketchy-Extended* dateset, we provide a test split which forces that test classes are not present in ImageNet in case of using pre-trained models. Finally, this dataset consists of 330,000 sketches and 204,000 photos moving towards a large-scale retrieval. We consider that this dataset will provide better insights about the real performance of ZS-SBIR in a real scenario.

[3] `https://quickdraw.withgoogle.com/`

Table 5.1 provides a comparison of the three benchmarks for the task of ZS-SBIR. To the best of our knowledge, this is the first time that a real large-scale problem is addressed providing 6 times more sketches and more than the double of photos per each class. Qualitatively QuickDraw-Extended provides a high abstraction level than previous benchmarks as it is shown in Figure 5.2.

5.4 A ZS-SBIR framework

5.4.1 Problem Formulation

Let $\mathscr{C}$ be the set of all possible categories in a given dataset; $\mathscr{X} = \{x_i\}_{i=1}^N$ and $\mathscr{Y} = \{y_i\}_{i=1}^M$ be the set of photos and sketches respectively; $l_x : \mathscr{X} \to \mathscr{C}$ and $l_y : \mathscr{Y} \to \mathscr{C}$ be two labelling functions for photos and sketches respectively. Such that give an input sketch an optimal ranking of gallery images can be obtained. In a *zero-shot* framework, training and testing sets are divided according to *seen* $C^s \subset \mathscr{C}$ and *unseen* $C^u \subset \mathscr{C}$ categories, where $\mathscr{C}^s \cap \mathscr{C}^u = \emptyset$. Thus, the model needs to learn an aligned space between sketches and photos to perform well on test data whose classes have never been used in training. We define the set of *seen* and *unseen* photos as $\mathscr{X}^s = \{x_i; l_x(x_i) \in \mathscr{C}^s\}_{i=1}^N$ and $\mathscr{X}^u = \mathscr{X} \setminus \mathscr{X}^s$. We define analogously the *seen* and *unseen* sets for sketches, denoted as $\mathscr{Y}^s$ and $\mathscr{Y}^u$.

The proposed framework is divided in two main components. The encoder transforms the input image to the corresponding embedding space. The second component is the cost function which guides the learning process to provide the embedding with the desired properties. Figure 5.3 outlines the proposed approach.

5.4.2 Encoder Networks

Given a distance function $d(\cdot,\cdot)$, the aim of our framework is to learn two embedding functions $\phi : \mathscr{X} \to \mathbb{R}^D$ and $\psi : \mathscr{Y} \to \mathbb{R}^D$ which respectively map the photo and sketch domain into a common embedding space. Later, these embedding functions are used in the retrieval task during the test phase, therefore, they should possess a ranking property related to the considered distance function. Hence, given two photos $x_1, x_2 \in \mathscr{X}$ and a sketch $y \in \mathscr{Y}$, we expect the embedding fulfils the following condition: $d(\phi(x_1), \psi(y)) < d(\phi(x_2), \psi(y))$, when $l_x(x_1) = l_y(y)$ and $l_x(x_2) \neq l_y(y)$. In a retrieval scenario, our system is able to provide a ranked list of images by the chosen distance function. In this framework, d has been set as ℓ_2-distance. During training, the two embedding $\phi(\cdot)$ and $\psi(\cdot)$ are trained with multi-modal information, therefore they presume to learn a modality free representation.

Our embedding functions $\phi(\cdot)$ and $\psi(\cdot)$ are defined as two CNNs with attention where the last fully-connected layer has been replaced to match the desired embedding size D. The *attention* [211] mechanism helps our system to localise the important

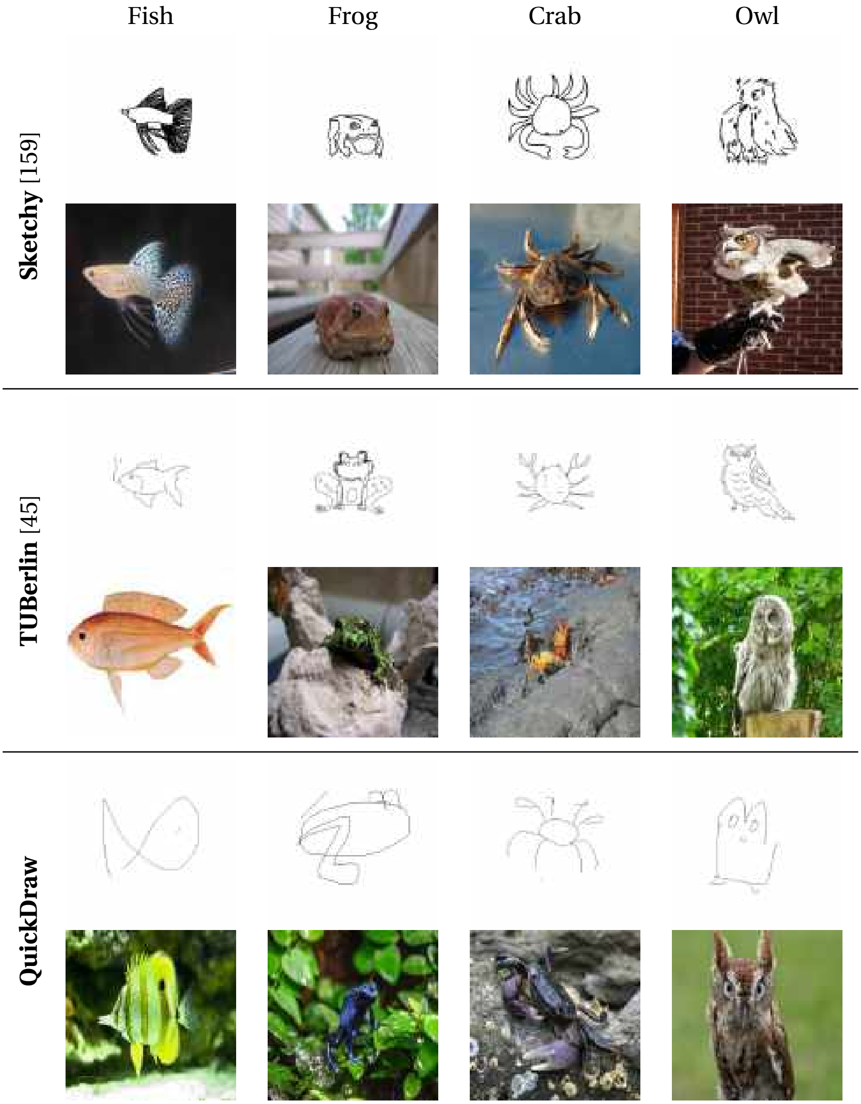


Figure 5.2: Qualitative comparison of the datasets. The different levels of abstraction in the sketches can be appreciated. From the top to the bottom, the figure also shows the decrease in the alignment between sketches and images.

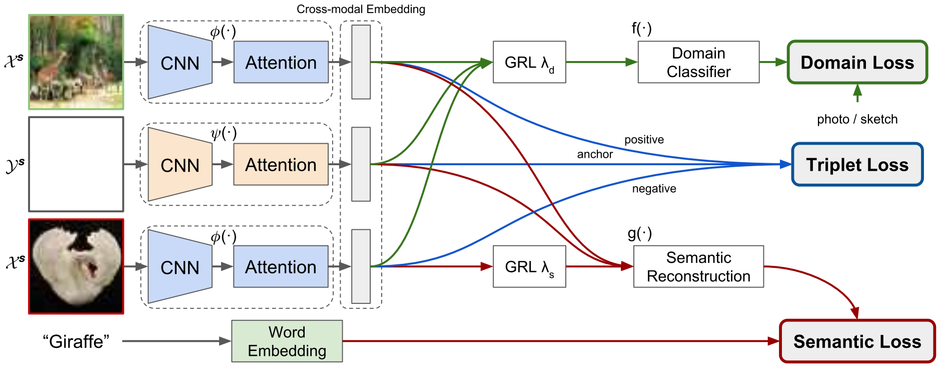


Figure 5.3: Proposed architecture for ZS-SBIR which maps sketches and photos in a common embedding space. It combines three losses: (i) triplet loss, to learn a ranking metric; (ii) domain loss to merge images and sketches to an indistinguishable space making use of a GRL; (iii) semantic loss forces the embeddings to contain semantic information by reconstructing the word2vec embedding of the class. It also helps to distinguish semantically similar classes by means of a GRL on the negative example (best viewed in color).

features in both modalities. Soft-attention is the widely used one because it is differentiable, and hence it can be learned end-to-end with the rest of the network. Our soft-attention model learns an attention mask which assigns different weights to different regions of an image given a feature map. These weights are used to highlight important features, therefore, given an attention mask att and a feature map f, the output of the attention module is computed by $f + f \cdot att$. The attention mask is computed by means of 1 × 1 convolution layers applied on the corresponding feature map.

5.4.3 Learning objectives

The learning objective of the proposed framework combines: (i) *Triplet Loss*; (ii) *Domain Loss*, (iii) *Semantic Loss*. These objective functions provide visual and semantic information to the encoder network. Let us consider a triplet $\{a, p, n\}$ where $a \in \mathcal{Y}^s$, $p \in \mathcal{X}^s$ and $n \in \mathcal{X}^s$ are respectively the *anchor*, *positive* and *negative* samples during the training. Moreover, $l_x(p) = l_y(a)$ and $l_x(n) \neq l_y(a)$.

Triplet Loss

This loss aims to reduce the distance between embedded sketch and image if they belong to the same class and increase it if they belong to different classes. For simplicity, if we define the distances between the samples as $\delta_+ = \left\|\psi(a) - \phi(p)\right\|_2$ and $\delta_- = \left\|\psi(a) - \phi(n)\right\|_2$ for the *positive* and *negative* samples respectively, then, the ranking loss for a particular triplet can be formulated as $\lambda(\delta_+, \delta_-) = \max\{0, \mu + \delta_+ - \delta_-\}$ where $\mu > 0$ is a margin parameter. Batch-wise, the loss is defined as:

$$\mathscr{L}_t = \frac{1}{N} \sum_{i=1}^{N} \lambda(\delta_+^i, \delta_-^i). \tag{5.1}$$

This loss measures the violation of the ranking order of the embedded features. Therefore, the order aimed by this loss is $\delta_- > \delta_+ + \mu$, if this is the case, the network is not updated, otherwise, the weights of the network are updated accordingly. Triplet loss provides a metric space with ranking properties based on visual features.

Domain Loss

Triplet loss mentioned above does not explicitly enforce the mapping of sketch and image samples to a common space. Therefore, at this end, to ensure that the obtained embedding belong to the same space, we propose to use a domain adaptation loss [61]. The basic idea of this loss is to obtain a domain-agnostic embedding that does not contain enough information to decide whether it comes from a sketch or photo. Given the embedding $\phi(\cdot)$ and $\psi(\cdot)$, we make use of a Multilayer Perceptron (MLP) as a binary classifier trying to predict which was the initial domain. Purposefully, in order to create indistinguishable embedding we use a *GRL* defined as $R_\lambda(\cdot)$, which applies the identity function during the forward pass $R_\lambda(x) = x$, whereas during the backward pass it multiplies the gradients by the meta-parameter $-\lambda$, $\frac{\mathrm{d}R_\lambda}{\mathrm{d}x} = -\lambda I$. This operation reverses the sign of the gradient that flows through the CNNs. In this way, we encourage our encoders to extract the shared representation from sketch and photo. For this loss, we define a meta-parameter λ_d that changes from 0 (only trains the classifier but does not update the encoder network) to 1 during the training according to a defined function. In our case it is defined according to the iteration i as $z_\lambda(i) = (i-5)/20$. Following the notation, $f : \mathbb{R}^D \to [0,1]$ be the MLP and $e \in \mathbb{R}^D$ an embedding coming from the encoders network. Then we can define the binary cross entropy of one of the samples as $l_t(e) = t\log(f(R_{\lambda_d}(e))) + (1-t)\log(1 - f(R_{\lambda_d}(e)))$, where e is the embedding obtained by the encoder network and t is 0 and 1 for sketch and photo domains respectively. Hence, the domain loss is defined as:

$$\mathscr{L}_d = \frac{1}{3N} \sum_{i=1}^{N} \big(l_0(\psi(a_i)) + l_1(\phi(p_i)) + l_1(\phi(n_i)) \big) \tag{5.2}$$

Semantic Loss

A decoder network trying to reconstruct the semantic information of the corresponding category from the generated embedding is proposed. This reconstruction forces that the semantic information is encoded in the obtained embedding. In this case, we propose to minimise the cosine distance with the reconstructed feature vector and the semantic representation of the category. Inspired by the idea presented by Gonzalez *et al.* [66] for cross-domain disentanglement, we propose to exploit the negative sample to foster the difference between similar semantic categories. Hence, we apply a GRL $R_{\lambda_s}(\cdot)$ to the negative sample at the input of the semantic decoder and we train it to reconstruct the semantics of the positive example. The idea is to help the encoder network to separate the semantic information of similar classes. In this case, we decided to keep the meta-parameter λ_s to a fixed value among all the training, in particular, it was set to 0.5.

