This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3148853, IEEE Transactions on Pattern Analysis and Machine Intelligence

Deep Learning for Free-Hand Sketch: A Survey

Peng Xu, Timothy M. Hospedales, *Member, IEEE,* Qiyue Yin, Yi-Zhe Song, *Senior Member, IEEE,* Tao Xiang, and Liang Wang, *Fellow, IEEE*

Abstract—Free-hand sketches are highly illustrative, and have been widely used by humans to depict objects or stories from ancient times to the present. The recent prevalence of touchscreen devices has made sketch creation a much easier task than ever and consequently made sketch-oriented applications increasingly popular. The progress of deep learning has immensely benefited free-hand sketch research and applications. This paper presents a comprehensive survey of the deep learning techniques oriented at free-hand sketch data, and the applications that they enable. The main contents of this survey include: (i) A discussion of the intrinsic traits and unique challenges of free-hand sketch, to highlight the essential differences between sketch data and other data modalities, *e.g.*, natural photos. (ii) A review of the developments of free-hand sketch research in the deep learning era, by surveying existing datasets, research topics, and the state-of-the-art methods through a detailed taxonomy and experimental evaluation. (iii) Promotion of future work via a discussion of bottlenecks, open problems, and potential research directions for the community.

Index Terms—Free-Hand Sketch, Deep Learning, Survey, Introductory, Taxonomy.

1 Introduction

Free-hand sketch is a universal communication and art modality that transcends barriers to link human societies. It has been used from ancient times to today, comes naturally to children before writing, and transcends language barriers. 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 free-hand 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 (Figure 1).

Sketches can convey many words, or even concepts that are hard to convey at all in words. Figure 1 shows several examples covering ancient and contemporary; literal and emotional; iconic and descriptive; abstract and concrete; and different media of drawing.

Free-hand sketch can be illustrative, despite its highly concise and abstract nature, making it useful in various scenarios such as communication and design. Therefore, free-hand sketch has been widely studied in computer vision and pattern recognition [1]–[5], computer graphics [6], [7], human computer interaction [8]–[11], robotics [12], and cognitive science [13] communities. In particular, early research can be traced back to the 1960s and 1970s [14], [15].

However, free-hand sketch is fundamentally different from natural photos[1]. Sketch images provide a special data modality/domain that has both domain-unique challenges (*e.g.*, highly sparse, abstract, artist-dependent) and advantages (*e.g.*, lack of background, use of iconic representation). It is also unique in that free-hand sketch can be stored and processed in multiple representations as its source is a dynamic 'pen' movement. These include static pixel space (when rendered as an image), dynamic stroke coordinate space (when considered as a time series), and geometric graph space (when considered topologically) – as discussed in Section 2. Thus, from a pattern recognition or machine intelligence perspective, these unique traits of free-hand 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. This survey will review these considerations and designs in detail.

Sketch research and applications in both industry and academia have boomed in recent years due to the prevalence of touchscreen devices (*e.g.*, smartphone, tablet) that make acquiring sketch data much easier than ever; as well as the rapid development of deep learning techniques that are achieving state-of-the-art performance in diverse artificial intelligence tasks. This boom has occurred on several fronts: (i) Some classic research topics (*e.g.*, sketch recognition, sketch-based image retrieval, sketch-based 3D shape retrieval) have been re-studied in a deep learning context [2]–[4], [7], [18], [19] resulting in significant performance improvements. (ii) Some brand-new topics have been proposed based on deep learning, *e.g.*, deep learning based sketch generation/synthesis [5], sketch-based model generation [20], reinforcement learning based sketch abstraction [21], adversarial sketch based image editing [22], graph neural network based sketch recognition [23], graph convolution-based sketch semantic segmentation [24], and sketch based software prototyping [9]. (iii) Beyond global representation based tasks (*e.g.*, sketch recognition), more instance-level and stroke-level tasks have been further studied or proposed, *e.g.*, instance-level sketch-based image retrieval [4], and deep stroke-level sketch segmentation [25]. (iv) Compared with the conventional approach of representing sketches as static images [1], 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

1. Images can include both free-hand sketch and natural photos, *etc*. In this survey, "photo" denotes natural photo images obtained by a camera such as in ImageNet unless otherwise specified.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3148853, IEEE Transactions on Pattern Analysis and Machine Intelligence

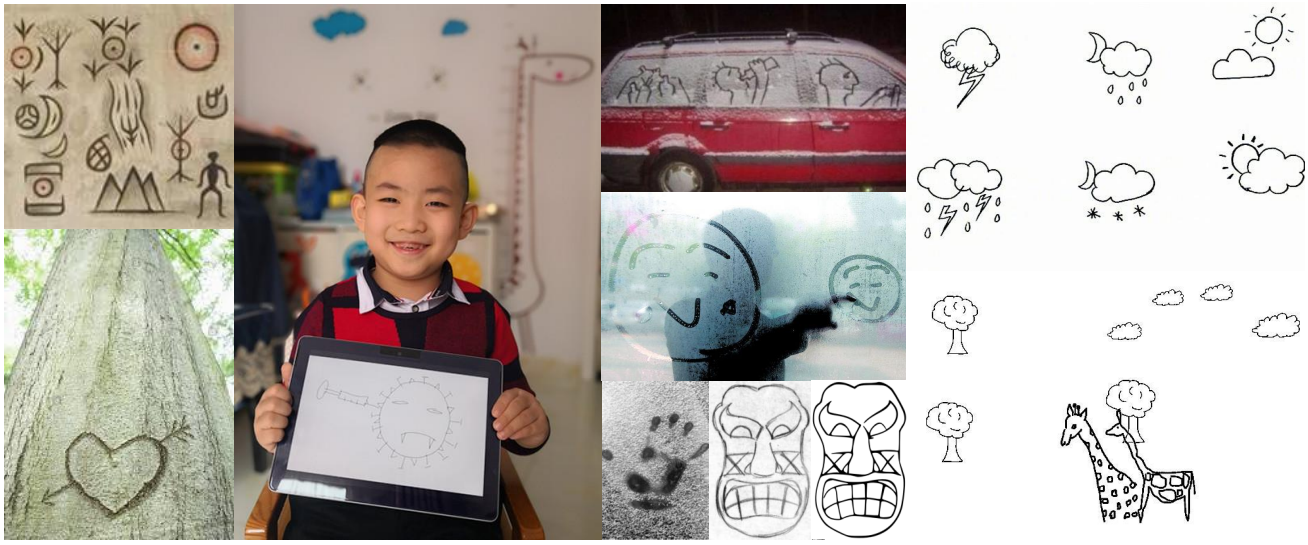


Fig. 1. Diverse free-hand sketches in human daily life. The masks (rough and simplified) on the bottom are from [16]. The scene-level sketch (cloud, trees, and giraffes) on the bottom right corner is from SketchyCOCO dataset [17].

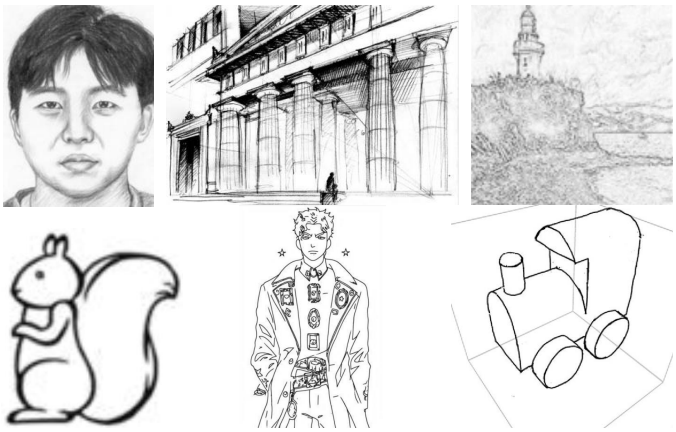


Fig. 2. Drawing samples out-of-scope of our focus on free-hand sketch.

SketchRNN [5], the sequential nature of free-hand sketches is now widely modeled by recurrent neural network (RNN). (v) More sketch-based applications have appeared, *e.g.*, the online sketch game QuickDraw[2] [5], and sketch-based commodity search engine[3] [4], [26]. (vi) Some large-scale sketch datasets have been collected, *e.g.*, Sketchy [7] and Google QuickDraw[4] [5] – a million-scale sketch dataset (50M+).

1.1 Overview

This survey aims to review the state of the free-hand sketch community in deep learning era, hoping to bring insights to researchers and assist practitioners aiming to build sketch-based applications.

Previous Surveys and Scope This survey focuses on free-hand sketches, not including professional (forensic) facial sketch [27]–[31], professional pencil sketch (professional line drawing/art) [32]–[38], professional landscape sketch [39], photo-like edge-maps (artificially rendered 'sketch') [40]–[42], cartoon/manga [43]–[45], well-drawn 3D sketch [46]. (See Figure 2.) In this survey, "sketch" refers to "free-hand sketch", unless otherwise specified.

To our knowledge, only a few survey papers [47]–[51] were published in the free-hand sketch community in the past decade. However, these survey papers [47]–[51]: (i) only focus on two research topics, *i.e.*, free-hand sketch based recognition and image/3D retrieval. (ii) mainly review classic non-deep techniques. In contrast, the current boom in advanced deep methodologies, techniques (hashing), representations (sequential, topological), and novel applications (generation, segmentation) makes it timely to provide an up-to-date survey of the big picture of research on free-hand sketch.

Contributions We provide a comprehensive survey reviewing the state of the field with regards to deep learning techniques, as well as applications of free-hand sketch. In particular: (a) We discuss the intrinsic traits, and unique challenges and opportunities posed when working with free-hand sketch data. (b) We provide a detailed taxonomy of both datasets, and applications covering both uni-modal (sketch alone) and multi-modal (relating sketches to photos, text, *etc*) cases. For each specific task, the contemporary landscape of deep learning solutions is summarized, and milestone works are described in detail. (c) We discuss current bottlenecks, open problems, and potential research directions for free-hand sketch.

Organization of This Survey The rest of this survey is organized as follows. Section 2 provides background on free-hand sketch, including intrinsic traits, domain-unique challenges, milestone techniques of the existing sketch-oriented deep learning works, *etc*. Section 3 summarizes representative free-hand sketch datasets. In Section 4, we provide a comprehensive taxonomy for various sketch-based tasks, and describe representative deep learning techniques in detail. Section 4 also presents some experiment comparisons

2. https://quickdraw.withgoogle.com
3. http://sketchx.eecs.qmul.ac.uk/demos/
4. https://github.com/googlecreativelab/quickdraw-dataset

TABLE 1
Notation and abbreviations used in this survey.

Notations	Descriptions
$\mathcal{X} = \{\mathbf{X}_n\}_{n=1}^N$	sketch sample set
$\mathbf{X}_m$, $\mathbf{X}_n$	m-th and n-th sketch samples in the sketch sample set $\mathcal{X}$
$\mathcal{Y} = \{y_n\}_{n=1}^N$	associated label set of $\mathcal{X}$
y_n	label of $\mathbf{X}_n$
$\mathcal{L}$	loss function
$\mathbf{\Theta}$	learnable parameters of neural network
$\mathcal{F}(\cdot)$	function mapping or feature extraction
$\mathcal{F}_\Theta(\cdot)$	neural network feature extraction, parameterized by Θ
$\mathcal{D}(\cdot,\cdot)$	distance metric, *e.g.*, ℓ_2 distance
λ	weighting factor
$\sum$	summation
α, β, γ	hyper parameters set manually

Abbreviated Terms	Descriptions
CNN	convolutional neural network
GNN	graph neural network
GCN	graph convolutional network
RNN	recurrent neural network
LSTM	Long Short Term Memory
GRU	Gated Recurrent Unit
BERT	Bidirectional Encoder Representations from Transformers
TCN	temporal convolutional neural network
GAN	generative adversarial network
VAE	Variational Auto Encoder
RL	Reinforcement Learning

based on `TorchSketch`[5] implementation. Section 5 discusses open problems, bottlenecks, and potential research directions before the survey concludes in Section 6.