Let $c \in \mathscr{C}^s$ be the corresponding category of the anchor a. The semantics of this category are obtained by the *word2vec* [124] embedding trained on part of Google News dataset (~ 100 billion words), *GloVe* [141] and *fastText* [11] (more results are available in supplementary materials). Let $g : \mathbb{R}^D \rightarrow \mathbb{R}^{300}$ be the semantic reconstruction network and $s = \text{embedding}(c) \in \mathbb{R}^{300}$ be the semantics of the given category. Hence, given an image embedding $e \in \mathbb{R}^D$ the cosine loss is defined as $l_c(e, s) = \frac{1}{2}\left(1 - \frac{g(e)s^t}{||g(e)||\cdot||s||}\right)$. The semantic loss is defined as follows:

$$\begin{aligned}\mathscr{L}_s = \frac{1}{3N}\sum_{i=1}^{N}\big(&l_c(\psi(a_i), s_i) + l_c(\phi(p_i), s_i)\\ &+ l_c(R_{\lambda_s}(\phi(n_i)), s_i)\big) \qquad (5.3)\end{aligned}$$

Therefore, the whole network will be trained by a combination of three proposed loss functions.

$$\mathscr{L} = \alpha_1\mathscr{L}_t + \alpha_2\mathscr{L}_d + \alpha_3\mathscr{L}_s, \qquad (5.4)$$

where the weighting factors α_1, α_2 and α_3 are equal in our model. Algorithm 1 presents the training algorithm followed in this work. $\Gamma(\cdot)$ denotes the optimiser function.

Algorithm 1 Training algorithm for the proposed model .

```
Input: Photo-Sketch data $\{\mathscr{X}, \mathscr{Y}\}$; Class semantics $\mathscr{S}$;
    $\lambda_s = 0.5$ and max training iterations $T$
Output: Encoder networks parameters $\{\Theta_\phi, \Theta_\psi\}$.
repeat
    Get a random mini-batch $\{y_i, x_i^p, x_i^n, s_i\}_{i=1}^{N_B}$; where
        $y_i$, $x_i^p$ belong to the same class and $x_i^n$ does not.
    $\lambda_d \leftarrow \text{clip}(z_\lambda(\cdot), \text{min} = 0, \text{max} = 1)$
    $\mathscr{L} \leftarrow$ Eq. 5.4
    $\Theta \leftarrow \Theta - \Gamma(\nabla_\Theta \mathscr{L})$
until Convergence or max training iterations $T$
```

Table 5.2: Comparison against the state-of-the-art with that of the proposed model. Note: the same train and test split are used for all experiments on CVAE [219] and ours. ZSIH [170] did not report the specific details on their split (other than 25 classes were used for testing), and we could not produce their results on *QuickDraw-Extended* due to the lack of publicly available code.

Method	**Sketchy-Extended** [159]			**TUBerlin-Extended** [45]			**QuickDraw-Extended**		
	mAP	mAP@200	P@200	mAP	mAP@200	P@200	mAP	mAP@200	P@200
ZSIH [170]	0.2540[a]	–	–	**0.2200**	–	–	Not able to produce		
CVAE [219]	0.1959	0.2250	0.3330	0.0050	0.0090	0.0030	0.0030	0.0060	0.0030
Ours	**0.3691**	**0.4606**	**0.3704**	0.1094	**0.1568**	**0.1208**	**0.0752**	**0.0901**	**0.0675**

[a] Using a random partition of 25 test categories following the setting proposed in [26], we obtained 0.3521 for our model.

Table 5.3: Ablation study for the proposed model. As baseline, the triplet loss is used and the different modules are incrementally added.

Attn.	**Dom.**	**Sem.**	**Sketchy-Extended** [159]			**TUBerlin-Extended** [45]			**QuickDraw-Extended**		
			mAP	mAP@200	P@200	mAP	mAP@200	P@200	mAP	mAP@200	P@200
-	-	-	0.3020	0.3890	0.3091	0.0590	0.1040	0.0682	0.0354	0.0546	0.0454
✓	-	-	0.3207	0.4150	0.3342	0.0729	0.1141	0.1002	0.0456	0.0635	0.0496
✓	✓	-	0.3256	0.4113	0.3444	0.0845	0.1264	0.1080	0.0651	0.0881	0.0615
✓	-	✓	0.3392	0.4146	0.3586	0.1055	0.1496	0.1115	0.0693	0.0896	0.0625
✓	✓	✓	**0.3691**	**0.4606**	**0.3704**	**0.1094**	**0.1568**	**0.1208**	**0.0752**	**0.0901**	**0.0675**

5.5 Experimental Validation

This Section experimentally validates the proposed ZS-SBIR approach on three benchmarks *Sketchy-Extended*, *TUBerlin-Extended* and *QuickDraw-Extended*, highlighting the importance of the newly introduced dataset which is more realistic for practical SBIR purpose. A detailed comparison with the state-of-the-art is also presented.

5.5.1 Zero-shot Experimental Setting

Implementation details

Our CNN-based encoder networks $\phi(\cdot)$ and $\psi(\cdot)$ make use of a ImageNet pre-trained VGG-16 [171] architecture. This can be replaced by any model to enhance the extracted feature quality. Both, domain classifier $f(\cdot)$ and semantic reconstruction $g(\cdot)$ of the proposed model makes use of 3 fully connected layers with ReLU activation functions. The whole framework was implemented with PyTorch [138] deep learning tool and is trainable on single Pascal Titan X GPU card.

Training setting

Our system uses triplets to utilise the inherent ranking order. The training batches are constructed in a way so that it can take the advantage of the semantic information in order to mine hard negative samples for a given anchor class. This implies that semantically closer classes will have a higher probability to be used during training and thus they are likely to be disjoint in the final embedding. We trained our model following an early stopping strategy in validation to provide the final test result. The model is trained end-to-end using the SGD [12] optimiser. The learning rate used throughout is $1e-4$. The epochs required to train the model on different dataset is around 40.

Evaluation protocol

The proposed evaluation uses the metrics used by Yelamarthi *et al.* [219]. Therefore, the evaluation is performed taking into account the top 200 retrieved samples. Moreover, we also provide metrics on the whole dataset. Images labelled with the same category as that of the query sketch, are considered as relevant. Note that this evaluation does not consider visually similar drawings that can be considered correct by human users. For the existing datasets, we used the proposed splits in [219, 170].

5.5.2 Model Discussion

This section presents a comparative study with the state-of-the-art followed by a discussion on the *TUBerlin-Extended* results and finally the ablative study. As mentioned,

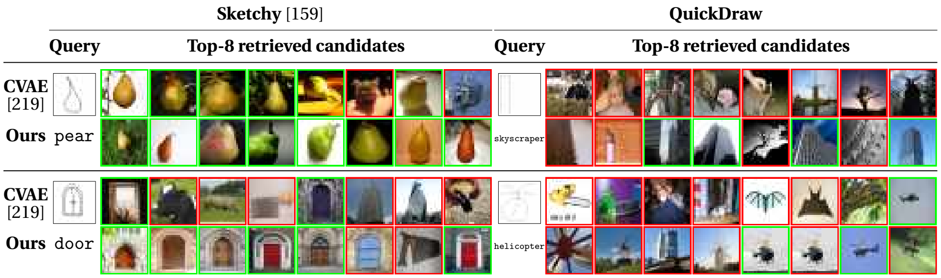


Figure 5.4: Top 8 image retrieval examples given a query sketch. All the examples correspond to a zero-shot setting, *i.e.* no example have been seen in training. First row provides a comparison with CVAE [219] method against our pipeline. Note that in some retrieval cases, for instance, `door` is confused with `window` images which can be true even for humans. Green and Red stands for correct and incorrect retrievals. (Better viewed in pdf)

our model is build on top of a triplet network. We take this as a baseline and study the importance of the different components of the full model which includes the *attention mechanism*, the *semantic loss* and the *domain loss*.

Comparison

Table 5.2 provides comparisons of our full model results against those of the state-of-the-art. We report a comparative study with regard to two methods presented in Section 5.2, namely ZSIH [170] and CVAE [219]. Note that we have not been able to reproduce the ZSIH model due to lack of technical implementation details and the code being unavailable. Hence, the results on *QuickDraw-Extended* dataset nor an evaluation using the top 200 retrieval could be computed. The last row of the Table 5.2 shows the result of our full model. From the Table 5.2 the results suggest the limitation of the previous models regarding their ability in an unconstrained domain where sketches have higher level of abstractions. The CVAE [219] method trained with sketch-image correspondence has difficulties to capture the intra-class variability, the domain gap and also the ability to infer unseen classes. The following conclusions are drawn: (i) our base model outperforms all the state-of-the-art methods in *Sketchy-Extended* Dataset; (ii) our model performs the best overall on each metric and on almost all the datasets; (iii) the gap between our model and the state-of-the-art datasets is almost double in *Sketchy-Extended* Dataset; (iv) the difference in the result in previous dataset points out the need of a new well structured dataset for ZS-SBIR (v) the new benchmark also provides the different aspects (*i.e* of semantics, mutual information) that can play important role in a real ZS-SBIR scenario; (vi) the evaluation shows the importance of going towards large-scale ZS-SBIR where the retrieval search space is in the range of 166 million comparisons (16 times of the current largest dataset).

Discussion on *TUBerlin-Extended*

As stated in Section 5.3, the results could be heavily affected by the chosen classes for experiments. Since [170] did not report specific details on their train and test split, we can not offer a fair comparison on TUBerlin-Extended. Instead, for both [219] and ours, we resort to the commonly accepted median over random splits setting. And it shows our method favourably beats [219] by a clear margin. We did however observe a high degree of fluctuation over the different splits on TUBerlin-Extended, which reaffirms our speculation on how the categories included in TUBerlin-Extended might not be optimal for the zero-shot setting (see Section 5.3). This could explain the superior performance of [170], yet more experiments are needed to confirm such suspicion. Unfortunately, again such experiments would not be possible without details on their train and test split.

Ablation study

Here, we investigate the contribution of each component to the model, as well as other issues of the architecture.

The first 5 rows of Table 5.3 present a study of the contribution of each component to the whole proposed model. From this Table we can draw the following conclusions: (i) attention plays a major role in improving the baseline result; (ii) the domain loss is able to alleviate to some extend the domain gap, this is more remarkable in those datasets where sketches are more abstract; (iii) as the difficulty of the dataset increases, the semantic and the domain losses start playing a major role in improving the baseline result; (iv) semantics provide better extrapolation to unseen data than domain loss which shows that either the mutual information is very less or that the semantic information is really needed in this extrapolation; (v) the poor performance in the *QuickDraw-Extended* dataset shows that the practical problem of ZS-SBIR is still indeed unsolved. It should be noted, that the best model makes use of the three losses.

5.5.3 Qualitative

Some retrieval results are shown in Figure 5.4 for *Sketchy-Extended* and *QuickDraw-Extended.* We also provide a qualitative comparison with CVAE proposed by Yelamarthi *et al.* [219]. The qualitative results reinforce that the combination of semantic, domain and triplet loss fairs well in a dataset with substantial variances on visual abstraction. We would also like to point out that the retrieved results for the class `skyscraper` show high visual shape similarity with rectangle *i.e.* `door` and `saw`. The retrieved circular `saw` could also might be retrieved because of the semantic rather than the visual similarity. Similar visual correspondences can also be noticed between the query sketch `helicopter` and the retrieved result `windmill`.

5.6 Further Experimentation

In this section further experiments have been done in order to validate the parameters of the proposed architecture and the choice of the reported results in-case of *TUBerlin-Extended*.

5.6.1 Embedding Size

Table 5.4 presents the results obtained changing the final embedding size. The results are presented in the *Sketchy-Extended* dataset. It is a well accepted dataset without confusing categories as in the case of TUBerlin-Extended [45]. The mAP on the full database reinforces our choice of 64 dimension as mentioned in Section 5 of the original paper. Though the mAP@200 has a very marginal improvement in case of lower dimension, we would like to keep a eye on the overall mAP as we are looking towards a large scale zero-shot image retrieval. The slight change in the mAP@200 can be credited to the non-deterministic behaviour of ranking-based metrics [130].

Table 5.4: Comparison against different sizes of embedding of our method on *Sketchy-Extended* dataset.

Dimension	**Sketchy-Extended** [45]		
	mAP	mAP@200	P@200
128	35.08	45.99	36.75
64	**36.91**	46.06	**37.04**
32	35.96	**46.91**	37.01

5.6.2 Leakage of class information

We previously followed common practice in [219], where an off-the-shelf word embedding was used without re-training. Yet, we do fully acknowledge the need to ensure no class information is leaked during training. For that, we first re-trained word2vec from scratch without the test classes from Sketchy-Extended, and obtained 36.1 mAP, which is slightly worse than the previously reported mAP of 36.9. We further trained two alternative word embeddings from scratch – GloVe and Fasttext – and report results in Table 5.5. It shows GloVe being superior to word2vec and Fasttext.

Table 5.5: Comparison against different sizes of embedding of our method on *Sketchy-Extended* dataset.