Throughout this survey, bold uppercase and bold lowercase characters denote matrices and vectors, respectively. Unless specified otherwise, mathematical symbols and abbreviated terms follow the conventions in Table 1.

2 Background

This section presents background knowledge, including: the intrinsic traits, and domain-unique challenges and opportunities of free-hand sketch. In particular, we cover the essential differences to natural photos; and a brief development history of deep learning for free-hand sketch, summarizing the milestone techniques.

2.1 Intrinsic Traits and Domain-Unique Challenges

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 (Figure 3, left images). Since sketch generation is a dynamic process, suitably captured sketches can also be represented as a sequence of strokes or pen coordinates (Figure 3, right). In this regard, sketches share similarities with hand-written characters, yet are fundamentally different given their highly abstract and free-style nature (c.f., alphabetic hand-writing is subject to specific rules and a teaching process). From another perspective, free-hand sketches can also be modeled as a sparsely connected graph where lines

5. An open source sketch-oriented deep learning software library. Please see its GitHub page for details https://github.com/PengBoXiangShang/torchsketch.

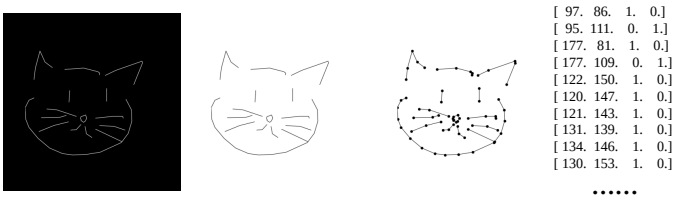


Fig. 3. Sketch-specific representations. Representations from left to right: sparse matrix (black background with white lines), dense picture (white background with black lines), graph, stroke sequence. Both graph and stroke sequence representations are based on the key stroke points. In stroke sequence, each key point is denoted as a four-tuple, where the first two entries and the last two entries represent the coordinates and pen state, respectively. See details in text.

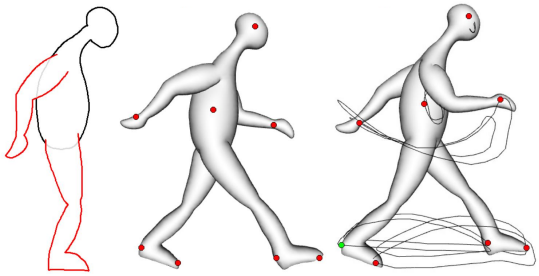
Fig. 4. Illustrations of the major domain-unique challenges of free-hand sketch. Each column is a photo-sketch pair. Sketch is highly abstract. A pyramid can be depicted as a triangle in sketch, and a few strokes depict a fancy handbag. Sketch is highly diverse. Different people draw distinctive sketches when given the identical reference, due to subjective salience (head vs. body), and drawing style.

are 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.

Unique Challenges and Opportunities The unique challenges of free-hand sketch can be summarized as follows: (i) Abstraction: 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 4, a pyramid can be depicted as a triangle in sketch, and few strokes can depict a fancy handbag. (ii) Diversity: Different people 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 concept "cat", people differ on choice of drawing with/without body (Figure 4). Finally, there is the mental viewpoint of different users, *e.g.*, whether they imagine an orthographic or perspective projection image. In Figure 4, we can see that different people draw differing perspective views of an identical slipper. (iii) Sparsity: No matter the representation, free-hand sketch is a highly sparse signal compared to photographs. (iv) Invariance: People can still recognize sketches after they are shifted, rescaled, rotated, or flipped. In Section 4.3.2, we conduct a robustness study to evaluate whether deep networks are sensitive to spatial transformations in sketch related tasks. (v) Finally, there are some unique challenges when collecting sketch, which will be discussed in detail as follows (see Section 3.2).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3148853, IEEE Transactions on Pattern Analysis and Machine Intelligence

(a) A sketch is drawn with the device on the back of the hand [52].



(b) A walking cycle sequence. Left: input hand-drawn sketch, middle: inflated 3D model with control points, right: walking cycle animation created by recording trajectories of individual control points specified by the user [53].

Fig. 5. Novel applications that sketch supports.

Sketch also provides some unique opportunities compared to photos: (i) As a counterpoint to the sparsity challenge, sketch often lacks distracting background clutter compared to photos, which can benefit automated analysis [3]. (ii) If captured appropriately, the sequential nature of sketch generation can further be exploited to benefit analysis compared to static images [54]. (iii) 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 [5], [55], [56]. (iv) Sketches can serve as a computer-interaction modality in a way that photos cannot [4], [54], due to the intuitive way humans can generate them without training. For example, people without professional painting training can do casual sketch-based design via sketch-to-photo generation techniques, *e.g.*, scene photo generation [17]. Figure 5(a) presents another example: People could sketch on the back of the hand to make notes conveniently, and the sketch could be shown and recorded in the watch [52]. (v) Sketches natively can express motion trajectories, thus can be applied to dynamic modeling. As shown in Figure 5(b), walking cycle animation can be created by recording trajectories of individual control points specified by the user [53].

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

2.2 A Brief History of Sketch in the Deep Learning Era

In the past five years, the free-hand sketch community has developed rapidly as summarized by Figure 6 from the perspectives of: tasks, datasets, representations and supervision. (i) In 2015, Sketch-a-Net [18] was proposed as a 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. (ii) In 2016, three fine-grained[6] sketch-based image retrieval (FG-SBIR) datasets were released, *i.e.*, QMUL Shoe [4] and Chair [4], and Sketchy [7]. Combined with deep triplet ranking [57], these fine-grained cross-modal datasets motivated a wave of follow-up FG-SBIR and other fine-grained tasks. (iii) 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 world, making it a rich and diverse dataset. Furthermore, based on the QuickDraw dataset, Ha *et al*. proposed "SketchRNN", a RNN-based deep Variational Auto Encoder (VAE) that can generate diverse sketches [5]. 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 [58] began to appear in the top conferences in computer vision. (iv) From 2018 to date, based on deep learning techniques, various novel methodologies – *e.g.*, sketch hashing [19], sketch transformers [23]; and applications – *e.g.*, sketch abstraction [21], sketch-based photo classifier generation [20], sketch perceptual grouping [59], and sketch vectorization [56] have been proposed. See Figure 6 for a chronological summary.

3 Free-Hand Sketch Datasets

In the last decade numerous new free-hand sketch datasets have been collected, to satisfy the need for large-scale deep network training, and the growing diversity of sketch-related tasks considered by the community. This section will summarize these datasets, and further discuss some of the unique challenges in sketch-related data collection.

3.1 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 pairing sketches with samples from 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 [6], QuickDraw [5]) are usually used for sketch recognition, sketch retrieval; while fine-grained datasets (*e.g.*, QMUL Shoe [4]) provide fine-grained visual details and manual annotations.

6. The phrase "fine-grained" in this paper has different meanings according to the context. For sketch tasks, fine-grained sketch-based image retrieval means instance-level sketch-photo matching, while other fine-grained tasks (*e.g.*, generation, segmentation) emphasize that machine needs to perceive sketches on stroke or part or group levels. For sketch datasets, "fine-grained" datasets mean that their sketches provide visual details and/or detailed manual annotations (*e.g.*, stroke/part/group level annotations, instance-level pairing information).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3148853, IEEE Transactions on Pattern Analysis and Machine Intelligence

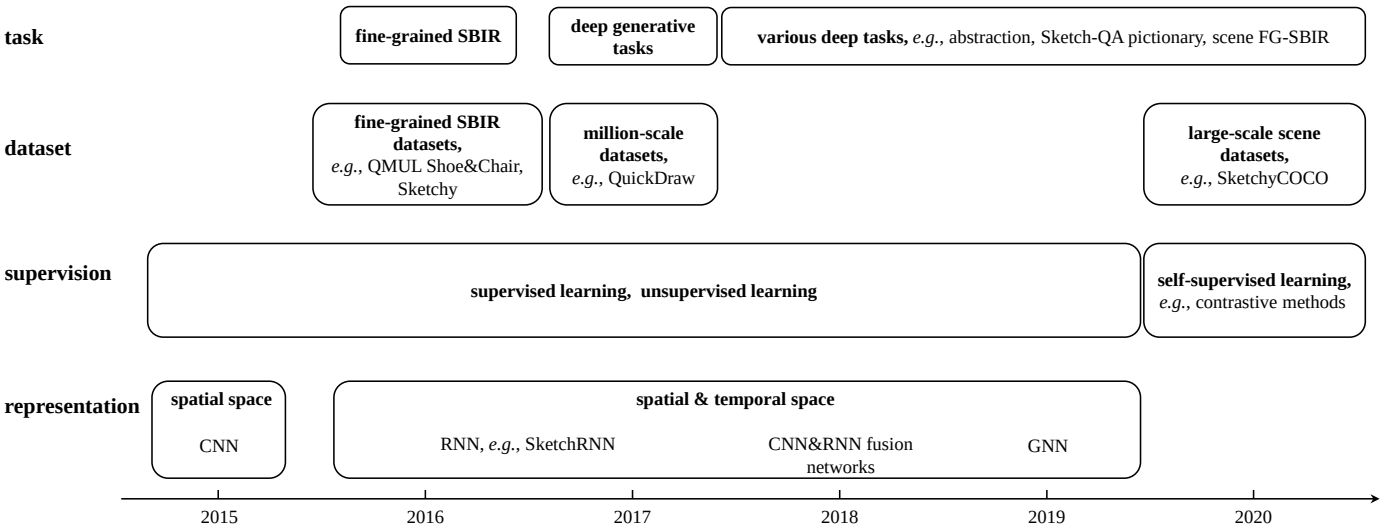


Fig. 6. Milestones of deep learning based free-hand sketch research, from the perspectives of task, dataset, supervision, and representation. Note that self-supervised learning is a branch of unsupervised learning.

TABLE 2
Summary of the representative sketch datasets. 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	Modalities & Sample Amount	Cat.	Stroke	Object/ Scene	Instance Pairing	Annotations	Remarks
TU-Berlin [6]		✓	20K sketches	250	✓	o	-	class	
QuickDraw [5]		✓	50M+ sketches	345	✓	o	-	class	
QuickDraw-5-step [60]			38M+ sketches	345		o	-	class	
SPG [59]	✓	✓	20K sketches	25	✓	o	-	class, grouping	
SketchSeg-150K [25]	✓		150K sketches	20	✓	o	-	class, segmentation	57 semantic labels
SketchSeg-10K [61]	✓	✓	10K sketches	10		o	-	class, segmentation	24 semantic labels
SketchFix-160 [62]		✓	3904 sketches	160	✓	o	-	class, eye fixation	
Sheep 10K [63]	✓	✓	10K sheep sketches	1	✓	o	-	class	
COAD [64]	✓	✓	620 sketches	20	✓	o	-	class	
Multi-Modal Datasets	**Fine-Grained**	**Public**	**Modalities & Sample Amount**	**Cat.**	**Stroke**	**Object/ Scene**	**Instance Pairing**	**Annotations**	**Remarks**
QMUL Shoe [4]	✓	✓	419 sketches, 419 photos	1		o	✓	pairing,triplet,attribute	21 binary attributes
QMUL Chair [4]	✓	✓	297 sketches, 297 photos	1		o	✓	pairing,triplet,attribute	15 binary attributes
QMUL Handbag [26]	✓	✓	568 sketches, 568 photos	1		o	✓	pairing	
Sketchy [7]	✓	✓	75K sketches, 12K photos	125	✓	o		class	12K objects
Sketch&UI [8]	✓		1998 sketches, 1998 photos	23		o	✓	class, pairing	UI
QuickDrawExtended [65]		✓	330K sketches, 204K photos	110		o		class	
SketchTransfer [66]			112.5K sketches, 90K CIFAR-10 photos	9		o		class	resolution of 32x32
TU-Berlin Extended [67]			20K sketches, 191K photos	250		o		class	
Sketch Flickr15K [1]		✓	330 sketches, 15K photos	33		o		class	
Aerial-SI [68], [69]			400 sketches, 3.3K photos	10		o, s		class	aerial scene
HUST-SI [70]		✓	20K sketches, 31K photos	250	✓	o		class	
SBSR [71]		✓	1814 sketches, 1814 3D models	161		o		class	
SHREC'13 [47]	✓	✓	7200 sketches, 1258 3D models	90		o		class	
SHREC'14 [72]	✓	✓	12680 sketches, 8987 3D models	171		o		class	
PACS DG [73]		✓	9991 (sketches, photos, cartoons, paintings)	7		o		class	domain generalization
Flickr1M [74]			500 sketches, 1.3M photos	100		o		class	
Cross-Modal Places [75]		✓	16K sketches, 11K descriptions, 458K spatial texts, 12K clip arts, 1.5M photos	205		s		class	
SketchyScene [76]	✓	✓	29K sketches, 7K photos			s	✓	pairing, segmentation	
DomainNet [77]		✓	0.6M (cliparts, infographs, paintings, QuickDraw skteches, real photos, professional pencil sketches)	345	✓	o		class	
SketchyCOCO [17]	✓	✓	14K+ (sketches, photos, edge-maps)	17	✓	o, s	✓	class, pairing, five-tuple, segmentation	3 background classes 14 foreground classes
SceneSketcher [78]	✓	✓	1225 scene sketch-photo pairs	14	✓	o, s	✓	class,pairing segmentation	