Dataset	**word2vec** [125]	**GloVe** [141]	**fastText** [11]
Sketchy [159]	36.9	40.1	33.1
TU-Berlin [45]	10.9	11.8	9.70

5.6.3 Discussion on *TUBerlin-Extended*

Table 5.6 shows the results we obtain in five different dataset split in *TUBerlin-Extended* [45] as discussed in the paper in Section 5. This in a way shows that the different splits in this dataset can have a huge effect on the zero-shot retrieval results. If seen carefully the same method performs much better in `Split-5` as compared to that in `Split-2` and `Split-3`.

Table 5.6: Performance on different dataset split in *TUBerlin-Extended* of our method.

Random Split	**TUBerlin-Extended** [159] mAP
Split-1	11.84
Split-2	5.3
Split-3	4.26
Split-4	15.74
Split-5	10.94
Median	**10.94**

5.7 Extra Qualitative Results

This section introduces an extended qualitative study on the three datasets. We choose 7 queries and their corresponding top-10 retrieval images. The chosen queries are the ones which best illustrates the results on the different parts of our method. These sketch queries were selected among approximately 12600, 2400 and 90000 test sketches in *Sketchy-Extended*, *TUBerlin-Extended* and *QuickDraw-Extended* dataset respectively.

Figure 5.5 presents the results for *Sketchy-Extended*. The results shows that our method performs really best due to the close correspondence of sketches and images. Like-wise the other state-of-the-art method [219] also performs good. Triplet loss helps our method to find an embedding space where the domain agnostic fea-

Figure 5.5: Top-10 image retrieval examples for *Sketchy-Extended* [159]. All the examples correspond to a zero-shot setting. First row provides a comparison with CVAE [219] method against our pipeline. Green and Red stands for correct and incorrect retrievals. (Better viewed in pdf)

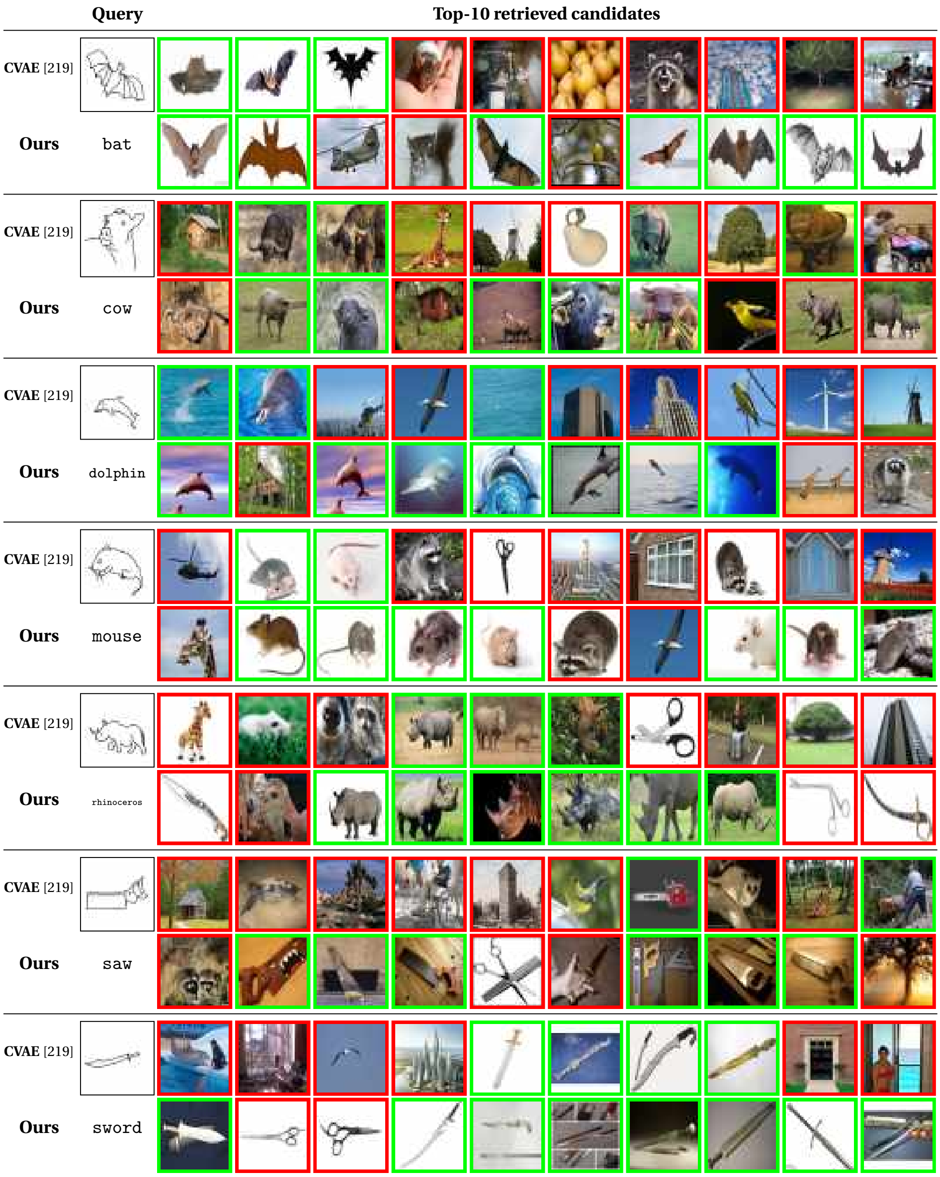


Figure 5.6: Top-10 image retrieval examples for *TUBerlin-Extended* [45]. All the examples correspond to a zero-shot setting. First row provides a comparison with CVAE [219] method against our pipeline. Green and Red stands for correct and incorrect retrievals. (Better viewed in pdf)

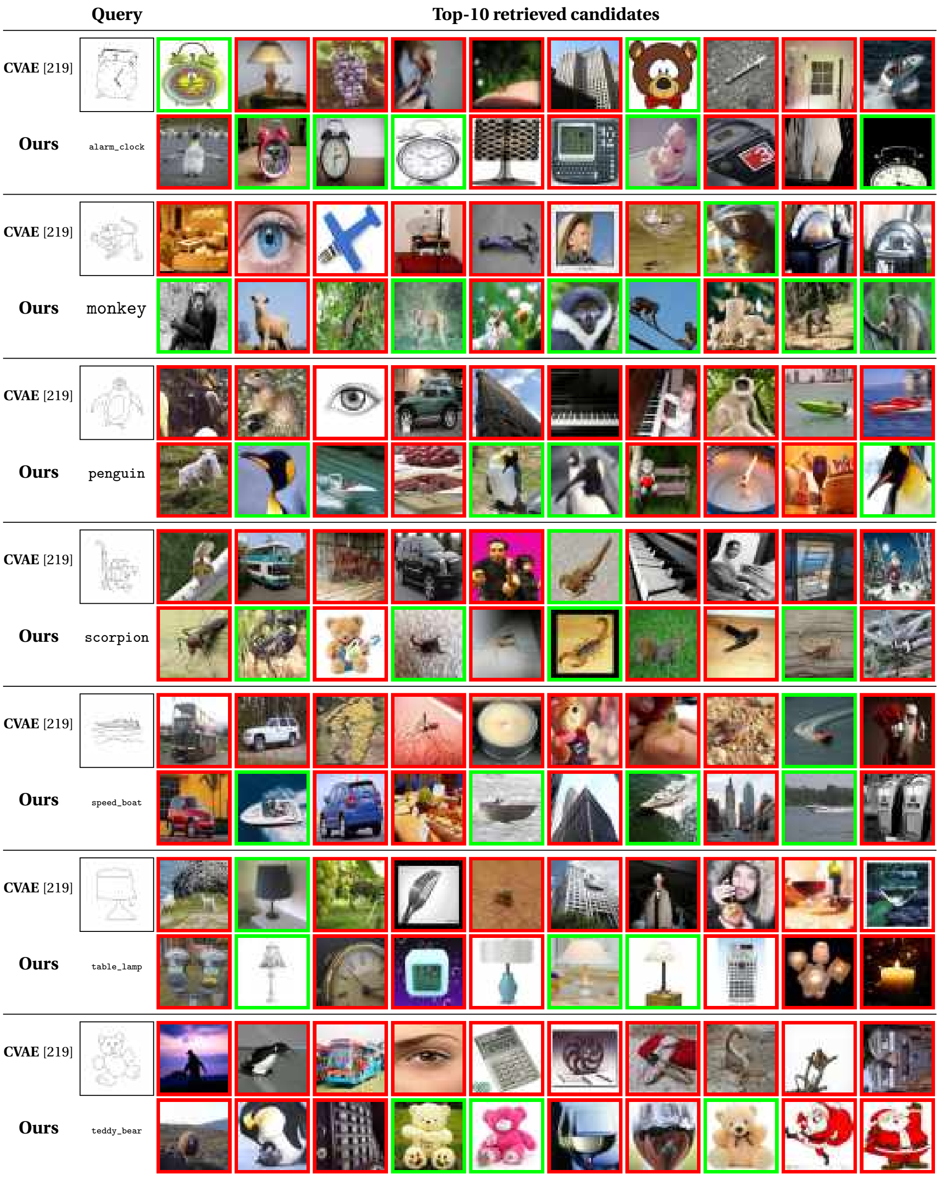
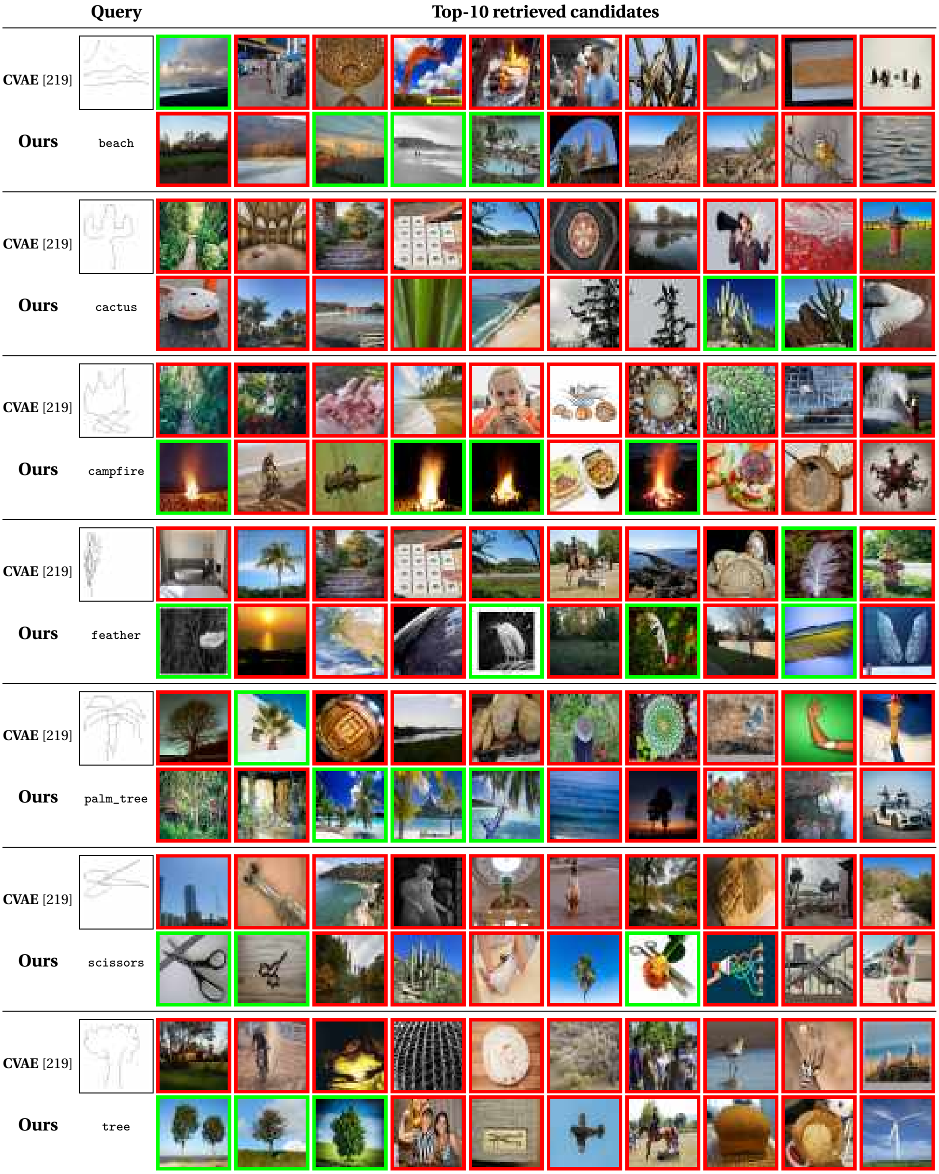


Figure 5.7: Top-10 image retrieval examples for *QuickDraw-Extended*. All the examples correspond to a zero-shot setting. First row provides a comparison with CVAE [219] method against our pipeline. Green and Red stands for correct and incorrect retrievals. (Better viewed in pdf)



tures are closely bounded for the same class. If observed carefully this helps to retrieve images which are visually similar to that of the sketches. In-case, of `bat` all the retrieved images have a completely spread wingspan similar to that of sketch. For `cows` and `rhinoceros` the front and the side face are really captured in the images. Even the bad retrievals in case of `mouse`, `rhinoceros` and `sword` have a lot of similar visual mappings.