More specifically, coarse-grained single-modal datasets [5], [71] support sketch recognition and retrieval; while coarse-grained multi-modal datasets (*e.g.*, QuickDraw-Extended [65]) support category-level sketch-based im-

age retrieval. Fine-grained single-modal datasets [25], [59] support perceptual grouping, segmentation, and parsing. Fined-grained multi-modal datasets (*e.g.*, QMUL Shoe [4]) provide the instance-level pairing information to support retrieval.

Table 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 static picture files such as JPEG and PNG, whilst these static files cannot store or provide the original drawing process (stroke ordering). We exclude some well-known but overly-small datasets such as [79].

3.2 Unique Challenges of Sketch Collection

Sketch Collection Strategies Existing collection approaches mainly include: (i) bespoke creation by researchers [25], [59], [61], (ii) crowd-sourcing selecting and matching on existing datasets, *e.g.*, Doodle2Sketch QuickDraw-Extended [65], (iii) crowd-sourcing drawing from scratch, *e.g.*, QMUL Shoe [4], Sketchy [7]. (iv) collecting via online drawing games, *e.g.*, Google QuickDraw [5]. (v) Web crawling of existing sketches [73], [77]. In particular, for fine-grained multi-modal sketch datasets, crowd-sourcing is widespread, since fine-grained drawing, selection, and matching are time-consuming. Note that a sketch dataset's potential applications are determined by both its collection and annotation protocol.

Sketch Collection Challenges Free-hand sketch poses some unique data collection challenges compared to other image types: (i) **Time-Sequence Nature** Sketching is a dynamic and temporally extended process. Thus, collecting sketches as static raster images (*e.g.*, JPEG, PNG) is very limiting, and recording vector representation (*e.g.*, SVG[7]) together with stroke position and timing is preferred to support research. As a consequence, this means that collecting by web-crawling (which typically retrieves raster images), is less useful. (ii) **Cross-Modal Pairing** Collecting cross-modal datasets provides the additional challenge of pairing sketches and associated data in other modalities. One can start with existing images and sketches and pair them [80], or draw sketches specifically corresponding to given examples in other modalities [4], [7]. But both options are time-consuming. (iii) **Demographic Information** Since sketches are human-created, rather than pixel-perfect images captured by camera, it is important to ensure that sketch datasets are created by diverse demographics to ensure they are representative (an even stronger version of the challenge [81] posed by photo images). Furthermore, meta-data about the artists (*e.g.*, gender, nationality, skill level) may be important to store, both in order to study how different humans sketch, and also to ensure that any sketch-based applications are fair and unbiased [82]. However to our knowledge, this kind of meta-data has not been recorded in existing datasets.

Discussion These challenges all mean that the standard computer vision approach of web-crawling is poorly suited to collection of sketch data: It does not typically retrieve vector-graphic and time-stamped sketch generation, does not make it easy to obtain matched cross-modal pairs, and does not come with any demographic information. For these reasons, bespoke creation, crowd-sourcing, and generation as a byproduct through gamification are the recommended methods of collection.

7. https://en.wikipedia.org/wiki/Scalable_Vector_Graphics

4 Tasks and Methodology Taxonomy

In this section, we aim to provide an overview of deep learning related tasks and methods from the perspective of the whole free-hand sketch research area, rather than categorizing methods for a specific task. We observe that several trends emerging in contemporary sketch research including: (i) More novel tasks are continually being proposed, each of which has different task-specific challenges, and thus different task-specific methods/designs. This inspires us to review the existing methods from a task-driven perspective. (ii) Free-hand sketch is often associated with data from other modalities. This inspires us to categorize the existing tasks on the basis of the data modalities involved.

According to the data modalities involved, free-hand sketch related tasks can be divided into single- and multi-modal tasks, with single-modality sketch analysis techniques 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 7 provides a tree diagram of the existing free-hand sketch tasks. The main advantages of this task taxonomy include: (i) Straightforward and efficient. The tree has balanced depth and width, and does not contain redundant nodes. (ii) Extensible. The single-modality and multi-modal sub-trees are uncoupled for independent update. E.g., the single-modality sub-tree can remain fixed when we insert new modalities in future.

Some uni-modal tasks listed in Figure 7 are also studied in the natural photo domain, while the sketch-based multi-modal tasks are unique.

4.1 Uni-Modal Tasks: Pure Sketch Analysis

These 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.

4.1.1 Recognition

Sketch recognition [6] aims to predict the class label of a given sketch, which is one of the most fundamental tasks in computer vision. It has a variety of practical applications including: interactive drawing systems that provide feedback to users [83], sketch-based science education [84], games [5], [54], *etc*. Both object [5], [6] and scene [85], [86] categories have been studied from a recognition perspective. Notably sketch recognition techniques underpin the popular web game QuickDraw and WeChat mini-app Caihua Xiaoge, both released by Google.

Sketch recognition can be categorized into (i) offline and (ii) online recognition settings. Offline recognition systems take the whole sketch as input and predict a class label based on the complete sketch. Online recognition systems take the accumulated sketch strokes and continuously predict the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3148853, IEEE Transactions on Pattern Analysis and Machine Intelligence

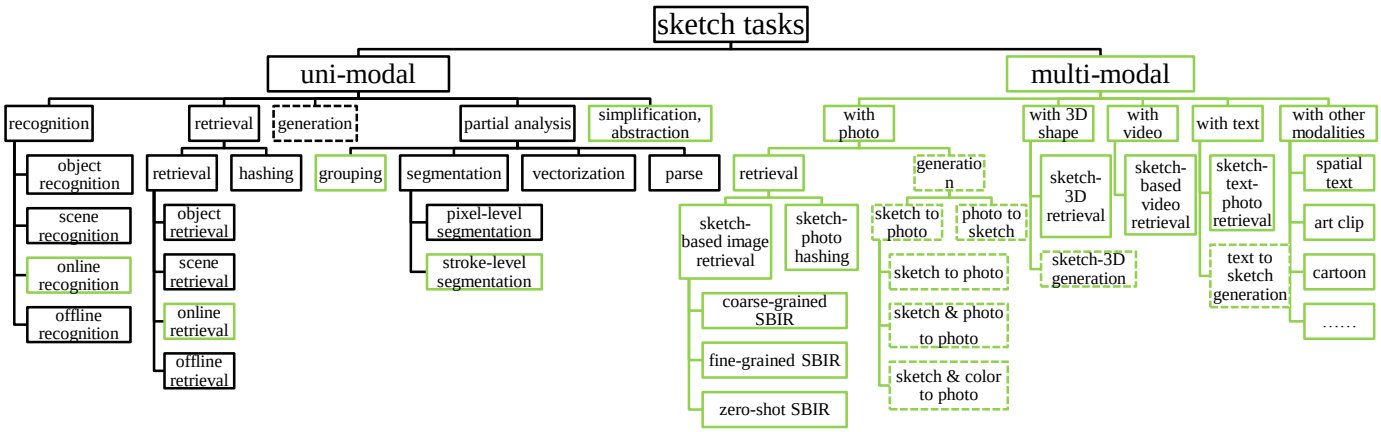


Fig. 7. A tree diagram of the sketch task taxonomy. Generative tasks are framed by dashed lines. Sketch domain-unique tasks are framed by green lines. Best viewed in color.

class label, during sketching. Offline recognition methods are more common, but online methods can be used in more interactive real-time applications such as real-time drawing guidance [60], tracing, and interactive sketch retrieval.

There are several current trends in sketch recognition, building on underpinning deep learning progress: (i) From raster image to sequence representation; (ii) From global representation to local analysis; (iii) From Euclidean (CNN, RNN based) to topological analysis (GNN based), (iv) From fully-supervised learning to self-supervised learning.

Numerous deep models have now been proposed for free-hand sketch recognition [67], [87]–[98]. In the following, we review these models from the perspectives of network architectures and loss functions. Data augmentations will be discussed separately in Section 4.1.6.

Networks Figure 6 (below) summarizes the evolution of the deep learning-based sketch representations. Moreover, Table 3 lists various networks that are engineered for free-hand sketches and capable of sketch recognition. We next introduce some representative networks.

(i) Sketch-a-Net [3], [18] was the first deep CNN designed for free-hand sketch. Compared with classic photo-oriented CNN architecture [105], the sketch-specific aspects of its architecture mainly include: (a) Considered the sparse low-texture nature of sketch, larger size (15×15) first layer filters are used to capture more context. (b) Local response normalization (LRN) [105] layers are removed for faster learning without sacrificing performance, since LRN is for "pixel brightness normalization", but most sketches are binary images. (c) Two novel sketch-specific data augmentation strategies are proposed, leveraging stroke appearance and stroke sequence. (d) Finally, an ensemble of Sketch-A-Net were combined using joint Bayesian fusion [106].

(ii) Sarvadevabhatla *et al*. [99] proposed a sketch recognition network to leverage the sequential process of sketching, where each training sketch is plotted as a continuous sequence of cumulative stroke pictures and the corresponding AlexNet [105] based deep features will be sent into a Gated Recurrent Unit (GRU) [107] network in sequence. This network is also able to work in online recognition mode, since it involves the intermediate status of the sketch. Furthermore, Jia *et al*. [101] proposed a multiple feature based model to improve this idea and obtained good performance on sketch recognition as reported in Table 3, where multiple GRU networks separately encode multiple features of the cumulative stroke groups and their outputs are combined by time-step-based weights.

(iii) Similarly, He *et al*. [102] proposed the deep visual-sequential fusion (DVSF) net to capture spatial and temporal patterns of sketches simultaneously. For each training sketch, its three accumulation sub-pictures (with 60%, 80%, 100% of strokes) go through three-way CNNs (ResNet-18 [108]) to produce deep features, which are fed into both visual and sequential networks. In particular, the visual and sequential networks are implemented by residual fully-connected (R-FC) and Residual Long Short Term Memory (R-LSTM) [109] layers respectively. The visual and sequential paths are integrated by a fusion layer for final recognition.

(iv) In 2017, Ha and Eck proposed the groundbreaking SketchRNN [5], which performs representation learning through its Variational Inference (VI) [110] based sequential sketch generation model. Distinctively different to the prior stroke accumulated sub-picture representations, the key points of sketch strokes are directly fed into the RNN backbone of SketchRNN. In particular, as illustrated in Figure 3, each key point is denoted as a vector consisting of two coordinate bits (*i.e.*, horizontal and vertical coordinates) and the corresponding flag bits. The flag bits indicate the start/end of a stroke by pen state. Although initially proposed for generative modeling, the encoder backbone of SketchRNN also performs well for sketch recognition[8].

(v) Xu *et al*. proposed the sketch hashing network SketchMate [19], where the backbone is a CNN-RNN dual branch network, using CNN to extract abstract visual concepts and RNN to model human temporal stroke order. The CNN branch takes in the raster pixel sketch pictures; and the RNN branch process the vector sketch (*i.e.*, key point coordinates). The branches are combined by a late-fusion layer. This network demonstrates the complementarity of visual and temporal embedding spaces of sketch representation. This CNN-RNN dual-branch modeling idea has been widely

8. https://github.com/payalbajaj/sketch_rnn_classification

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3148853, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 3
Comparison of representative sketch recognition networks. "–" indicates not mentioned or unclear in the original paper. Reported performance is top-1 accuracy. Abbreviations in this table: "stroke accu. pic.": stroke accumulated pictures; "R-FC": residual fully-connected layer; "pad.": padding; "tru.": truncation; "augm.": specific augmentations; "tran.": transformer.

Year	Model	Architecture	Layers	Params	Ensemble	Pretrain	Input	Preprocess	Dataset	Accuracy
2015	Sketch-a-Net [18]	CNN	5 conv.	8.5M	✓		picture	augm. [18]	TU-Berlin [6] 250 cat.	0.7490
2016	AlexNet-FC-GRU [99]	CNN-to-RNN cascaded	–	–			stroke accu. pic.	–	TU-Berlin 160 cat.	0.8510
2018	SketchMate [19], [100]	RNN	2 GRU	–			stroke vector	tru. & pad. [19]	QuickDraw 3.8M [19]	0.7788
2018	SketchMate [19], [100]	CNN-RNN dual-branch	5 conv. & 2 GRU	–			picture & stroke vector	tru. & pad. [19]	QuickDraw 3.8M [19]	0.7949
2017	Jia *et al*. [101]	RNN-RNN dual branch	–	–	✓	✓	CNN features of stroke accu. pic.	reflection, rotation, *etc*.	TU-Berlin	0.9220
2017	DVSF [102]	R-FC and RNN dual branch	–	–		✓	CNN features of stroke accu. pic.	–	TU-Berlin	0.7960
2018	FBin DAB-Net [103]	binary CNN	–	–			picture	–	TU-Berlin	0.7370
2018	RNN→CNN [104]	RNN-to-CNN cascaded	2 LSTM & 5 conv.	–		✓	stroke vector	augm. [3]	TU-Berlin	0.7849
2019	multi-graph tran. [23]	GNN	4 tran.	10M			stroke vector	–	QuickDraw subset [23]	0.7070

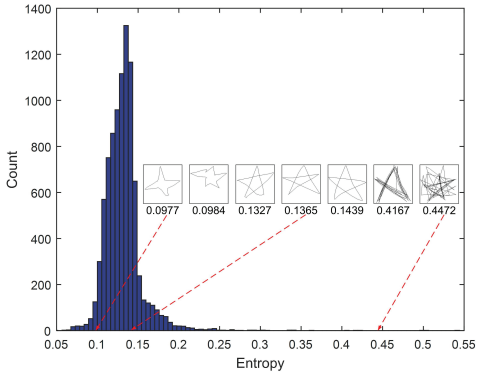


Fig. 8. Image entropy histogram of 9K sketch 'stars' [19]. The blue bars denote the bin counts within different entropy ranges. Some representative sketches corresponding to different entropy values are illustrated.

applied to other sketch tasks, *e.g*., SPFusionNet [61] for sketch semantic segmentation. In addition to the parallel pipelines of CNN and RNN, some cascaded pipelines (*e.g*., RNN-to-CNN [104]) have also been studied.

(vi) A sketch can be represented as the sparsely connected graphs in topological space. Multi-Graph Transformer (MGT) [23] is a GNN model that learns both geometric structure and temporal information from sketch graphs. MGT injects domain knowledge into Graph Transformers[9] through sketch-specific graphs. In particular, MGT represents each sketch as multiple intra-stroke and extra-stroke graphs, to model its local and global topological stroke structure, respectively.

(vii) While prior approaches to sketch recognition are based on supervised learning (*e.g*., [3], [99], [102]), [111] provided the first investigation of self-supervised representation learning for sketch, proposing a rotation- and deformation- based deep self-supervised model (rot. & def. model). This model uses multi-branch CNN and TCN network to represent sketch in a self-supervised setting.

Loss Functions Most of the previous deep sketch recognition methods use cross-entropy softmax loss to train deep neural networks. An active research question is whether sketch-specific loss functions can help to further improve recognition performance. To that end, Xu *et al*. proposed the sketch-specific center loss [19] for million-scale sketches, based on a staged-training strategy. The basis is that the image entropy distribution of each sketch category is a truncated Gaussian distribution (see Figure 8 for an example). Inspired by classic Bayesian decision theory [112], Mishra *et al*. proposed a novel metric loss to drive the pretrained deep neural network to minimize the Bayesian risk of misclassifying sketch pairs that were randomly selected within each mini-batch [113]. Based on this Bayesian risk loss, sketch recognition needs two-stage training. After obtaining the features, a linear SVM [114] is trained as the classifier.

Summary and Discussion In this subsection, we reviewed sketch recognition related deep learning works, from the perspective of architecture and loss functions. Promising areas of future work include: Online sketch recognition, motivated by practical human-computer interaction applications; going beyond the mainstream line of supervised sketch recognition to investigate semi-supervised, self-supervised [115] and unsupervised learning for recognition; zero-shot [86] and few-shot sketch recognition [116]; and multi-task learning [93], [117] to simultaneously solve sketch recognition with other tasks.

4.1.2 Retrieval and Hashing

Sketch retrieval [60], [118], [119] methods aim to use a query sketch to retrieve similar samples from a gallery or database of sketches. Sketch retrieval is a challenging task due to abstraction, intra-class variation, drawing style variation, and feature sparsity. These properties make it difficult to localize repeatable feature points across sketches (*e.g*., the manner of SIFT [120]) in order to perform classic interest-point based retrieval approaches [120]. In the deep learning era, end-to-end feature learning has outperformed shallow features on various retrieval tasks in computer vision, and CNNs have also been used for sketch retrieval.

Common practice in image retrieval is to use CNNs to learn a vector embedding, and then perform retrieval/matching as Nearest Neighbor search. Most existing deep sketch retrieval models work in a similar metric learn-

9. Transformer is essentially a GNN that encodes input as a fully-connected graph.

ing manner, with research focusing on CNN architectures and loss function designs for effective sketch matching. Wang *et al*. [118] proposed a representative sketch retrieval pipeline, which has two key components: A pure convolutional layer based Siamese CNN backbone, and an ℓ_1 norm distance based pair-wise loss between query and gallery images. The idea is two-fold: (i) Use the convolutional feature map to preserve the spatial information for sketches without point correspondence. (ii) Compute distance in feature space, and optimize for similar pairs to be nearby while different pairs to be far apart.

Given the growing number of images available, there is also an increasing concern about scalability of retrieval, leading to studies of hashing-based methods where all sketches are encoded and searched as binary hash code vectors, rather than real-valued vectors. Xu *et al*. [19] proposed the first deep sketch-hashing model. Their deep sketch hashing used a dual-branch CNN-RNN network to exploit both global appearance and local sequential stroke information, as well as a new center loss variant to ensure the learned embedding is more semantically meaningful.

The sketch retrieval/hashing methods mentioned so far exploit supervised information. If class labels are unavailable, adversarial training can be used to learn a feature representation for sketches. Based on the Generative Adversarial Network (GAN) [121], Creswell *et al*. [119] proposed Sketch-GAN for unsupervised sketch retrieval, where both the query and gallery sketches are represented by the output features of the discriminator network.

4.1.3 Generation

Sketch generation [5], [56], [126]–[131] has grown rapidly in recent years as deep learning-based approaches easily outperform earlier classic sketch generators [132], [133]. Sketch generation has several practical applications, *e.g*., synthesizing novel pictures, assisting artist design, and finishing incomplete sketches and games [131]. It can be addressed using various deep learning tools, *e.g*., VAE [5], [126], [130], [131], [134], GAN [122], VAE-GAN [122], Bidirectional Encoder Representations from Transformers (BERT) [135], and reinforcement learning (RL) [122], [125]. We compare these pipelines and their representative models in Table 4. Most of these pipelines are flexible and able to use GRU, LSTM, Transformer as backbones to achieve the stroke-by-stroke generation.

The seminal model SketchRNN [5] is a sequence-to-sequence VAE for conditional and unconditional generation of vector sketches. Its encoder and decoder are implemented by bidirectional RNN [136] and unidirectional RNN, respectively. As stated earlier, free-hand sketches can be represented as a sequence of keypoints 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.

As shown in Figure 9, the VAE encoder of SketchRNN takes vector sketches as input, and encodes it as a vector $\mathbf{h}$, 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 $N(\mu, \sigma)$, from which a latent vector $\mathbf{z}$ will be sampled. Then, the LSTM based VAE decoder will generate the coordinates and pen states of the key stroke points, conditional on $\mathbf{z}$. 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*. [137] 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 [56], [130], [138] or reinforcement learning [122], [125], [131], [139], [140] to train policies that draw sketches iteratively according to different criteria such as adversarial training against human sketches [139]. 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 [141], [142].

Going forward, there are several emerging trends in sketch generation, notably: (i) Fine-grained sketch generation [134]. (ii) A novel evaluation metric "Ske-score" [122], aims to provide a better metric to quantify the goodness of generated vector sketches. (iii) Transformer-based architectures [123], [143] are being applied to sketch generation. (iv) Competitive generation, which aims to render understandable sketches in the fewest possible strokes [131]. (v) Finally there is scaleable vector-graphic generation [56], [130], which aims to generate sketch strokes via parametric strokes rather than standard waypoint lists.

4.1.4 Grouping, Segmentation, and Parsing

Compared with sketch recognition, retrieval, and generation, there are several more fine-grained single-modal sketch analysis tasks: perceptual grouping, segmentation, and parsing. These tasks need sketch analysis at the local (stroke) level. Besides their intrinsic interest, these local sketch understanding techniques can also benefit other global tasks such as sketch-based image retrieval, sketch-based video retrieval [144], and sketch generation/synthesis. We next review recent advances in these areas.

Sketch Perceptual Grouping (SPG) Humans have the ability to perceptually group visual cues into semantic object parts/components, which has been widely researched in Gestalt psychology area [145], [146]. Humans are able to perceptually group sketch strokes into semantic parts, *e.g*., airplane strokes grouped into fuselage and wings. Thus, sketch perceptual grouping (SPG) is to imitate the human ability to group strokes into semantic parts. SPG has been studied with pre-deep learning methods [147]–[149], however progress has advanced rapidly since then. One representative application of SPG is to simplify sketches [150]. Moreover, SPG can also be used for sketch recognition [151], sketch semantic segmentation, synthesis [133], retrieval, fine-grained sketch-based image retrieval (FG-SBIR), sketch-based video retrieval [144], *etc*.

Li *et al*. [59], [152] contributed the largest SPG dataset to date of $20,000$ manually-annotated sketches across 25 object categories, and propose a universal deep grouper that can be applied to sketches of any category. Specifically, this deep universal grouper is also a sequence-to-sequence VAE with both generative and discriminative objectives: (i) Its generative loss provides the ability to handle unseen object

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3148853, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 4
Comparison of the representative sketch generation deep models. "gen.", "rec.", and "com." denote "generation", "reconstruction", and "completion", respectively. "RNN+VAE" means "RNN backbone based VAE".