Figure 5.6 presents the results for *TUBerlin-Extended*. Having the visual features nicely mapped in commmon embedding space we would like to demonstrate that including the semantic information ensures that the space also has some semantic correspondence. `speed_boat`, `scorpion` and `monkey` all of these sketches retrieves images that are either semantically or visual cues close to them.

Figure 5.7 presents the results for *QuickDraw-Extended*. Though this dataset set has the most abstractions and domain gap, our method is still able to fetch images corresponding to either the shape or the semantic feature of the queried sketch. `beach`, `cactus` and `tree` where the visual features precedes the semantic information, but in `palm_tree` the semantic information seems to play a huge part. For `feather` the retrieved images has a `bird` which are correctly in the database as it has feathers. If observed carefully the last retrieved image do contain `feather` but also contains a `fire_hydrant`. In annotation this was labelled as `fire_hydrant`. Also in case of `scissors` the fifth retrieved image has `scissors` in it, though the image is annotated as `bandage`.

5.8 Conclusions

This chapter represents a first step towards a practical ZS-SBIR task. Previous works on this task do not address some of the important challenges that appear when moving to an unconstrained retrieval and do not tackle with the large domain gap between amateur sketch and photo. In this scenario, to overcome the lack of proper data, we have contributed to the community a specifically designed large-scale ZS-SBIR dataset, *QuickDraw-Extended* which provides highly abstract amateur sketches collected with the Google Quick, Draw! game. Then, we have proposed a novel ZS-SBIR system that combines visual as well as semantic information to generate an image embedding. We experimentally show that this novel framework overcomes recent state-of-the-art methods in the ZS-SBIR setting.

Chapter 6

Conclusions and Future Work

There is no real ending. It's just the place where you stop the story.
– Dune, 1965 by Frank Herbert

In this chapter, we summarize the contributions of this thesis to the pattern recognition and computer vision fields and in particular, its application to sketch based image retrieval problems. We also highlight the main achievements and limitations of the proposed approaches. Finally, we lead the reader towards possible new research lines and natural extensions of the proposed methodologies.

In this thesis, we have introduced a study on how to retrieve images using sketch based queries by understanding and conglomerating different pattern recognition and machine learning strategies. In particular the huge vastness of digital imagery present now a days, it has become essential to have a highly efficient and intuitive retrieval system. All the variation in different natural language expression not always encompasses the true potential of the present day image retrieval systems. Image retrieval systems still based on textual queries do not take into account the very nascent as well as a primitive way of communication (which was hieroglyphs). Though the natural language processing research has improved a lot but still it lacks in translating many languages properly. This is where sketches play an important part where even if people are not able to understand each other verbally they can communicate drawing sketches and expressing themselves through their abstraction of the world. This abstraction of world objects may differ from people to people, this is where the real challenge lies to understand a person through its mind's eye for the world.

6.1 Summary of the Contributions

With the advent of touch screen devices it has become an utmost requirement to provide people with more flexibility to search in their vast image gallery the images of their choices. Besides voice and text, the sketch is one such modality of query which is quite native to the human mode of expression. After an introduction to the concept of sketches and the motivation in Chapter 1, we decided to make the readers familiar with the state of the art of different concepts and ideas in Chapter 2. Research-wise, we opted to tackle the problem of flexibility of the queries for image retrieval in many different experimental settings on one hand. On the other hand, we also took into account the new advancements in machine learning to incorporate those strategies to tackle the problem of lack of availability of human annotated data to make a better SBIR model.

The contributions presented in this work are enumerated in three points. Moreover, even though the focus of this thesis is the development of SBIR methodologies, some of the contributions are generic algorithms for computer vision multi-modal data. Let us briefly summarize these three contributions:

- **Cross-modal SBIR**: Single object sketch based queries were used for SBIR. In this chapter we have demonstrated that our framework can allow users to retrieve images queried by multiple objects of the same modality. In particular, we have shown competitive results when comparing the same model while attending different objects. No matter what the input query modality was our framework performed better than the state of the art and even at par at some specific experimental situations

- **Multi-modal and Multi-object SBIR**: While moving forward with SBIR from the previous chapter we thought of mixing the modalities of query for image retrieval(text and sketch). Previously we had been using the same modality in the particular query. We also designed a mechanism to put attention on the specific zones of the images depending on the inter-related objects which usually cooccur in nature. We also generalised Hungarian Loss using a different loss function that permits encoding the object based features. Consecutively we also validate the performance of our approach on standard single/multi-object datasets, showing state-of-the-art performance in every dataset. We also contributed with a dataset specially curated for multi-modal and multi-object SBIR.

- **Practical Zero-Shot SBIR**: Keeping the same flavour of the theme of the dissertation we realised that this supervised learning of world objects paired with sketches in the real scenario was pretty impossible. We represented a first step towards a practical ZS-SBIR task. We have contributed to the community a specifically designed large-scale ZS-SBIR dataset, and a novel ZS-SBIR system that combines visual as well as semantic information to generate an image embedding.

6.2 Discussions

This thesis has made several contributions in SBIR fields using the machine learning strategies. Although the application domain that has driven our research is SBIR, most of our contributions are transferal and can be formulated in terms of cross-modal retrieval, aligning salient objects, multi modal retrieval, and zero-shot learning. For example, Chapter 5, introduces a novel ZSL method which has proved to be state-of-the-art not only for SBIR but could be used as different piece of building blocks for other architecture.

Even though during this thesis deep learning have been used in a wide range of applications, there are particular cases where they are not the adequate representational framework. For example in Chapter 3 LSTM was used to remember the attention for different objects which can be easily overcome by transformers which much memory now a days. Remember we used in this chapter both vision and language interpretation. An important we had to face in this work is the time complexity involved in the loss calculation.

Chapter 4 proposed solutions to scale it to multi-modal query and use Hungarian loss to speed up the retrieval time. Nonetheless, researchers have proposed hashing techniques to allow these set of approaches to work at real time, a setting where this type of representations are very likely to be used. Both the chapters while using the vision and language representation to narrow the domain gaps, we always felt that an external knowledge in terms of structured information could help to bridge the gap between different objects in the nature.

We did use the external information as well tricks to reduce the domain gap in Chapter 5 while at the same time trying to address the zero shot problem, we realised more and more the necessity to move towards symbolic programming to give a more reasoning capability to our model. Having said that we did find some very interesting observation such as variation in level of sketching can affect the performance of the model a lot. We also realised mind abstraction for different object for different person is also very difficult for human to guess the intent of the sketch. While working on the real problem of lack of sketch data for all the objects and abstraction problem we made a benchmark for ZS-SBIR which could give a equated setting for all the future work on practical ZS-SBIR.

Hence in these scenario, we started some works in this regard towards neuro-symbolic AI which we would discuss in the following section.

6.3 Future Work

Along the thesis we have already stressed upon some open questions worth considering as unexplored lines. Moreover, taking into account the power of the deep learning methods, we are convince that there is still a wide variety of opportunities for improving and advancing our work. Also note that the new methodologies derived from the deep learning field has opened several research lines that were not covered in this dissertation. Deep learning is experiencing an evolution from the point of view of the learning strategies. The huge amount of data required for the supervision of new models causes a huge bottleneck dealing with new problems. Therefore, self-supervision and reinforcement learning strategies are gaining popularity among the machine learning community. We hypothesize that similar approaches will be able to deal properly with sketch data. As a matter of fact, we performed some tests on the neuro-symbolic ai setting in order to combine, both the power neural networks (NN) and symbolic AI. The NN finds the statistical correlations, can directly learn from the raw data and also capture complex patterns, reaching human level performance in many recognition and control tasks. But where it struggles to learn is the compositional and casual structure in how concepts are formed. This kind of notion is very useful for sketches which should be:

- Able to used as one shot classification,
- Should be parsed/ segmented into smaller parts,
- Which in turn can be useful for new exemplar generation,
- And also generation of new concepts.

In our case the lack of sketch data for all world objects is a huge drawback. The only way forward is to generate sketches in one shot manner. We did some preliminary experiments on human level concept learning for sketches using Bayesian probabilistic programming [98]. People learning new concepts can often generalize successfully from just a single example, yet machine learning algorithms typically require tens or hundreds of examples to perform with similar accuracy. This a computational model that captures these human learning abilities for a large class of simple visual concepts. The model represents concepts as simple programs that best explain observed examples under a Bayesian criterion in Equation 6.1. The Bayesian programming automates probabilistic reasoning. Bayesian Programming (BP) is a formalism and a methodology to specify probabilistic models and solve problems when all the necessary information are not available.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{6.1}$$

BP can restate many classical probabilistic models. The key idea behind this BP are compositionality, causality, and learning to learn. Decomposing the rich concepts into simpler primitives, use probabilistic semantics to handle noise and capture abstract casual structure, then constrict programs under a Bayesian criterion Equation 6.2.

$$P(\psi) = P(\kappa) \prod_{i=1}^{\kappa} P(S_i)\, P(R_i \mid S_1, \ldots, S_{i-1}) \tag{6.2}$$

In equation mentioned before ψ is a concept type, κ number of sub-parts, S_i is the sub-part and R_i is the relation between S_i and previous sub-parts. Thus $\psi = (\kappa, S, R)$. At first similar to the original paper [98] we generate concept tokens and types for train image. Extracting character skeleton and generate random parses and also search for sub-strokes(primitives) using the Equation 6.3.

$$P\left(x_i^{(m)}, y_i^{(m)}, z_i\right) = P(z_i) \prod_{j=1}^{n_i} P\left(y_{ij}^{(m)} \mid y_{ij}\right) P\left(y_{ij} \mid z_{ij}\right) \int P\left(x_{ij}^{(m)} \mid x_{ij}\right) P\left(x_{ij} \mid z_{ij}\right) \mathrm{d}x_{ij} \tag{6.3}$$

Then $\theta^{(a)}, \psi$ for each parse can be specified and choose some best parses by $P\left(\theta^{(m)} \mid \psi\right) P(\psi)$ where $\theta^{(m)} = \left\{L^{(m)}, x^{(m)}, y^{(a)}, R^{(a)}, A^{(n)}, \varepsilon^{(m)}, \sigma_b^{(m)}\right\}$ After this we compute posterior probability of test image belonging to each class and argmax : Assume K best match parses for test images:

$$\psi^{[1]}, \theta^{(m)[1]}, \ldots, \psi^{[K]}, \theta^{(m)[K]} \tag{6.4}$$

These matches approximate posterior with a discrete distribution:

$$P\left(\psi, \theta^{(m)} \mid I^{(m)}\right) \approx \sum_{i=1}^{K} w_i \delta\left(\theta^{(m)} - \theta^{(m)[i]}\right) \delta\left(\psi - \psi^{[i]}\right) \tag{6.5}$$

It does statistics on background data-sets to obtain various distributions, conditional distributions and Gaussian noise parameters.

Then, it runs probabilistic generative program according to statistics learned in the previous step. It learns abstract concept generation which is very useful in case of sketches.

$$\begin{gathered}
\text{where } \delta\left(x - x^{[i]}\right) = P\left(x \mid x^{[i]}\right) \quad \theta^{(m)} = \left\{L^{(m)}, x^{(m)}, y^{(m)}, R^{(m)}, A^{(m)}, \varepsilon^{(m)}, \sigma_b^{(m)}\right\} \\
\sum_i w_i = 1, w_i \text{ is proportional to parse score} \\
\text{In details, } w_i \propto P\left(\theta^{(m)|i|} \mid \psi^{[il}\right) P\left(\psi^{(i)}\right) \\
P\left(\theta^{(m)} \mid \psi\right) = P\left(L^{(m)} \mid \theta^{(m)}_{\setminus L^{(-)}}, \psi\right) \prod_i P\left(R_i^{(m)} \mid R_i\right) P\left(y_i^{(m)} \mid y_i\right) P\left(x_i^{(m)} \mid x_i\right) P\left(A^{(m)} \mathscr{E}^{(m)}, \sigma_b^{(m)}\right) \\
P(\psi) = P(\kappa) \prod_{i=1}^{K} P(S_i) P(R_i \mid S_1, \ldots, S_{i-1})
\end{gathered} \tag{6.6}$$

To get a better approximation take N sample from

$$P\left(\psi, \theta^{(m)} \mid I^{(m)}\right) \approx Q\left(\psi, \theta^{(m)}, I^{(m)}\right) = \sum_{i=1}^{K} w_i \delta\left(\theta^{(m)} - \theta^{(m)[i]}\right) \frac{1}{N} \sum_{i=1}^{N} \delta\left(\psi - \psi^{[ij]}\right) \tag{6.7}$$

$$\begin{aligned}\log P\left(I^{(T)} \mid I^{(c)}\right) &\approx \log \int P\left(I^{(T)} \mid \theta^{(T)}\right) P\left(\theta^{(T)} \mid \psi\right) Q\left(\theta^{(c)}, \psi, I^{(c)}\right) \mathrm{d}\psi \mathrm{d}\theta^{(c)} \mathrm{d}\theta^{(T)} \\ &\approx \log \sum_{i=1}^{K} w_i \max_{\theta^{(T)}} P\left(I^{(T)} \mid \theta^{(T)}\right) \frac{1}{N} \sum_{j=1}^{N} P\left(\theta^{(T)} \mid \psi^{[ij]}\right)\end{aligned} \tag{6.8}$$

Proceeding to one-shot generation, sampling from this posterior predictive distributions:

$$\begin{aligned}P\left(I^{(2)}, \theta^{(2)} \mid I^{(1)}\right) &= \int P\left(I^{(2)}, \theta^{(2)} \mid \theta^{(1)}, \psi\right) P\left(\theta^{(1)}, \psi \mid I^{(1)}\right) \mathrm{d}\left(\psi, \theta^{(1)}\right) \\ &= \int P\left(I^{(2)} \mid \theta^{(2)}\right) P\left(\theta^{(2)} \mid \psi\right) P\left(\theta^{(1)}, \psi \mid I^{(1)}\right) \mathrm{d}\left(\psi, \theta^{(1)}\right) \\ &\approx \int P\left(I^{(2)} \mid \theta^{(2)}\right) P\left(\theta^{(2)} \mid \psi\right) Q\left(\theta^{(1)}, \psi, I^{(1)}\right) \mathrm{d}\left(\psi, \theta^{(1)}\right) \\ &= \sum_{i=1}^{K} \sum_{j=1}^{N} \frac{w_i}{N} P\left(I^{(2)} \mid \theta^{(2)}\right) P\left(\theta^{(2)} \mid \psi^{[ij]}\right)\end{aligned} \tag{6.9}$$

Figure 6.1 shows artificially generated sketches when the model was asked to produce one shot new exemplars of the sketches it has never seen before.

Though the results were quite good. The symbolic AI(BPL) does have certain drawbacks which we realised too in the sketches produced. Namely,

- Probabilistic models make simplifying and rigid parametric assumptions, and when the assumptions are wrong, they create bias,
- The construction of the structured hypothesis spaces requires significant domain knowledge,
- Symbolic probabilistic models can not learn directly from raw data (Contrary to humans and NN).

Specifically for the above reasons we moved towards Neuro-symbolic AI which leverages both symbolic and NN paradigms. We did also apply Generative Neuro-Symbolic (GNS) model for sketch concept. The work is still in its preliminary stages.

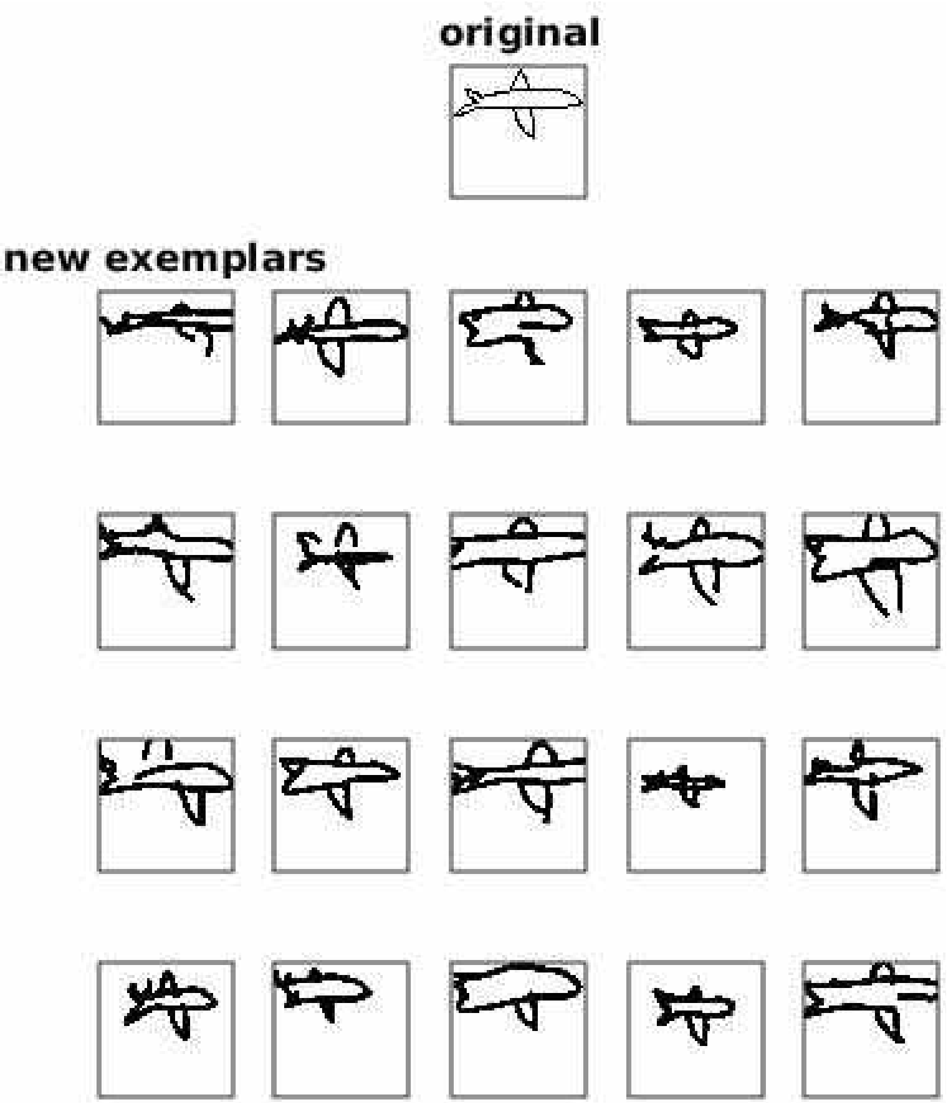


Figure 6.1: BPL method was given an image of a novel sketch(top) and asked to produce new exemplars. The twenty-sketch grids were generated by the method

Appendix

Along the chapters of this dissertation, we have conducted several experiments to evaluate the performance of proposed approaches. Therefore, several evaluation metrics and benchmarks have been used. Here, we present an appendix detailing carefully the dataset proposed by us to asses the performances

A Datasets

You should take the approach that you're wrong. Your goal is to be less wrong
– by Elon Musk

Along this thesis, several databases have been used to evaluate the performance of the proposed frameworks. Although this thesis focuses on SBIR problems, in order to test the generality of the proposed approaches, we have also experimented in other datasets made as contribution to the community. In particular, apart from the standard SBIR datasets, we make use of already well known computer vision datasets namely MS-COCO to make multi-object dataset to retrieve images corresponding to the multi-object query. Moreover, the dataset to benchmark ZS-SBIR is also discussed a bit.

A.1 MS-COCO modified

The MS-COCO modified dataset is curated from what was originally a large scale object detection, segmentation, and captioning dataset. It is a database of images containing multiple objects. As the label number for each image also varies considerably, rendering MS-COCO is even more challenging. We use the class names of the Sketchy dataset and take all possible combinations by taking two, three, four, five class names. We do this to keep the sketches available coherent with the multiple object dataset. Afterwards, we download the images belonging to these combined classes, and use them for training and retrieval. Few combined classes having less than 10 images are eliminated, leaving 125 number of combined classes for the experiment with at least 900 images. Some such images is also shown in the Figure 2, Figure 3 and Figure 4 for further reference. The dataset is readily available online.

A.2 QuickDraw-Extended

This dataset was proposed keeping in mind the limitations of the datasets present at the time of presenting the paper in a ZS-SBIR scenario. It adds to the community a novel large-scale dataset for the same. The challenges of practical ZS-SBIR it covers are as follows:

1. the large domain gap between amateur sketch and photo
2. move towards large-scale retrieval
3. large intra-class variability provided by the high abstraction level of different people drawing.

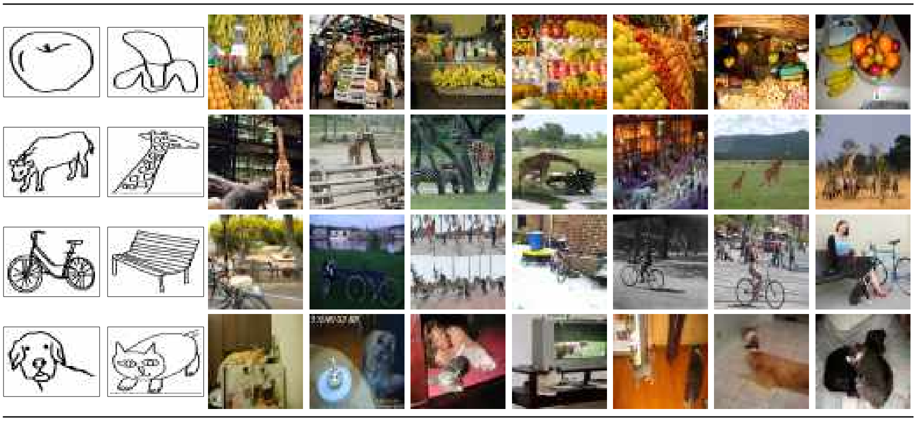


Figure 2: MS-COCO modified dataset [37]. Each row displays the multiple class label (2 in this case) in form of sketches(within the black frame) and then 'seven' such images associated with it, available in the dataset. (Better viewed in pdf)

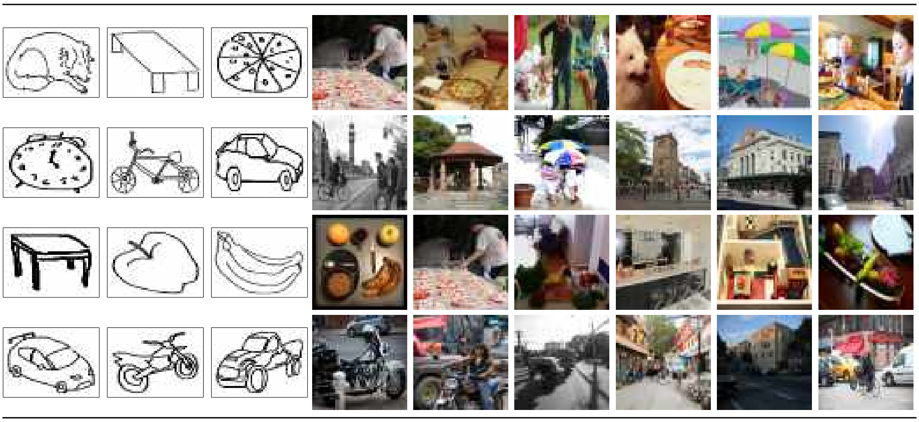


Figure 3: MS-COCO modified dataset [37]. Each row displays the multiple class label (3 in this case) in form of sketches(within the black frame) and then 'six' such images associated with it, available in the dataset. (Better viewed in pdf)

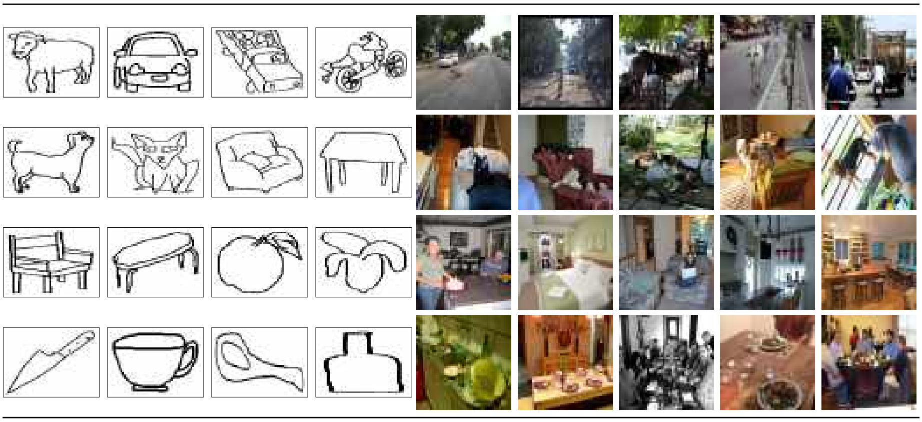


Figure 4: MS-COCO modified dataset [37]. Each row displays the multiple class label (4 in this case) in form of sketches(within the black frame) and then ’five’ such images associated with it, available in the dataset. (Better viewed in pdf)

We used the Google Quick, Draw! data data which is a huge collection of drawings (50 millions) belonging to 345 categories. We make use of a subset of sketches to construct a novel dataset for large-scale ZS-SBIR containing 110 categories (80 for training and 30 for testing). As a retrieval gallery, we provide images extracted from *Flickr* tagged with the corresponding label. Manual filtering is performed to remove outliers. Finally, this dataset consists of 330, 000 sketches and 204, 000 photos moving towards a large-scale retrieval. We consider that this dataset will provide better insights about the real performance of ZS-SBIR in a real scenario. Figure 5 shows a glimpse of the dataset collected.