Pipelines	Representative Ref.	Applications	Advantages & Disadvantages
RNN+VAE	SketchRNN [5]	gen., rec., com.	A: brief, flexible. D: scribble effect, single-class gen.
RNN+GAN	SkeGAN [122]	gen., com.	A: less scribble effect, faster convergence. D: single-class gen.
RNN+VAE+GAN	VASkeGAN [122]	gen., rec., com.	A: less scribble effect. D: high complexity, single-class gen., slower convergence
BERT	Sketchformer [123] Sketch-BERT [124]	gen., rec., com.	A: good at handling longer stroke sequences D: more parameters, high complexity
CNN+RL	Doodle-SDQ [125]	imitating a reference	A: can handle unseen classes. D: hybrid training (supervised learning & RL), high complexity
VAE+Renderer	Cloud2Curve [56]	gen. rec.	A: Scaleable vector sketch generation. Long sketches. D: High complexity.

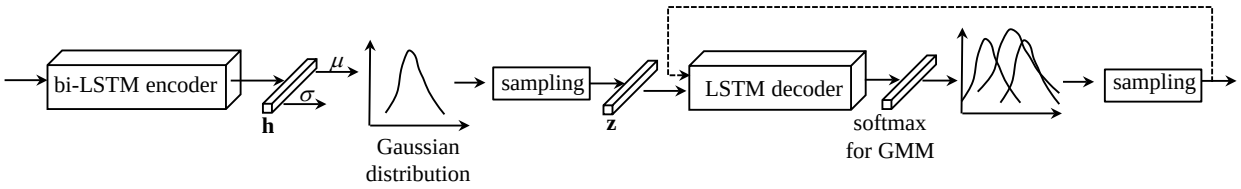


Fig. 9. The pipeline of SketchRNN [5]. The dashed arrow line denotes the recurrent processing of LSTM decoder. For simplicity, the recurrent processing of bi-LSTM encoder is not shown here.

categories and datasets. (ii) Its discriminative loss consists of local and global grouping losses, to guarantee both local and global consistency in the grouping outputs.

Discussion Shallow grouping methods mainly relied on thresholding low-level geometric properties among the strokes, often resulting in strokes with similar geometry but different semantics being grouped. Contemporary SPG methods consider more high-level semantic and temporal information due to their deep and recurrent representations.

Sketch Semantic Segmentation (SSS) Sketch semantic segmentation has drawn attention [147], [153], [154] in the free-hand sketch community as a classic topic prior to deep learning.

Sketch semantic segmentation can potentially be addressed by conventional photo segmentation CNNs. However, these do not exploit the vector representation of strokes or their temporal patterns, loading to the development of sketch-specific segmentation models.

Existing deep models for sketch semantic segmentation [24], [25], [61], [155]–[157] can be grouped according to architectures: CNN, RNN, GCN based models, *etc*.

Li *et al*. [158] trained a CNN-based network to transfer well-annotated segmentation and labels from a 3D dataset to sketch domain. They used annotated 3D data [153], [159] to produce edge-maps with partial annotations as the synthetic sketch data to train the segmentation network.

Qi *et al*. proposed SketchSegNet [157] and SketchSegNet+ [25]. SketchSegNet+ [25] considers sketch stroke orderings and is able to process multiple object categories. In particular, SketchSegNet and SketchSegNet+ work in an RNN-based VAE pipeline, where the Gaussian mixture model (GMM) layers of SketchRNN are replaced with fully-connected softmax layers to predict the part labels. StrokeRNN [156] uses the same encoder as SketchRNN but extends the decoder to predict segmentation.

Besides the RNN backbone, other backbones have also been explored on sketch segmentation. SPFusionNet [61] uses late fusion of CNN-RNN branches to represent sketches for segmentation. SketchGCN [24] is a graph convolutional neural network for sketch semantic segmentation. It uses a mixed pooling block to fuse the intra-stroke and inter-stroke features from its two-branch architecture.

Discussion SPG essentially performs stroke-level clustering, while SSS provides stroke-level classification. *i.e*., SSS provides explicit part labels (category names) for each stroke, while grouping only provides aggregation relationships. SSS and SPG are analogous to the classic (supervised) semantic segmentation [160] and unsupervised segmentation [161] respectively [59]. Therefore, SSS needs stronger supervision during training, namely stroke categories, in contrast to SPG's stroke grouping.

Vectorization Sketch vectorization is a widely studied topic for well-drawn pencil sketches (particularly pencil-and-paper scanned sketches) [142], [155], [162]–[167]; with a few studies begining to address it for free-hand sketches [56], [115], [130]. It aims to generate vector representations for raster sketch photographs. Sketch vectorization is essentially different from sketch semantic segmentation, in that sketch vectorization aims for instance segmentation on the stroke level, rather than semantic classification for a stroke or a stroke group.

Sketch Parsing Recently, a new concept of "sketch pars-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3148853, IEEE Transactions on Pattern Analysis and Machine Intelligence

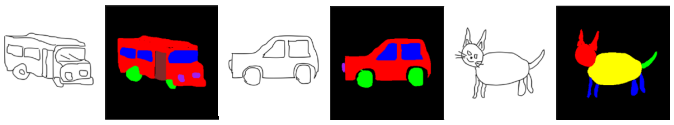


Fig. 10. Sketches (bus, car, cat) and ground truth annotations selected from sketch parsing paper [171]. The semantic parts and background are annotated by colors. Best viewed in color.

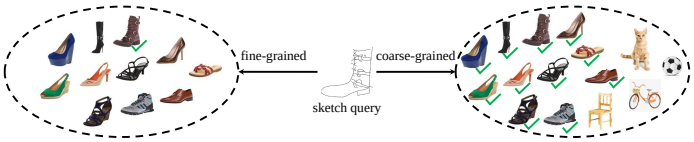


Fig. 11. Comparison between fine-grained (instance-level) and coarse-grained (category-level) sketch-based image retrieval. True match photos are ticked.

ing" [168]–[171] has gained traction. As a kind of fine-grained semantic understanding of sketch, sketch parsing has already been applied to assist other sketch tasks [168], *e.g.*, sketch-based image retrieval (SBIR). Sketch parse is related to sketch semantic segmentation. However, as shown in Figure 10, the goal is to perform pixel-wise segmentation of the semantic regions defined by the sketch, rather than the sketch strokes as in SSS. Existing models for sketch parsing thus far only use CNN-base networks to represent sketch the, *e.g.*, SFSegNet [169] uses Deep Fully Convolutional Networks (FCN) [172].

4.1.5 Simplification and Abstraction

Sketch simplification has been widely studied [150], [173]–[175] by the computer graphics community to simplify sketches by merging redundant strokes [175]. A typical pipeline [176] is two-stage: geometrically clustering strokes into groups (*e.g.*, by Gestalt principles); and generating a new line to replace each group.

With the prevalence of deep learning, CNN-based sketch representations have been used for sketch simplification. Simo-Serra *et al.* [16] proposed a fully-convolutional network (FCN) for simplifying sketches directly from raster images of rough sketches, where a pixel level mean square error (MSE) loss is used to compare the training pairs of rough (input) and simplified (target) sketches. This approach is fully automatic and requires no user intervention. Thanks to these advantages, this FCN based model was further studied. However, this fully supervised approach needs large amounts of supervised pairs of rough sketches and their corresponding sketch simplifications as annotation. To alleviate this limitation, Simo-Serra *et al.* [177] integrated their FCN model into the generative adversarial pipeline, where a fully-convolutional network works as the generator. This upgraded model can be jointly trained from both supervised and unsupervised data, and obtained significant performance improvements. To further study how discriminator networks can improve sketch simplification, Xu *et al.* [178] proposed a multi-layer discriminator by fusing all VGG [179] feature layers to differentiate sketches and simplify lines, where weights used in layer fusing are automatically optimized via an intelligent adjustment mechanism. The experimental results demonstrate that this multi-layer discriminator helps the FCN based generator further improve its simplification performance. Comparing the experimental results of these three representative methods [16], [177], [178], suggests that pixel-level losses (*i.e.*, MSE loss) and *Vanilla* discriminator loss may fail to provide adequate supervision to help models retain semantically meaningful details when simplifying relatively complicated sketches.

A different take on simplification focuses on the epitome of a sketch [180]. This was recently studied under the guise of "stroke-level sketch abstraction" in the free-hand sketch community. Stroke-level abstraction [21], [181] aims to abstract sketches by removing strokes that do not affect the recognizability of the sketch. Solving this problem provides several benefits: (1) It learns stroke saliency as a byproduct [21] – strokes that contribute the most to recognizability are the most salient. (2) It can be used to synthesize sketches of variable abstraction for generation, or data augmentation of discriminative sketch models [21]. (3) It can be used for summarization and compression more broadly [181].

The stroke-level abstraction task can be seen as a discrete combinatorial optimization problem, and thus is intractable to solve with traditional methods. This was tackled in [21] by training a reinforcement learning (RL) policy to include/exclude each stroke in the sequence, while trading off between the number of included strokes and recognizability. The RL-based abstraction idea was extended by [181] to reorder input strokes, rather than being constrained to the original input sequence; and further to enable customizing of the abstraction goal to preserve different aspects of 'recognizability' such as category vs. attributes.

Discussion Despite this good progress, simplification through *merging* multiple strokes into a coarser replacement, rather than simply filtering them, remains an open question for deep learning-based sketch analysis.

Discussion A bottleneck for sketch simplification is that in the existing literature the experimental results are mainly evaluated by visual comparison and user studies. Defining a good quantifiable and automatic metric to evaluate simplified sketches is an open problem and a big challenge. A good metric would be of great help in designing more well-defined loss functions for this task.

4.1.6 Data Augmentations

The sketch-specific data augmentation methods discussed in this subsection can be applied to both sketch recognition and all the other sketch involved tasks (both single-modal and multi-modal), *e.g.*, sketch-based image retrieval, sketch-related generation.

(i) When represented as raster pictures, most common data augmentations designed for natural photos [182] can be applied to sketch, *e.g.*, horizontal reflection/mirroring, rotation, horizontal shift, vertical shift, central zoom. These augmentations have already been evaluated by the early sketch-oriented deep learning works [18], [183]. However, random cropping is likely unsuitable for sketch since partial sketches are often too sparse to recognize even for humans, and some image enhancement methods based on statistics like contrast/histogram/brightness enhancement cannot be applied to sketches.

(ii) Stroke thickening/dilation can be used for free-hand sketch. As discussed in some previous works on spatially-sparse convolutional neural networks [184], the subtle details of sparse sketch strokes can be lost after multiple layers of convolution. Thus, this can be useful for deep neural networks that process sketches as image inputs.

(iii) Yu *et al*. [3] proposed to remove the strokes to obtain more diverse sketches. Based on the observation [6] that humans tend to draw outlines first before the detail, heuristics can be proposed to remove strokes with probability dependent on their sequential order.

(iv) Zheng *et al*. [185] proposed a Bezier pivot based deformation (BPD) strategy and a mean stroke reconstruction (MSR) approach. These do not need any temporal information in the sketch. The main idea of MSR is to generate novel sketches with smaller intra-class variance.

(v) Liu *et al*. [186] proposed two sketch-specific data augmentation strategies: (a) Manually extract some strokes from sketch SVG files to construct noise stroke masks. Then, randomly apply the noise stroke masks to the original sketches to synthesize augmented sketches. (b) Randomly extract a patch from one sketch, and attach it to a given sketch.

(vi) Muhammad *et al*. [21] applied reinforcement learning to learn a sketch abstraction model that preserves the semantics of the sketch. Once trained, this model can be applied to generate augmentations of an input sketch at different abstraction levels.

Discussion Compared with augmentations on full images (*e.g*., rotation, shift), the augmentation strategies above make better use of stroke information both locally and globally. However, only [21] makes (limited) use of human sketch variability to perform augmentation. In future an interesting direction is to learn from the variation in sketch style between different humans, and treat sketch augmentation as a cross-human style transfer problem.

4.2 Multi-Modal Tasks: Sketch with Other Modalities

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 text domains.

Nowadays, most visual retrieval approaches work under the "query-by-example" (QBE) [187] 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 [188], 3D shape, video, *etc*.