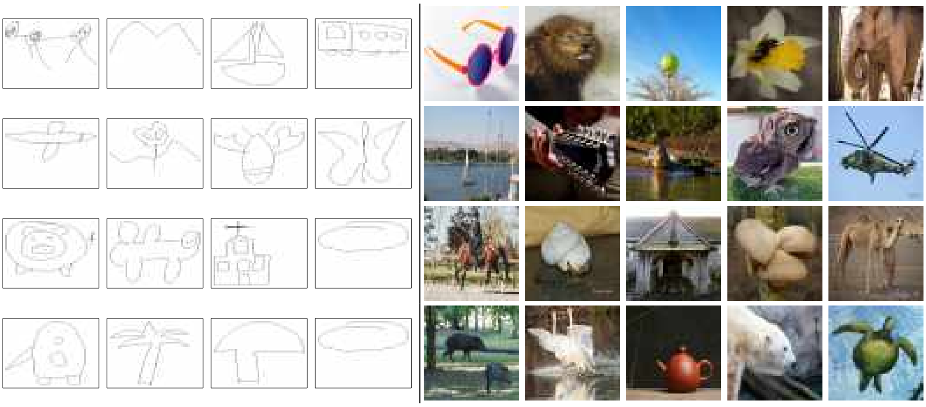


Figure 5: QuickDraw-Extended dataset [40]. The left hand side has the free hand sketches (within the black frame) mined from Quick Draw! and then images mined from Flickr as mentioned in the Chapter 5 is on the right hand side. (Better viewed in pdf)

List of Contribution

Coming together is a beginning; keeping together is progress;
working together is success.
by Henry Ford

In this chapter, we summarize the contributions of this dissertation to the pattern recognition and computer vision fields and in particular, its application to sketch based image retrieval problems. We also highlight the main achievements and limitations of the proposed approaches. Finally, we lead the reader towards possible new research lines and natural extensions of the proposed methodologies.

Topics

The main topic of this dissertation, is the development of better sketch based image retrieval systems. However, this thesis has also generated other side contributions in other topics that had raised our attention on the same field.

- **Sketch-based Image Retrieval** : This topic investigates the problem of sketch-based image retrieval, where human sketches are used as queries to conduct retrieval of photos. We focus on the problem from the supervised learning, query flexibility and lack of human annotated data availability
- **Document Image Analysis** : This task refers to algorithms and techniques that are applied to images of documents to obtain a computer-readable description from pixel data. In particular different handwritten documents both historical and recent, signature verification.
- **Vision and Language** : It is a task of visual recognition and language understanding which are two challenging tasks in artificial intelligence. Particularly, we a study of research at the intersection of vision and language.

International Journals

- Andres Mafla, Rubèn Tito, **Sounak Dey***, Lluis Gomez, Marçal Rossinyol and Dimosthenis Karatzas, "Real-Time Lexicon-Free Scene Text Retrieval", in *Pattern Recognition*, 2020.
- **Sounak Dey***, Anguelos Nicolaou, Josep Lladós and Umapada Pal,"Evaluation of word spotting under improper segmentation scenario", in *International Journal on Document Analysis and Research (IJDAR)*, 2019.
- **Sounak Dey***, Palaiahnakote Shivakumara, KS Raghunandan, Umapada Pal, Tong Lu, G Hemantha Kumar and Chee Seng Chan, "Script independent approach for multi-oriented text detection in scene image", in *Neurocomputing*, 2017.

International Conferences

- Andres Mafla, **Sounak Dey***, Ali Furkan Biten, Lluis Gomez and Dimosthenis Karatzas, "Multi-Modal Reasoning Graph for Scene-Text Based Fine-Grained Image Classification and Retrievals", submitted in *Winter Application in Computer Vision (WACV)*, 2021.
- Andres Mafla, **Sounak Dey***, Ali Furkan Biten, Lluis Gomez and Dimosthenis Karatzas, "Fine-grained Image Classification and Retrieval by Combining Visual and Locally Pooled Textual Features", in *Winter Application in Computer Vision (WACV)*, 2020.
- **Sounak Dey***, Pau Riba, Anjan Dutta, Josep Lladós and Yi-Zhe Song, "Doodle to search: Practical zero-shot sketch-based image retrieval", in *Computer Vision and Pattern Recognition (CVPR)* (***ORAL**), 2019.
- **Sounak Dey***, Anjan Dutta, Suman K Ghosh, Ernest Valveny, Josep Lladós and Umapada Pal, "Aligning Salient Objects to Queries: A Multi-modal and Multi-object Image Retrieval Framework", in *Asian Conference on Computer Vision (ACCV)*, 2018.
- **Sounak Dey***, Anjan Dutta, Suman K Ghosh, Ernest Valveny, Josep Lladós and Umapada Pal,"Learning Cross-Modal Deep Embeddings for Multi-Object Image Retrieval using Text and Sketch", in *International Conference on Pattern Recognition (ICPR)*, 2018.
- J Ignacio Toledo, **Sounak Dey***, Alicia Fornés and Josep Lladós, "Handwriting recognition by attribute embedding and recurrent neural networks", in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017.
- Pau Riba, Anjan Dutta, **Sounak Dey***, Josep Lladós and Alicia Fornés, "Improving Information Retrieval in Multiwriter Scenario by Exploiting the Similarity Graph

of Document Terms", in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017.

International Workshops

- Anguelos Nicolaou, **Sounak Dey***, Vincent Christlein, Andreas Maier and Dimosthenis Karatzas, "Non-deterministic Behavior of Ranking-based Metrics when Evaluating Embeddings", in *International Workshop on Reproducible Research in Pattern Recognition (ICPR)*, 2018.
- **Sounak Dey***, Anjan Dutta, Josep Lladós and Umapada Pal, "Bringing Back Hieroglyph", in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017.
- **Sounak Dey***, Anjan Dutta, Josep Lladós, Alicia Fornés and Umapada Pal, "Shallow neural network model for hand-drawn symbol recognition in multi-writer scenario", in *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 2017.
- **Sounak Dey***, Anguelos Nicolaou, Josep Llados and Umapada Pal, "Local binary pattern for word spotting in handwritten historical document", in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern, Recognition (SSPR)*, 2017.

arXiv

- **Sounak Dey***, Anjan Dutta, J Ignacio Toledo, Suman K Ghosh, Josep Lladós and Umapada Pal, "Signet: Convolutional siamese network for writer independent offline signature verification", in *arXiv*, 2017.

Bibliography

[1] Facebook has a quarter of a trillion user photos. http://mashable.com/2013/09/16/facebook-photo-uploads. Accessed: 30-04-2018.

[2] Instagram study 2014 q3. http://get.simplymeasured.com/rs/simplymeasured/images/InstagramStudy2014Q3.pdf. Accessed: 30-04-2018.

[3] The state of social marketing. https://get.simplymeasured.com/rs/135-YGJ-288/images/SM_StateOfSocial-2017.pdf. Accessed: 30-04-2018.

[4] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *IJCV*, 123(1), 2017.

[5] Julien Ah-Pine, Gabriela Csurka, and Stéphane Clinchant. Unsupervised visual and textual information fusion in cbmir using graph-based methods. *ACM TOIS*, 33(2):9, 2015.

[6] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *European conference on computer vision*, pages 584–599. Springer, 2014.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[8] Sreyasee Das Bhattacharjee, Junsong Yuan, Weixiang Hong, and Xiang Ruan. Query adaptive instance search using object sketches. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1306–1315. ACM, 2016.

[9] Sreyasee Das Bhattacharjee, Junsong Yuan, Yicheng Huang, Jingjing Meng, and Lingyu Duan. Query adaptive multiview object instance search and localization using sketches. *IEEE Transactions on Multimedia*, 20(10):2761–2773, 2018.

[10] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9779–9788, 2020.

[11] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 2017.

[12] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*. 2010.

[13] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *ECCV*, pages 730–746, 2016.

[14] Tu Bui, L Ribeiro, Moacir Ponti, and John Collomosse. Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network. *Computer Vision and Image Understanding*, 164:27–37, 2017.

[15] Tu Bui, Leonardo Ribeiro, Moacir Ponti, and John Collomosse. Generalisation and sharing in triplet convnets for sketch based visual search. *arXiv preprint arXiv:1611.05301*, 2016.

[16] Tu Bui, Leonardo Ribeiro, Moacir Ponti, and John Collomosse. Deep manifold alignment for mid-grain sketch based image retrieval. In *Asian Conference on Computer Vision*, pages 314–329. Springer, 2018.

[17] Sema Candemir, Eugene Borovikov, KC Santosh, Sameer Antani, and George Thoma. Rsilc: rotation-and scale-invariant, line-based color-aware descriptor. *Image and Vision Computing*, 42:1–12, 2015.

[18] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[19] Nan Cao, Xin Yan, Yang Shi, and Chaoran Chen. Ai-sketcher: A deep generative model for producing high-quality sketches. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2564–2571, 2019.

[20] Yang Cao, Hai Wang, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang. Mindfinder: interactive sketch-based image search on millions of images. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1605–1608. ACM, 2010.

[21] Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2940–2949, 2016.

[22] Abdolah Chalechale, Golshah Naghdy, and Alfred Mertins. Sketch-based image matching using angular partitioning. *IEEE Transactions on Systems, Man, and Cybernetics-part a: systems and humans*, 35(1):28–41, 2004.

[23] Soravit Changpinyo, Wei-Lun Chao, and Fei Sha. Predicting visual exemplars of unseen classes for zero-shot learning. In *ICCV*, 2017.

[24] Yin Chans, Zhibin Lei, Daniel P Lopresti, and Sun-Yuan Kung. Feature-based approach for image retrieval by sketch. In *Multimedia Storage and Archiving Systems II*, volume 3229, pages 220–231. International Society for Optics and Photonics, 1997.

[25] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu. Sketch2photo: Internet image montage. In *ACM transactions on graphics (TOG)*, number 5, page 124. ACM, 2009.

[26] Jungwoo Choi, Heeryon Cho, Jinjoo Song, and Sang Min Yoon. Sketchhelper: Real-time stroke guidance for freehand sketch retrieval. *IEEE Transactions on Multimedia*, 21(8):2083–2092, 2019.

[27] Sumit Chopra, Raia Hadsell, Yann LeCun, et al. Learning a similarity metric discriminatively, with application to face verification. In *CVPR (1)*, pages 539–546, 2005.

[28] John Collomosse, Tu Bui, and Hailin Jin. Livesketch: Query perturbations for guided sketch-based visual search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2887, 2019.

[29] John Collomosse, Tu Bui, Michael J Wilber, Chen Fang, and Hailin Jin. Sketching with style: Visual search with sketches and aesthetic context. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2660–2668, 2017.

[30] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[31] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *CVIU*, 163:90–100, 2017.

[32] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):5, 2008.

[33] Alberto Del Bimbo and Pietro Pala. Visual image retrieval by elastic matching of user sketches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):121–132, 1997.

[34] Alberto Del Bimbo, Pietro Pala, and Simone Santini. Visual image retrieval by elastic deformation of object sketches. In *Proceedings of 1994 IEEE Symposium on Visual Languages*, pages 216–223. IEEE, 1994.

[35] Alberto Del Bimbo, Pietro Pala, and Simone Santini. Image retrieval by elastic matching of shapes and image patterns. In *Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems*, pages 215–218. IEEE, 1996.

[36] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[37] Sounak Dey, Anjan Dutta, Suman K Ghosh, Ernest Valveny, Josep Lladós, and Umapada Pal. Aligning salient objects to queries: A multi-modal and multi-object image retrieval framework. In *Asian Conference on Computer Vision*, pages 241–255. Springer, 2018.

[38] Sounak Dey, Anjan Dutta, Suman K Ghosh, Ernest Valveny, Josep Lladós, and Umapada Pal. Learning cross-modal deep embeddings for multi-object image retrieval using text and sketch. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 916–921. IEEE, 2018.

[39] Sounak Dey, Anjan Dutta, Suman Kumar Ghosh, Ernest Valveny, Josep Lladós, and Umapada Pal. Learning cross-modal deep embeddings for multi-object image retrieval using text and sketch. In *ICPR*, pages 916–921, 2018.

[40] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2179–2188, 2019.

[41] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[42] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5089–5098, 2019.

[43] Titir Dutta and Soma Biswas. Style-guided zero-shot sketch-based image retrieval. In *BMVC*, page 209, 2019.

[44] Vage Egiazarian, Oleg Voynov, Alexey Artemov, Denis Volkhonskiy, Aleksandr Safin, Maria Taktasheva, Denis Zorin, and Evgeny Burnaev. Deep vectorization of technical drawings. *arXiv preprint arXiv:2003.05471*, 2020.

[45] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? In *SIGGRAPH*, 2012.

[46] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph.*, 31(4):44–1, 2012.

[47] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. A descriptor for large scale image retrieval based on sketched feature lines. In *SBM*, pages 29–36, 2009.

[48] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. An evaluation of descriptors for large-scale image retrieval from sketched feature lines. *Computers & Graphics*, 34(5):482–498, 2010.

[49] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE transactions on visualization and computer graphics*, 17(11):1624–1636, 2010.

[50] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE transactions on VCG*, pages 1624–1636, 2011.

[51] Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE TVCG*, pages 1624–1636, 2011.

[52] Mathias Eitz, Ronald Richter, Tamy Boubekeur, Kristian Hildebrand, and Marc Alexa. Sketch-based shape retrieval. *ACM Trans. Graph.*, 31(4):31–1, 2012.

[53] Kevin Ellis, Daniel Ritchie, Armando Solar-Lezama, and Josh Tenenbaum. Learning to infer graphics programs from hand-drawn images. In *Advances in neural information processing systems*, pages 6059–6068, 2018.

[54] Christos Faloutsos, Ron Barber, Myron Flickner, Jim Hafner, Wayne Niblack, Dragutin Petkovic, and William Equitz. Efficient and effective querying by image content. *Journal of intelligent information systems*, 3(3-4):231–262, 1994.

[55] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *CVPR*, pages 1473–1482, 2015.

[56] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, et al. Query by image and video content: The qbic system. *computer*, 28(9):23–32, 1995.

[57] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[58] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-view zero-shot learning. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2332–2345, 2015.

[59] Yanwei Fu, Tao Xiang, Yu-Gang Jiang, Xiangyang Xue, Leonid Sigal, and Shaogang Gong. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine*, 35(1):112–125, 2018.

[60] Yaroslav Ganin, Tejas Kulkarni, Igor Babuschkin, SM Eslami, and Oriol Vinyals. Synthesizing programs for images using reinforced adversarial learning. *arXiv preprint arXiv:1804.01118*, 2018.

[61] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, 2015.

[62] Lianli Gao, Jingkuan Song, Fuhao Zou, Dongxiang Zhang, and Jie Shao. Scalable multimedia retrieval by deep learning hashing with relative similarity learning. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 903–906. ACM, 2015.

[63] Yongsheng Gao and Maylor KH Leung. Face recognition using line edge map. *IEEE transactions on pattern analysis and machine intelligence*, 24(6):764–779, 2002.

[64] Danilo Gasques, Janet G Johnson, Tommy Sharkey, and Nadir Weibel. What you sketch is what you get: Quick and easy augmented reality prototyping with pintar. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

[65] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *European conference on computer vision*, pages 392–407. Springer, 2014.

[66] Abel Gonzalez-Garcia, Joost van de Weijer, and Yoshua Bengio. Image-to-image translation for cross-domain disentanglement. *NeurIPS*, 2018.

[67] Albert Gordo, Jon Almazán, Naila Murray, and Florent Perronin. Lewis: Latent embeddings for word images and their semantics. In *ICCV*, pages 1242–1250, 2015.

[68] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pages 241–257. Springer, 2016.

[69] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *ECCV*, pages 241–257, 2016.

[70] Albert Gordo and Diane Larlus. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6589–6598, 2017.

[71] Longteng Guo, Jing Liu, Yuhang Wang, Zhonghua Luo, Wei Wen, and Hanqing Lu. Sketch-based image retrieval using generative adversarial networks. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1267–1268. ACM, 2017.

[72] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017.

[73] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *NC*, 16(12):2639–2664, 2004.

[74] Christopher F Herot. Graphical input through machine recognition of sketches. In *Proceedings of the 3rd annual conference on Computer graphics and interactive techniques*, pages 97–102, 1976.

[75] Kyoji Hirata and Toshikazu Kato. Query by visual example. In *international conference on extending database technology*, pages 56–71. Springer, 1992.

[76] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[77] Conghui Hu, Da Li, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-classifier: sketch-based photo classifier generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9136–9144, 2018.

[78] Rui Hu, Mark Barnard, and John Collomosse. Gradient field descriptor for sketch based retrieval and localization. In *ICIP*, pages 1025–1028, 2010.

[79] Rui Hu and John Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding*, 117(7):790–806, 2013.

[80] Rui Hu, Tinghuai Wang, and John Collomosse. A bag-of-regions approach to sketch-based image retrieval. In *2011 18th IEEE International Conference on Image Processing*, pages 3661–3664. IEEE, 2011.

[81] Fei Huang, Cheng Jin, Yuejie Zhang, Kangnian Weng, Tao Zhang, and Weiguo Fan. Sketch-based image retrieval with deep visual semantic descriptor. *Pattern Recognition*, 76:537–548, 2018.

[82] Forrest Huang, John F Canny, and Jeffrey Nichols. Swire: Sketch-based user interface retrieval. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2019.

[83] Cisco Visual Networking Index. Forecast and methodology, 2016–2021. *White paper, Cisco public*, 6, 2017.

[84] Stuart James and John Collomosse. Interactive video asset retrieval using sketched queries. In *Proceedings of the 11th European Conference on Visual Media Production*, page 11. ACM, 2014.

[85] Natasha Jaques, Jennifer McCleary, Jesse Engel, David Ha, Fred Bertsch, Rosalind Picard, and Douglas Eck. Learning via social awareness: Improving a deep generative sketching model with facial feedback. *arXiv preprint arXiv:1802.04877*, 2018.

[86] Andrin Jenal, Nikolay Savinov, Torsten Sattler, and Gaurav Chaurasia. Rnn-based generative model for fine-grained sketching. *arXiv preprint arXiv:1901.03991*, 2019.

[87] Tianbi Jiang, Gui-Song Xia, and Qikai Lu. Sketch-based aerial image retrieval. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3690–3694. IEEE, 2017.

[88] Jonas Jongejan, Henry Rowley, Takashi Kawashima, Jongmin Kim, and Nick Fox-Gieg. The quick, draw! - a.i. experiment. *https://quickdraw.withgoogle.com*, 2016.

[89] Nour Karessli, Zeynep Akata, Bernt Schiele, Andreas Bulling, et al. Gaze embeddings for zero-shot image classification. In *CVPR*, 2017.

[90] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[91] Toshikazu Kato, Takio Kurita, Nobuyuki Otsu, and Kyoji Hirata. A sketch retrieval method for full color image database-query by visual example. In *[1992] Proceedings. 11th IAPR International Conference on Pattern Recognition*, pages 530–533. IEEE, 1992.

[92] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, abs/1412.6980, 2014.

[93] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *CVPR*, pages 4437–4446, 2015.

[94] Shin-ichiro Kondo, Masahiro Toyoura, and Xiaoyang Mao. Sketch based skirt image retrieval. In *Proceedings of the 4th Joint Symposium on Computational Aesthetics, Non-Photorealistic Animation and Rendering, and Sketch-Based Interfaces and Modeling*, pages 11–16, 2014.

[95] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, pages 32–73, 2017.

[96] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[97] Kin Chung Kwan and Hongbo Fu. Mobi3dsketch: 3d sketching in mobile ar. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[98] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

[99] Alex Lamb, Sherjil Ozair, Vikas Verma, and David Ha. Sketchtransfer: A new dataset for exploring detail-invariance and the abstractions learned by deep networks. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 963–972, 2020.

[100] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE PAMI*, 36(3):453–465, 2014.

[101] Tian Lan, Weilong Yang, Yang Wang, and Greg Mori. Image retrieval with structured object queries using latent ranking svm. In *ECCV*, pages 129–142, 2012.

[102] Jianjun Lei, Yuxin Song, Bo Peng, Zhanyu Ma, Ling Shao, and Yi-Zhe Song. Semi-heterogeneous three-way joint embedding network for sketch-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[103] Jianjun Lei, Kaifu Zheng, Hua Zhang, Xiaochun Cao, Nam Ling, and Yonghong Hou. Sketch based image retrieval via image-aided cross domain learning. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3685–3689. IEEE, 2017.

[104] Bo Li, Yijuan Lu, Afzal Godil, Tobias Schreck, Benjamin Bustos, Alfredo Ferreira, Takahiko Furuya, Manuel J Fonseca, Henry Johan, Takahiro Matsuda, et al. A comparison of methods for sketch-based 3d shape retrieval. *Computer Vision and Image Understanding*, 119:57–80, 2014.

[105] Bo Li, Yijuan Lu, Chunyuan Li, Afzal Godil, Tobias Schreck, Masaki Aono, Martin Burtscher, Hongbo Fu, Takahiko Furuya, Henry Johan, et al. Shrec'14 track: Extended large scale sketch-based 3d shape retrieval. In *Eurographics workshop on 3D object retrieval*, volume 2014, 2014.

[106] Jiangtong Li, Zhixin Ling, Li Niu, and Liqing Zhang. Bi-directional domain translation for zero-shot sketch-based image retrieval. *arXiv preprint arXiv:1911.13251*, 2019.

[107] Ke Li, Kaiyue Pang, Jifei Song, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Honggang Zhang. Universal sketch perceptual grouping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 582–597, 2018.

[108] Ke Li, Kaiyue Pang, Yi-Zhe Song, Timothy Hospedales, Honggang Zhang, and Yichuan Hu. Fine-grained sketch-based image retrieval: The role of part-aware attributes. In *WACV*, pages 1–9, 2016.

[109] Yanan Li, Donghui Wang, Huanhang Hu, Yuetan Lin, and Yueting Zhuang. Zero-shot recognition using dual visual-semantic mapping paths. In *CVPR*, 2017.

[110] Yi Li, Timothy M Hospedales, Yi-Zhe Song, and Shaogang Gong. Fine-grained sketch-based image retrieval by matching deformable part models. 2014.

[111] Yi Li, Yi-Zhe Song, Timothy M Hospedales, and Shaogang Gong. Free-hand sketch synthesis with deformable stroke models. *International Journal of Computer Vision*, 122(1):169–190, 2017.

[112] Hangyu Lin, Yanwei Fu, Peng Lu, Shaogang Gong, Xiangyang Xue, and Yu-Gang Jiang. Tc-net for isbir: Triplet classification network for instance-level sketch based image retrieval. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1676–1684, 2019.

[113] Kevin Lin, Huei-Fang Yang, Jen-Hao Hsiao, and Chu-Song Chen. Deep learning of binary hash codes for fast image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 27–35, 2015.

[114] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[115] Chenxi Liu, Junhua Mao, Fei Sha, and Alan L Yuille. Attention correctness in neural image captioning. In *AAAI*, pages 4176–4182, 2017.

[116] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, pages 2862–2871, 2017.

[117] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2862–2871, 2017.

[118] Yang Long, Li Liu, Yuming Shen, Ling Shao, and J Song. Towards affordable semantic searching: Zero-shot. retrieval via dominant attributes. In *AAAI*, 2018.

[119] David G Lowe et al. Object recognition from local scale-invariant features. In *iccv*, volume 99, pages 1150–1157, 1999.

[120] Peng Lu, Gao Huang, Yanwei Fu, Guodong Guo, and Hangyu Lin. Learning large euclidean margin for sketch-based image retrieval. *arXiv preprint arXiv:1812.04275*, 2018.

[121] Long Mai, Hailin Jin, Zhe Lin, Chen Fang, Jonathan Brandt, and Feng Liu. Spatial-semantic image search by visual feature synthesis. In *CVPR*, pages 1121–1130, 2017.

[122] Yusuke Matsui. Challenge for manga processing: Sketch-based manga retrieval. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 661–664, 2015.

[123] John FJ Mellor, Eunbyung Park, Yaroslav Ganin, Igor Babuschkin, Tejas Kulkarni, Dan Rosenbaum, Andy Ballard, Theophane Weber, Oriol Vinyals, and SM Eslami. Unsupervised doodling and painting with improved spiral. *arXiv preprint arXiv:1910.01007*, 2019.

[124] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[125] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[126] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, pages 3111–3119, 2013.

[127] J. Munkres. Algorithms for the assignment and transportation problems. *JSIAM*, pages 32–38, 1957.

[128] Vinod Nair and Geoffrey E Hinton. Inferring motor programs from images of handwritten digits. In *Advances in neural information processing systems*, pages 515–522, 2006.

[129] Carlton Wayne Niblack, Ron Barber, Will Equitz, Myron D Flickner, Eduardo H Glasman, Dragutin Petkovic, Peter Yanker, Christos Faloutsos, and Gabriel Taubin. Qbic project: querying images by content, using color, texture, and shape. In *Storage and retrieval for image and video databases*, volume 1908, pages 173–187. International Society for Optics and Photonics, 1993.

[130] Anguelos Nicolaou, Sounak Dey, Vincent Christlein, Andreas Maier, and Dimosthenis Karatzas. Non-deterministic behavior of ranking-based metrics when evaluating embeddings. *arXiv preprint arXiv:1806.07171*, 2018.

[131] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014.

[132] Anubha Pandey, Ashish Mishra, Vinay Kumar Verma, and Anurag Mittal. Adversarial joint-distribution learning for novel class sketch-based image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[133] Anubha Pandey, Ashish Mishra, Vinay Kumar Verma, Anurag Mittal, and Hema Murthy. Stacked adversarial network for zero-shot sketch based image retrieval. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2540–2549, 2020.

[134] Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Generalising fine-grained sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 677–686, 2019.

[135] Kaiyue Pang, Yi-Zhe Song, Tony Xiang, and Timothy M Hospedales. Cross-domain generative learning for fine-grained sketch-based image retrieval. In *BMVC*, 2017.