4.2.1 Sketch-Photo Retrieval

Sketch-photo retrieval is also known as sketch-based image retrieval (SBIR) [1], [4], [189]–[195][10]. SBIR is challenging for all the reasons that sketch-analysis in general is challenging (sparse and abstract input). It is particularly challenging 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 7 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 [196], and fine-grained SBIR (*i.e*., instance-level SBIR). FG-SBIR is essentially a kind of instance-level retrieval [197]. From the perspective of retrieval embedding space, SBIR can be 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.

Category and Instance Level SBIR In coarse-grained SBIR, given a sketch as query, a ranked list of images 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. Figure 11 provides an illustration. Based on SBIR ideas, several sketch-based commodity search engines have been implemented, *e.g*., sketch-based skirt image retrieval [198], fine-grained sketch-based shoe [4], [26], chair [4], [26], and handbag [26] retrieval systems.

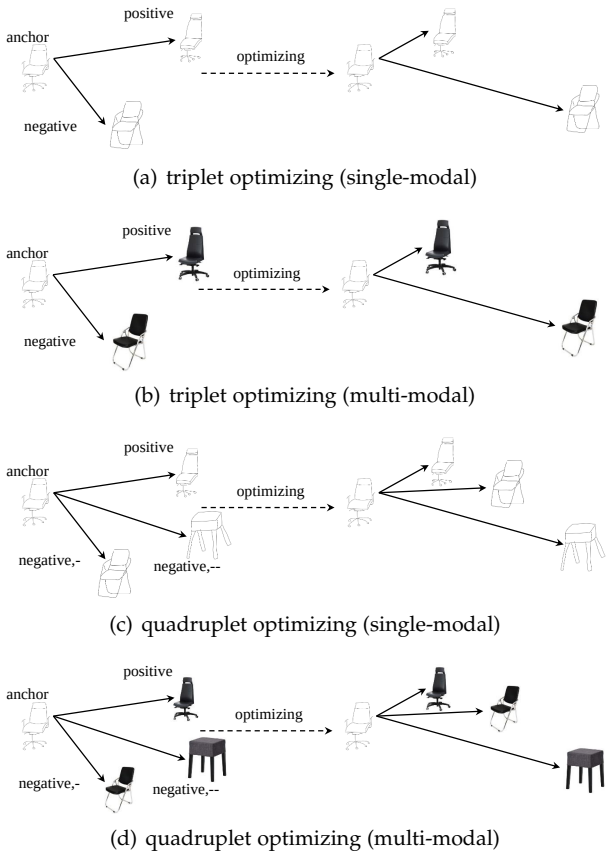
Some previous SBIR works [199]–[201] have used edge-maps (image contours) of photos as an approximation to corresponding sketch images in order to perform matching. Canny edge detector [202], Edge Boxes toolbox [203], and holistically-nested edge detection (HED) [204] were usually used to extract the edges from natural photos. However, this kind of hand-designed process is now commonly replaced by end-to-end deep feature learning.

Deep sketch-based image retrieval (SBIR) has been widely studied [4], [7], [205]–[217] 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 [118]), (b) ranking based methods [4], [7], (c) reinforcement learning based methods [218], (d) deep canonical correlation analysis (DCCA) [219] based methods [220], (e) cross-domain dictionary learning [221], *etc*. The most widely-studied methods are ranking-based, including triplet ranking [209], [222], [223] and quadruplet ranking [224].

Ranking-Based SBIR We next introduce the popular triplet- and quadruplet-ranking SBIR methods in detail. As shown in Figure 12, given a sketch anchor $\mathbf{X}_n$ and its positive and

10. Note that the common setting for SBIR is sketch as a query modality for images, but most methods enable either modality to be used as a query if desired.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3148853, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) triplet optimizing (single-modal)

(b) triplet optimizing (multi-modal)

(c) quadruplet optimizing (single-modal)

(d) quadruplet optimizing (multi-modal)

Fig. 12. Illustration of triplet and quadruplet ranking based optimization objectives. The lengths of solid arrows illustrate the distances in embedding spaces. In the quadruplet illustration, the "$negative, -$" sample denotes the negative sample from the anchor category, while the "$negative, --$" one denotes the negative sample from the remaining categories.

negative photo retrieval candidates ($\mathbf{X}_{n,+}$, $\mathbf{X}_{n,-}$), the goal of triplet ranking is

$$\mathcal{D}(\mathcal{F}(\mathbf{X}_n), \mathcal{F}(\mathbf{X}_{n,+})) < \mathcal{D}(\mathcal{F}(\mathbf{X}_n), \mathcal{F}(\mathbf{X}_{n,-})). \quad (1)$$

where $\mathcal{D}(\cdot,\cdot)$ is a distance metric (*e.g.*, ℓ_2 distance). In common FG-SBIR practice [4], [209], the negative sample is usually selected from the same class as the anchor.

For quadruplet ranking [224], 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. As illustrated in Figure 12, the goal of quadruplet ranking is to ensure

$$\begin{aligned}\mathcal{D}(\mathcal{F}(\mathbf{X}_n), \mathcal{F}(\mathbf{X}_{n,+})) &< \mathcal{D}(\mathcal{F}(\mathbf{X}_n), \mathcal{F}(\mathbf{X}_{n,-})) \\ &< \mathcal{D}(\mathcal{F}(\mathbf{X}_n), \mathcal{F}(\mathbf{X}_{n,--})).\end{aligned} \quad (2)$$

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. For example, Seddati *et al*. [224] constructed three triplets from each quadruplet, including $triplet_a = \{\mathbf{X}_n, \mathbf{X}_{n,+}, \mathbf{X}_{n,-}\}$, $triplet_b = \{\mathbf{X}_n, \mathbf{X}_{n,+}, \mathbf{X}_{n,--}\}$, and $triplet_c = \{\mathbf{X}_n, \mathbf{X}_{n,-}, \mathbf{X}_{n,--}\}$. Therefore, the quadruplet ranking loss is defined as

$$\mathcal{L}_{quadruplet} = \mathcal{L}_{triplet_a} + \lambda_b \mathcal{L}_{triplet_b} + \lambda_c \mathcal{L}_{triplet_c}, \quad (3)$$

Fig. 13. Examples of SceneSketcher dataset [78], a fine-grained scene-level sketch dataset.

where λ_b, and λ_c are the weights.

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

Recently, Liu *et al*. released the first scene-level fine-grained SBIR dataset, *i.e.*, SceneSketcher [78] (see Figure 13). This seminal work opens a novel direction for the future SBIR research, and contributes an effective solution to fine-grained scene sketch cross-modal matching that uses a GCN to encode the layout information of scene sketches in fine-grained triplet ranking.

Comment The essential principle of triplet loss is using local partial orderings to establish a global ordered relationship in the embedding space. Thus triplet ranking [57] can be understood as Topological Sorting[11]. The triplet annotations work as a partially ordered set. Compared with other loss functions, the main advantages of triplet loss are: (i) It helps to involve more local partial orderings and annotations to learn more fine-grained embedding space. (ii) 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 networks on smaller sketch datasets. It should be noted that the performance of triplet loss is heavily dependent on (a) the choice of margin parameter and (b) the triplet construction strategy.

Comment We remark that ranking-based SBIR models can also be improved by multi-task training along with classification [7], [225], [226]. Furthermore, rather than purely discriminative training, SBIR can also be tackled by generating one modality from the other [227], *e.g.*, using conditional GAN [228]; or using generative losses to regularize discriminative training [229]. SBIR training can also be combined with post-processing re-ranking [199], [201], [230] to refine the initially 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 14, both triplet and quadruplet ranking models can use backbone networks that are Siamese, semi-heterogeneous, or heterogeneous. (i) Siamese networks [4] use full weight-parameter sharing across branches. (ii) Semi-heterogeneous network [231], [232] use partial weight sharing across the branches. Typically early layers are modality-specific, and weight-tied layers are at deeper layers. (iii) In heterogeneous networks [233], the sketch branch uses independent parameters to the photo branches. The trade-offs underlying these architectures are that sharing weights enables more data (both sketch and photo) to be used to

11. https://en.wikipedia.org/wiki/Topological_sorting

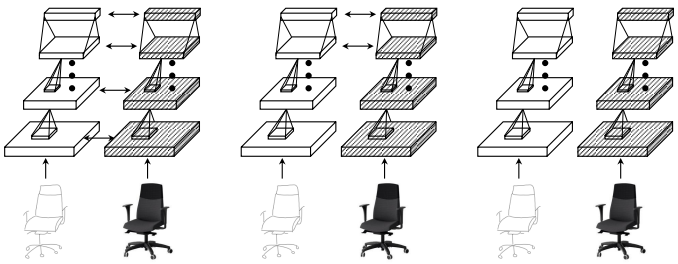


Fig. 14. Different weight sharing manners (left: Siamese, middle: semi-heterogeneous, right: heterogeneous) for CNN-based cross-modal networks. The hollow and shaded networks denote the branches for sketches and photos, respectively. The double sided arrows indicate sharing of weights.

estimate parameters, reducing overfitting. 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 [225].

Hashing-Based SBIR In order to achieve faster sketch-based image retrieval, recent research has studied optimizing the feature coding (*e.g.*, sketch-image hashing [234], [235]), and the feature map (*e.g.*, Asymmetric Feature Maps [236]).

In particular, sketch-image hashing (or hashing SBIR) has gained attention. Liu *et al*. [234], propose the first deep hashing model for SBIR, which is a classic deep hashing pipeline including: (i) feature extractor network, (ii) hashing layer with binary constraints, and (iii) hashing loss. This classic pipeline has been widely studied in photo-oriented deep hashing [237], [238], 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 benchmarks, *i.e.*, Sketchy [7] (75K sketches) and TU-Berlin Extended [67] (20K). The scale of these benchmarks is not yet large enough to thoroughly test hashing SBIR methods.

Discussion Current issues in SBIR include: Self-supervised pre-training for SBIR [115], [239], optimizing SBIR for early retrieval using partially drawn sketches, for example using reinforcement learning [218]; investigating whether costly sketch-photo annotation pairs can be replaced with edge-maps [240] and cross-category generalization of SBIR which is discussed next.

4.2.2 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 setting for supervised photo retrieval [241], zero-shot sketch-based image retrieval (ZS-SBIR) has also been studied [65], [216], [242]–[255]. Similar to natural photo zero-shot learning/recognition [256]–[258], 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.

Discussion ZS-SBIR systems can follow conventional zero-shot learning methods [256]–[258] in exploiting auxiliary knowledge such as word vectors [259], attributes [260], 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 [256] and inconvenient need to specify nameable categories at testing-time [256]). 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* [73] manner. That is, training a matching network on the training categories that is robust enough to support direct application to unseen testing categories. Thus, common approaches are to train ranking [65], [251] or generative [245], [246] models for retrieval, which are enhanced and made robust by constraints such as domain-alignment losses [65], [245], [251] and auxiliary semantic knowledge reconstruction [65], [245]. In these cases the auxiliary semantic knowledge is only used to constrain representation learning at train time and is not required during testing 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 [245], [251]; extending sketch-photo hashing to the zero-shot setting [261]; and training SBIR without paired samples [244].

4.2.3 Sketch-Photo Generation

Sketch and photo based mutual generation (translation/synthesis) is a classic cross-modal topic of sketch research covering both: (i) sketch-to-photo generation [17], [273], [274], (ii) photo-to-sketch generation [55], [271], [272], [275]. In particular, sketch-to-photo generation methods have addressed: (a) sketch to photo [267], (b) sketch & photo to photo [22], [263], (c) sketch/edge & color to photo [58]. Sketch and photo based generation can be used to help users to create or design novel images in various practical applications: sketch-based photo editing [22], [263], [264], sketch to painting generation [276], cloth design [277], [278], sketch to natural photo generation [266], *etc*. In some cases, sketch-photo generation also involves style transfer [268], [279].

Note that: (i) Sketch-to-photo generation aims to solve cross-modal translation from abstract and sparse line drawings to pixel space, different to well-drawn sketch colorization [280]–[282]. (ii) Photo-to-sketch generation does not refer to extracting edge-map from natural photos [203], [283] (edge-maps of literal perspective projections), but instead needs models that learn to mimic human sketching and abstract drawing style [55].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3148853, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 5
Comparison of the representative pipelines of deep sketch-photo generation. "s → p" and "p → s" denote "sketch → photo" and "photo → sketch", respectively. The backbones of the representative references here are implemented by CNNs.

Tasks	Pipelines	Representative Ref.	Sketch-Specific Designs	(Dis)Advantages
s → p	GAN	[262]–[264]	Sketch-specific designs are generally injected in the generators.	A: simple, end-to-end training D: suboptimal performance
s → p	GAN #1 → GAN #2	[265], [266]	GAN #1 for stroke refinement, GAN #2 for photo synthesis	A: clear motivation D: multi-stage training
s → p	CGAN	[267], [268]	Specific designs and domain knowledge can be injected as conditions.	A: clear motivation D: sensitive to conditions
s → p	ContextualGAN	[269]	learns joint distribution of sketch-photo pairs	A: appearance freedom, less strict alignment D: multi-stage training
s → p	TextureGAN	[270]	supports local texture constraints	A: more fine-grained D: high complexity
p → s	CGAN	[271]	Specific designs and domain knowledge can be injected as conditions.	A: clear motivation D: sensitive to conditions
p → s	conditional encoder-decoder	[272]	Conditions a convolutional decoder on a class prior.	A: simple, end-to-end training D: suboptimal performance
p → s	VAE (CNN encoder + RNN decoder)	[55]	shortcut cycle consistency, RNN decoder	A: stroke by stroke, end-to-end training D: suboptimal for long strokes

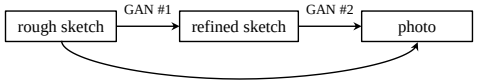


Fig. 15. Illustration of two-stage two-GAN based sketch-to-photo cross-modal generation idea [265], [266]. The arrow lengths denote the distances of cross-domain gap. See text for details.

Sketch and photo generation methods have been widely studied [186], [262], [265], [284], [285] based on various GAN [121] variants including conditional GAN [286], cycle GAN [284], and texture GAN [270]. Meanwhile, VAE also works well for sketch-photo generation. We compare the existing deep sketch-photo generation pipelines and their representative models in Table 5. The domain gap from sketch to photo is large. As demonstrated in Figure 15, an intuitive idea is to decompose the large gap into two smaller gaps. Some previous works [265], [266] proposed to use two GANs to achieve sketch-to-photo generation in two steps: (i) use the first GAN to refine rough sketches, *e.g.*, generating good contours [266], and (ii) input the refined sketches into the second GAN to generate the target photos.

Discussion Sketch-photo cross-modal generation is distinctively different from conventional photo-to-photo translation [285]. Existing photo-to-photo models can assume pixel-wise alignment between inputs and outputs. However, this requirement is strongly violated in the case of sketches and photos. Indeed no simple warping can provide pixel-wise alignment between sketch and photo given the potentially abstract or iconic nature of sketches. Thus, existing sketch-photo synthesis work has made efforts to work around this issue, *e.g.*, using contextual GAN [269].

Discussion Generating sketch images as pixels suffers from the problem that blurriness in an image that should be made of sharp edges is very visible. However, an important difference between sketch-photo translation and photo-photo translation is that the raw format of sketch data is often a time-series of way-points. If such raw sketch representation is to be used, the encoder or decoder should be an RNN rather than CNN. The first example of such sequential vectorized photo-sketch translation is in [55], where sharp sketches are sequentially produced using a recurrent decoder.

Discussion Other current considerations are similar to that in image-image translation including building sketch-photo translation models that work based on unpaired samples.

Discussion Sketch inpainting is an interesting task related to sketch generation, with a wide range of practical applications in engineering [287], [288], medicine [288], agriculture [288]–[290], *etc*. This task can be considered as a single-modal generation task from deteriorated sketch (with discontinuities strokes) to a restored complete sketch. *e.g.*, old sketches [33] and engineering sketches [287]. Sketch inpainting is also studied for generating/parsing the sketch structures from non-sketch images (*e.g.*, retina [288], plant roots [288]–[290], road networks [288] from satellites), which is a cross-modal photo-to-sketch generation task. How to solve this problem in end-to-end framework is still challenging and under-studied.

4.2.4 *Sketch-3D Retrieval*

Sketch-3D retrieval refers to using sketch as query to retrieve 3D models [291]. Compared to SBIR, 3D retrieval is more challenging due to the larger domain gap between 2D sketch and 3D model.

Sketch-3D retrieval was studied well before the deep learning era [47]. Classic approaches often proceeded in a two-stage manner [292]: (i) View selection: Use an automatic procedure to select representative viewpoints of a given 3D model, hoping that one of the selected viewpoints is similar to that of the query sketches; and (ii) Projection and Matching: Project each selected view of the 3D model into 2D space by line rendering algorithm [293]. Then match the sketch against the 2D projections of the model based on predefined features such as SIFT [120]. See Figure 16 for an illustration. However, as argued in some previous work [2], view selection is a bottleneck of the two-stage approach as the "best" views are subjective and ambiguous. Moreover, matching based on hand-crafted features is inaccurate.

Gradually, sketch based 3D model retrieval has been studied within the end-to-end deep learning paradigm [2], [294]–[299]. As a representative method, Wang *et al*. [2] proposed to use two Siamese networks to learn the sketch and projected views directly in the end-to-end manner, which takes a quadruplet of sketches and projected viewpoints as input, and uses multiple pair-wise losses. In each

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3148853, IEEE Transactions on Pattern Analysis and Machine Intelligence

Fig. 16. Illustration of view matching across sketch and 3D shape. Images (from left to right: sketch, 3D shape, three random views of the 3D shape) are selected from [294].

quadruplet, two sketches ($\mathbf{X}_1$, $\mathbf{X}_2$) and two viewpoints ($\mathbf{V}_1$, $\mathbf{V}_2$) are randomly selected from sketch and 3D domains, respectively. For simplicity they assume that $\mathbf{X}_1$ and $\mathbf{X}_2$ are from the same category sharing a Siamese network, while $\mathbf{V}_1$ and $\mathbf{V}_2$ are also from the same category sharing another Siamese network. The quadruplet loss is

$$\begin{aligned}\mathcal{L}(\mathbf{X}_1, \mathbf{X}_2, \mathbf{V}_1, \mathbf{V}_2) = \mathcal{L}_{pair}(\mathbf{X}_1, \mathbf{X}_2) \\ + \mathcal{L}_{pair}(\mathbf{V}_1, \mathbf{V}_2) + \mathcal{L}_{pair}(\mathbf{X}_1, \mathbf{V}_1),\end{aligned} \tag{4}$$

The loss function terms $\mathcal{L}_{pair}(\mathbf{X}_1, \mathbf{X}_2)$ and $\mathcal{L}_{pair}(\mathbf{V}_1, \mathbf{V}_2)$ enable the network to learn category-level similarity within each domain; while the $\mathcal{L}_{pair}(\mathbf{X}_1, \mathbf{V}_1)$ term forces the network to learn cross-modal similarity. Given input samples a and b, the pair-wise loss function is defined as:

$$\mathcal{L}_{pair}(a, b) = \begin{cases} \alpha\mathcal{D}(\mathcal{F}_a(a), \mathcal{F}_b(b)), & \text{if } y_a \neq y_b, \\ \beta e^{\gamma\mathcal{D}(\mathcal{F}_a(a), \mathcal{F}_b(b))}, & \text{otherwise}, \end{cases} \tag{5}$$

where y_a and y_b are the corresponding class labels, and $\mathcal{F}_a(\cdot)$ and $\mathcal{F}_b(\cdot)$ denote the feature extractions that have been applied to a and b, respectively.

Besides this pair-wise deep metric learning, other deep metric learning methods also can be applied to Sketch-3D matching, *e.g.*, triplet ranking [300], [301] and deep correlation metric learning [302], [303].

Moreover, some previous works also studied how to represent 3D models more comprehensively in sketch based 3D retrieval tasks. For instance, Xie *et al.* [304] proposed to represent 3D models by computing the Wasserstein distance [305] based barycenters of multiple projections of 3D models.

4.2.5 Sketch-3D Generation

Sketch to 3D model generation is also an interesting cross-modal research topic analogous to the sketch-to-image generation discussed in Section 4.2.3. Using sketch to generate 3D models/shapes [306], [307] is extremely challenging but has important applications such as sketch-based product design [308], [309]. Compared to the other tasks discussed, this is relatively under-studied thus far. Most of the existing deep learning based sketch-to-3D generation models are engineered for highly well-drawn or professional pencil sketches [310], [311]. Recently, 3D-to-sketch [312] generation has also been explored in a deep learning manner.

DeepSketchHair [308] is a representative deep model that generates realistic 3D hairstyle models from 2D sketches. As shown in Figure 17, given a 3D bust model as a reference, the system takes in a user-drawn sketch (consisting of hair contour (red lines) and a few strokes indicating the hair growing directions (blue lines) within a hair region), and automatically generates a 3D hair model, which matches the input sketch both globally (for contour) and locally

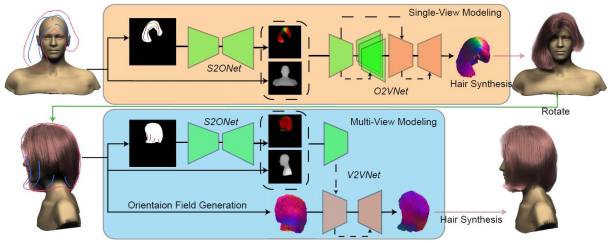


Fig. 17. Pipeline of a sketch-based 3D hair modeling deep model [308]. This system takes 2D sketch as input and generates a realistic 3D hairstyle.

(for growing directions). This model solves two challenging cross-domain mappings: (i) mapping sketch to dense 2D hair orientation field, by S2ONet, and (ii) mapping 2D orientation field to 3D vector field, by O2VNet.

4.2.6 Sketch-Video Retrieval

Sketch based video retrieval (SBVR) has been studied [313] prior to contemporary deep learning. SBVR is also highly challenging due to the huge domain gap between free-hand sketch and video. In SBVR applications using sketch as query has the advantage that humans can use lines or arrow vectors to describe moving or other dynamic scenes. Thus, sketch can be used to not only depict static objects and scenes, but also motion information.

Performing SBVR using sketches conveying both appearance and motion is challenging due to the need to segment the motion and appearance information from the sketch and use it to address the corresponding channels in the video gallery for retrieval. To address this challenge, a multi-stream multi-modality deep network was proposed in [314]. This study further extended SBVR to fine-grained SBVR to perform instance-level retrieval of videos given sketch queries.

Finally, we note that, motion sketch based crowd video retrieval [315], [316] has been studied recently, which is useful for video surveillance analysis.

Discussion Which kinds of videos are appropriate and feasible for sketch-based retrieval is still an open question. As are human-computer-interaction questions around how time, shot change, and motions should be depicted in sketch.

4.2.7 Other Sketch-Related Multi-Modal Tasks

Recently some other interesting sketch based multi-modal tasks have emerged, *e.g.*, text-to-sketch generation [317] (*e.g.*, text instruction based conversational authoring of sketches [318]), sketch-based photo classifier generation [20], sketch-based segmentation model generation [319], sketch-based pictionary games [54], [131], [320], sketch to photo contour transfer [321], and sketch-guided object localization in natural photos [322].

4.3 Experimental Comparison

4.3.1 Representing Sketch

As discussed in Section 2.1, sketch can be represented in several diverse formats such as raster image, and way-point sequence. These representations lend themselves to

different neural network architectures. We therefore take the opportunity to use `TorchSketch` to perform the first thorough comparison of neural network architectures for sketch recognition/representing.