[136] Kaiyue Pang, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10347–10355, 2020.

[137] Sarthak Parui and Anurag Mittal. Similarity-invariant sketch-based image retrieval in large databases. In *European Conference on Computer Vision*, pages 398–414. Springer, 2014.

[138] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. Pytorch, 2017.

[139] Mattis Paulin, Julien Mairal, Matthijs Douze, Zaid Harchaoui, Florent Perronnin, and Cordelia Schmid. Convolutional patch representations for image retrieval: an unsupervised approach. *IJCV*, 121:149–168, 2017.

[140] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019.

[141] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[142] Tiziano Portenier, Qiyang Hu, Attila Szabo, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. Faceshop: Deep sketch-based face image editing. *arXiv preprint arXiv:1804.08972*, 2018.

[143] Yonggang Qi, Yi-Zhe Song, Honggang Zhang, and Jun Liu. Sketch-based image retrieval via siamese convolutional neural network. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2460–2464. IEEE, 2016.

[144] Yonggang Qi and Zheng-Hua Tan. Sketchsegnet+: An end-to-end learning of rnn for multi-class sketch semantic segmentation. *Ieee Access*, 7:102717–102726, 2019.

[145] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European conference on computer vision*, pages 3–20. Springer, 2016.

[146] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. Deep shape matching. In *Proceedings of the european conference on computer vision (eccv)*, pages 751–767, 2018.

[147] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[148] R Kumar Rajendran and Shih-Fu Chang. Image retrieval with sketches and compositions. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, volume 2, pages 717–720. IEEE, 2000.

[149] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPRW*, pages 512–519, 2014.

[150] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, pages 49–58, 2016.

[151] Umar Riaz Muhammad, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning deep sketch abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8014–8023, 2018.

[152] Leo Sampaio Ferraz Ribeiro, Tu Bui, John Collomosse, and Moacir Ponti. Sketchformer: Transformer-based representation for sketched structure. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14153–14162, 2020.

[153] Yong Rui, Thomas S Huang, and Shih-Fu Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of visual communication and image representation*, 10(1):39–62, 1999.

[154] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, pages 211–252, 2015.

[155] Jose M Saavedra. 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)*, pages 2998–3002. IEEE, 2014.

[156] Jose M Saavedra, Juan Manuel Barrios, and S Orand. Sketch based image retrieval using learned keyshapes (lks). In *BMVC*, volume 1, page 7, 2015.

[157] Jose M Saavedra, Juan Manuel Barrios, and S Orand. Sketch based image retrieval using learned keyshapes (lks). In *BMVC*, volume 1, pages 1–10, 2015.

[158] Alexandre Sablayrolles, Matthijs Douze, Nicolas Usunier, and Hervé Jégou. How should we evaluate supervised hashing? In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1732–1736. IEEE, 2017.

[159] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *SIGGRAPH*, 2016.

[160] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):119, 2016.

[161] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2017.

[162] KC Santosh, Bart Lamiroy, and Laurent Wendling. Symbol recognition using spatial relations. *Pattern Recognition Letters*, 33(3):331–341, 2012.

[163] KC Santosh, Bart Lamiroy, and Laurent Wendling. Dtw–radon-based shape descriptor for pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 27(03):1350008, 2013.

[164] Ravi Kiran Sarvadevabhatla, Isht Dwivedi, Abhijat Biswas, Sahil Manocha, et al. Sketchparse: Towards rich descriptions for poorly drawn sketches using multitask hierarchical deep networks. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 10–18. ACM, 2017.

[165] Ravi Kiran Sarvadevabhatla, Shiv Surya, Trisha Mittal, and R Venkatesh Babu. Pictionary-style word guessing on hand-drawn object sketches: Dataset, analysis and deep network models. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):221–231, 2018.

[166] Kazuma Sasaki and Tetsuya Ogata. Adaptive drawing behavior by visuomotor learning using recurrent neural networks. *IEEE Transactions on Cognitive and Developmental Systems*, 11(1):119–128, 2018.

[167] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[168] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[169] Omar Seddati, Stéphane Dupont, and Saïd Mahmoudi. Quadruplet networks for sketch-based image retrieval. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 184–191. ACM, 2017.

[170] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, 2018.

[171] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[172] Saurabh Singh, Derek Hoiem, and David Forsyth. Learning to localize little landmarks. In *CVPR*, pages 260–269, 2016.

[173] Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning to sketch with shortcut cycle consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 801–810, 2018.

[174] Jifei Song, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Xiang Ruan. Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval. In *BMVC*, volume 1, page 3, 2016.

[175] Jifei Song, Yi-Zhe Song, Tony Xiang, and Timothy M Hospedales. Fine-grained image retrieval: the text/sketch input dilemma. In *BMVC*, volume 1, page 2, 2017.

[176] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, 2017.

[177] R. Stewart, M. Andriluka, and A. Y. Ng. End-to-end people detection in crowded scenes. In *CVPR*, pages 2325–2333, 2016.

[178] Sarah Suleri, Vinoth Pandian Sermuga Pandian, Svetlana Shishkovets, and Matthias Jarke. Eve: A sketch-based software prototyping workbench. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

[179] Ivan E Sutherland. Sketchpad a man-machine graphical communication system. *Simulation*, 2(5):R–3, 1964.

[180] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[181] Yap-Peng Tan, Sanjeev R Kulkarni, and Peter J Ramadge. A framework for measuring video similarity and its application to video query by example. In *Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348)*, volume 2, pages 106–110. IEEE, 1999.

[182] Xiaoou Tang and Xiaogang Wang. Face sketch recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):50–57, 2004.

[183] William Thong, Pascal Mettes, and Cees GM Snoek. Open cross-domain visual search. *arXiv preprint arXiv:1911.08621*, 2019.

[184] Giorgos Tolias and Ondrej Chum. Asymmetric feature maps with application to sketch based retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2377–2385, 2017.

[185] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *JMLR*, pages 2579–2605, 2008.

[186] V Varshaneya, Vineeth N Balasubramanian, and S Balasubramanian. Teaching gans to sketch in vector format. *arXiv*, 2019.

[187] Vinay Kumar Verma, Aakansha Mishra, Ashish Mishra, and Piyush Rai. Generative model for zero-shot sketch-based image retrieval. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 704–713. IEEE, 2019.

[188] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.

[189] Ji Wan, Dayong Wang, Steven Chu Hong Hoi, Pengcheng Wu, Jianke Zhu, Yongdong Zhang, and Jintao Li. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 157–166. ACM, 2014.

[190] Changhu Wang, Zhiwei Li, and Lei Zhang. Mindfinder: image search by interactive sketching and tagging. In *Proceedings of the 19th international conference on World wide web*, pages 1309–1312. ACM, 2010.

[191] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1875–1883, 2015.

[192] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *CVPR*, pages 1875–1883, 2015.

[193] Fei Wang, Shujin Lin, Xiaonan Luo, Hefeng Wu, Ruomei Wang, and Fan Zhou. A data-driven approach for sketch-based 3d shape retrieval via similar drawing-style recommendation. In *Computer Graphics Forum*, number 7, pages 157–166. Wiley Online Library, 2017.

[194] Fei Wang, Shujin Lin, Hefeng Wu, Hanhui Li, Ruomei Wang, Xiaonan Luo, and Xiangjian He. Spfusionnet: Sketch segmentation using multi-modal data fusion. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1654–1659. IEEE, 2019.

[195] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *ICCV*, 2017.

[196] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014.

[197] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *CVPR*, pages 2285–2294, 2016.

[198] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.

[199] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, pages 5005–5013, 2016.

[200] Shu Wang, Jian Zhang, Tony X Han, and Zhenjiang Miao. Sketch-based image retrieval through hypothesis-driven object boundary selection with hlr descriptor. *IEEE Transactions on Multimedia*, 17(7):1045–1057, 2015.

[201] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

[202] Xinggang Wang, Xiong Duan, and Xiang Bai. Deep sketch feature for cross-domain image retrieval. *Neurocomputing*, 207:387–397, 2016.

[203] Yanfei Wang, Fei Huang, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval. *Pattern Recognition*, 100:107148, 2020.

[204] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *ICCV*, 2017.

[205] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Hcp: A flexible cnn framework for multi-label image classification. *PAMI*, pages 1901–1907, 2016.

[206] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, volume 11, pages 2764–2770, 2011.

[207] Sabine Wieluch and Friedhelm Schwenker. Strokecoder: Path-based image generation from single examples using transformers. *arXiv preprint arXiv:2003.11958*, 2020.

[208] Changcheng Xiao, Changhu Wang, Liqing Zhang, and Lei Zhang. Sketch-based image retrieval via shape words. In *ACM ICMR*, pages 571–574, 2015.

[209] Jin Xie, Guoxian Dai, Fan Zhu, and Yi Fang. Learning barycentric representations of 3d shapes for sketch-based 3d shape retrieval. In *CVPR*, pages 3615–3623, 2017.

[210] Hao Xu, Jingdong Wang, Xian-Sheng Hua, and Shipeng Li. Interactive image search by 2d semantic map. In *ACM ICWWW*, pages 1321–1324, 2010.

[211] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[212] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015.

[213] Peng Xu, Yongye Huang, Tongtong Yuan, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, Zhanyu Ma, and Jun Guo. Sketchmate: Deep hashing for million-scale human sketch retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8090–8098, 2018.

[214] Peng Xu, Chaitanya K Joshi, and Xavier Bresson. Multi-graph transformer for free-hand sketch recognition. *arXiv preprint arXiv:1912.11258*, 2019.

[215] Xing Xu, Fumin Shen, Yang Yang, Dongxiang Zhang, Heng Tao Shen, and Jingkuan Song. Matrix tri-factorization with manifold regularizations for zero-shot learning. In *CVPR*, 2017.

[216] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. Exploit bounding box annotations for multi-label object recognition. In *CVPR*, pages 280–288, 2016.

[217] Lumin Yang, Jiajie Zhuang, Hongbo Fu, Kun Zhou, and Youyi Zheng. Sketchgcn: Semantic sketch segmentation with graph convolutional networks. *arXiv preprint arXiv:2003.00678*, 2020.

[218] Meng Ye and Yuhong Guo. Zero-shot classification with discriminative semantic representation learning. In *CVPR*, 2017.

[219] Sasikiran Yelamarthi, Shiva Krishna Reddy M, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018.

[220] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, pages 4651–4659, 2016.

[221] Deng Yu, Lei Li, Youyi Zheng, Manfred Lau, Yi-Zhe Song, Chew-Lan Tai, and Hongbo Fu. Sketchdesc: Learning local sketch descriptors for multi-view correspondence. *arXiv preprint arXiv:2001.05744*, 2020.

[222] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[223] Qian Yu, Xiaobin Chang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching. *arXiv preprint arXiv:1711.08106*, 2017.

[224] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 799–807, 2016.

[225] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. *International journal of computer vision*, 122(3):411–425, 2017.

[226] Qian Yu, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Sketch-a-net that beats humans. *arXiv preprint arXiv:1501.07873*, 2015.

[227] H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, and X. Cao. Sketchnet: Sketch classification with web images. In *CVPR*, 2016.

[228] Jingyi Zhang, Fumin Shen, Li Liu, Fan Zhu, Mengyang Yu, Ling Shao, Heng Tao Shen, and Luc Van Gool. Generative domain-migration hashing for sketch-to-image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 297–314, 2018.

[229] Jingyi Zhang, Fumin Shen, Li Liu, Fan Zhu, Mengyang Yu, Ling Shao, Heng Tao Shen, and Luc Van Gool. Generative domain-migration hashing for sketch-to-image retrieval. In *ECCV*, 2018.

[230] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015.

[231] Liang Zheng, Yi Yang, and Qi Tian. Sift meets cnn: A decade survey of instance retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1224–1244, 2017.

[232] Ningyuan Zheng, Yifan Jiang, and Dingjiang Huang. Strokenet: A neural painting environment. In *International Conference on Learning Representations*, 2018.

[233] Tao Zhou, Chen Fang, Zhaowen Wang, Jimei Yang, Byungmoon Kim, Zhili Chen, Jonathan Brandt, and Demetri Terzopoulos. Learning to doodle with stroke demonstrations and deep q-networks. In *BMVC*, page 13, 2018.

[234] Fan Zhu, Jin Xie, and Yi Fang. Learning cross-domain neural networks for sketch-based 3d shape retrieval. In *AAAI*, pages 3683–3689, 2016.

[235] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, pages 4995–5004, 2016.

[236] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014.

This work has been partially supported by the Spanish project TIN2015-70924-C2-2-R, the European project ERC- 2010-AdG-20100407-269796, the grants 2013-DI-067 and 2016-DI-095 from the Secretaria dUniversitats i Recerca del Departament dEconomia i Coneixement de la Generalitat de Catalunya, the Ramon y Cajal Fellowship RYC-2014-16831, the project RecerCaixa (XARXES, 2016ACUP-00008), a research program from Obra Social "La Caixa" with the collaboration of the ACUP, and the CERCA Programme/Generalitat de Catalunya. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.