To this end we follow [23] in using 414K sketches drawn from QuickDraw. These are organized into training, validation, and test splits composed of 1000, 100, and 100 sketches respectively from each of the 345 QuickDraw categories. Following [19], we truncate or pad sketch samples to a uniform length of 100 key points/steps to facilitate efficient training of RNN-, GNN-, and TCN-based models, where each time-step is a $4D$ input (*i.e.*, two coordinates and two pen state bits). Sketch recognition is a fundamental topic within the field, and the QuickDraw sketches are subject to realistic abstraction, noise, and drawing diversity. We therefore hope that this benchmark can help practitioners while supporting future research in the field.

The recognition accuracy across these architectures, as implemented by `TorchSketch`, are reported in Table 6. We can analyze these results with comparisons within and across architecture categories. Within each architecture, we can observe from Table 6: (i) For CNNs, the deeper networks (*e.g.*, DenseNet-161) have no obvious advantage compared to the shallower networks (*e.g.*, AlexNet, VGG). This is likely due to the sparsity of sketch where redundant convolution and pooling operations lose information about the sparse pixels. (ii) For RNNs, bidirectional networks outperform the unidirectional networks by a clear margin (0.66+ vs. 0.60+). This makes sense as human sketch ordering is only loosely consistent [133]. (iii) For GNNs, multi-graph transformer (MGT) outperforms graph convolutional network (GCN) and graph attention network (GAT).

Across the architectures, we can observe: (i) Thus far the best CNN networks (InceptionV3) outperform the best sequential networks. This may be because the sequential networks (RNN, GNN, and TCN) truncate the input coordinate sequences. (ii) Among GNNs, the multi-graph transformer [23] comes closest to matching peak CNN performance. (iii) Compared with CNNs and GNNs, RNNs and TCN have significantly fewer parameters. However (iv) TCN, performs unsatisfactorily in this fully-supervised setting.

4.3.2 Robustness Study on Spatial Transformation

As discussed in Section 2.1, even if sketches are shifted, rescaled, rotated, or flipped, they still can be recognized easily by people. It will be interesting to evaluate how sensitive the current deep networks are to the spatial transformations on different sketch tasks and datasets.

To facilitate comparison, we choose three quantitatively comparable tasks as target tasks involving both single-modal and multi-modal settings, *i.e.*, sketch recognition, coarse-grained SBIR, fine-grained SBIR, which will be respectively conducted on three commonly used datasets, *i.e.*, TU-Berlin, TU-Berlin Extended, QMUL Shoe. To make it more intuitive, we choose a CNN backbone, *i.e.*, ResNet-18.

Cross-entropy loss and triplet loss are used for recognition and SBIR tasks, respectively. To perform early-stopping and select models based on validation performance, we need to split the datasets: (i) For TU-Berlin, 40, 20, and 20 sketches per category are randomly selected for training, validation, and testing, respectively. (ii) For TU-Berlin Extended, both sketches and photos are randomly divided into training, validation, and testing sets as a ratio of $2:1:1$. (iii) For QMUL Shoe, 50 sketch-photo pairs are selected randomly from its training set for validation. To fully verify the sensitivity, we did not adopt any data augmentations in the training stage.

We choose five representative randomly spatial transformations for robustness testing, including position shift, scale, horizontal flip ($p = 0.5$), vertical flip ($p = 0.5$), and center rotation ($-45°$ to $45°$). Considering the randomness, testing was repeated 10 times with each spatial transformation, and both mean ($\%$) and standard deviation for each indicator are reported in Table 7, where top-K accuracy and rank accuracy are used as metrics for recognition and retrieval. Moreover, mean average precision (mAP) is also reported to evaluate the performance for coarse-grained SBIR as multiple true matches are provided by gallery. For a clear comparison, the testing results without any spatial transformations are provided in the bottom row of Table 7, while we draw boxplots (Figure 18) based on "acc.@1", "mAP", and "rank@1" for the selected tasks, respectively.

In Table 7, we observed that: (i) Deep sketch models are vulnerable to the spatial transformations that can result in performance degradation. (ii) Compared with shift and horizontal flip, the three other transformations cause more noticeable performance changes. In particular, for TU-Berlin sketches, horizontal flip causes very small performance changes. (iii) On QMUL Shoe, spatial transformation based perturbations have relatively large standard deviations. Moreover, it is interesting to see that shift transformation slightly improves fine-grained SBIR accuracy. This is also demonstrated by Figure 18. This is likely due to the small sample size (only 115 sketch-photo pairs for testing).

These observations indicate that sketch based spatial transformations are able to attack deep networks. The current deep learning technique still needs to be improved to achieve sketch-oriented robustness.

5 Discussion

5.1 Open Problems

5.1.1 Deep Learning vs. Traditional Methods

In recent years, deep learning methods have achieved the state-of-the-art in all the sketch tasks. However, some open problems still need further study, *e.g.*, (i) What are the advantages and disadvantages of deep learning and traditional methods on sketch? (ii) Why do deep learning networks work well on sketch? (iii) How are sketch-unique characteristics modelled by various deep learning networks?

Generally traditional methods work in two stages, *i.e.*, feature engineering from sketch space to feature space, mapping from feature space to target space, while generally deep learning methods do end-to-end mapping directly from sketch space to target space. Thus the essential opportunity for exploiting human insight is in hand-designing network structures vs. hand-designing features. Due to the sketch domain challenges (*e.g.*, abstract, noisy, sparse, diverse), the difficulty of feature engineering for sketch is one of the main bottlenecks of the traditional methods. This is

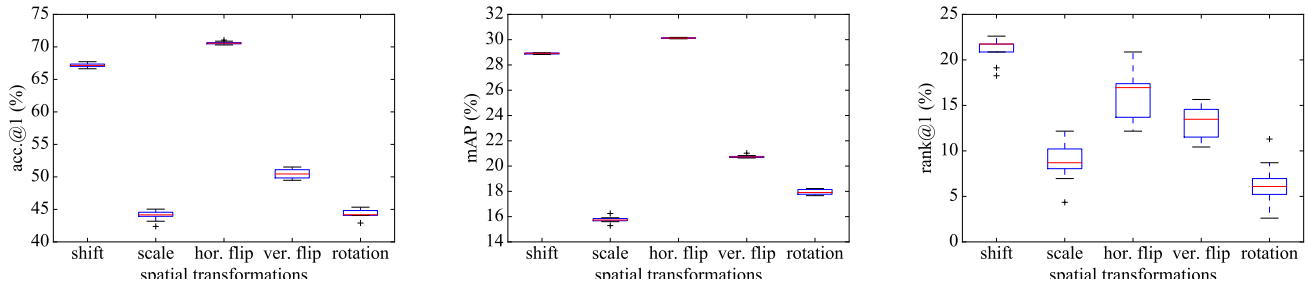
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3148853, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 6
Comparison of recognition accuracy for different network architectures on a subset [23] of QuickDraw [5].

Architecture & Network		Input	Recognition Accuracy			Parameter Amount
			acc.@1	acc.@5	acc.@10	
Convolutional Neural Networks (CNNs)	AlexNet [105]	picture	0.6808	0.8847	0.9203	58,417,305
	VGG-11 [179]		0.6743	0.8814	0.9191	130,179,801
	VGG-13 [179]		0.6808	0.8881	0.9232	130,364,313
	VGG-16 [179]		0.6837	0.8889	0.9253	135,674,009
	VGG-19 [179]		0.6908	0.8839	0.9208	140,983,705
	Inception V3 [323]		0.7422	0.9189	0.9437	25,315,474
	ResNet-18 [108]		0.7164	0.9072	0.9381	11,353,497
	ResNet-34 [108]		0.7154	0.9083	0.9375	21,461,657
	ResNet-50 [108]		0.7043	0.8987	0.9303	24,214,937
	ResNet-101 [108]		0.7071	0.8992	0.9317	43,207,065
	ResNet-152 [108]		0.6924	0.8973	0.9312	58,850,713
	DenseNet-161 [324]		0.7008	0.8971	0.9302	27,234,105
	DenseNet-169 [324]		0.7173	0.9050	0.9358	13,058,905
	DenseNet-201 [324]		0.7050	0.9013	0.9331	18,755,673
	MobileNet V2 [325]		0.7310	0.9161	0.9429	2,665,817
Recurrent Neural Networks (RNNs)	LSTM	stroke vector	0.6068	0.8416	0.8931	2,593,881
	Bi-directional LSTM		0.6665	0.8820	0.9189	5,553,241
	GRU		0.6224	0.8574	0.9055	2,000,473
	Bi-directional GRU		0.6768	0.8854	0.9234	5,419,097
Graph Neural Networks (GNNs)	Graph Convolutional Network (GCN) [326]	stroke vector	0.6800	0.8869	0.9224	6,948,441
	Graph Attention Network (GAT) [327]		0.6977	0.8952	0.9298	11,660,889
	Vanilla Transformer [328]		0.5249	0.7802	0.8486	14,029,401
	Multi-Graph Transformer (Base) [23]		0.7070	0.9030	0.9351	10,096,601
	Multi-Graph Transformer (Large) [23]		0.7280	0.9106	0.9387	39,984,729
Textual Convolutional Network (TCN)	TCN [111]	stroke vector	0.5511	0.8020	0.8646	2,750,873

TABLE 7
Robustness of ResNet-18 CNN backbone to perturbations in the form of random spatial transformations.
Each setting is tested 10 times, and mean (%) and standard deviation of performance are reported.

Spatial Transform	Recognition on TU-Berlin			Coarse-Grained SBIR on TU-Berlin Extended				Fine-Grained SBIR on QMUL Shoe		
	acc.@1	acc.@5	acc.@10	rank@1	rank@5	rank@10	mAP	rank@1	rank@5	rank@10
shift	67.16±0.31	88.20±0.24	92.82±0.32	51.17±0.36	71.04±0.31	77.33±0.21	28.91±0.05	21.13±1.42	50.17±2.42	67.39±2.10
scale	44.07±0.80	68.49±0.62	77.08±0.48	31.24±0.33	51.02±0.56	59.06±0.50	15.75±0.24	8.78±2.15	25.74±2.10	39.65±3.02
horizontal flip	70.60±0.22	89.58±0.12	93.88±0.14	52.30±0.13	72.05±0.21	78.32±0.20	30.12±0.04	16.08±2.85	44.00±2.63	58.61±2.70
vertical flip	50.44±0.74	70.77±0.59	78.00±0.36	36.25±0.33	52.57±0.42	58.32±0.32	20.75±0.11	13.04±1.88	40.35±3.18	56.17±3.36
rotation	44.34±0.69	68.38±0.44	76.83±0.50	32.84±0.66	50.14±0.48	57.14±0.73	17.94±0.21	6.35±2.39	17.48±2.64	27.92±3.19
None	72.06	90.1	94.02	53.12	73.14	78.94	30.62	20.00	53.91	68.70



(a) Recognition on TU-Berlin (b) SBIR on TU-Berlin Extended (c) Fine-Grained SBIR on QMUL Shoe

Fig. 18. Boxplots for robustness evaluation on spatial transformations.

somewhat ameliorated by deep learning methods which have the capacity to learn strong feature representations given sufficient data to model their variability and diversity. Furthermore, deep learning networks have various motivations and mechanisms to handle the sketch-unique characteristics: (i) CNNs use convolution filters to imitate the human reception field, and treat sketch as a binary matrix in pixel space. (ii) RNNs imitate the temporally-extended human sketching process to recurrently capture spatial and temporal patterns of stroke, (iii) GNNs represent sketches as graphs, and can encode both topological/spatial and temporal patterns of stroke sequences, and (iv) TCNs imitate a temporal reception field on stroke sequence.

We summarize the advantages and disadvantages of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2022.3148853, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 8
Comparison of deep learning and traditional approaches to sketch-related tasks.

Method	Advantages	Disadvantages
Deep Learning	relatively superior performance end-to-end mapping larger model capacity to support big data more training data brings higher performance in general	manually-designed network structures more parameters/resource cost vulnerable to overly fit less rigorous in mathematical expression and interpretability
Traditional	clear in mathematical expression and interpretability fewer parameters/resource less over-fitting	relatively suboptimal performance manually-designed features under-parameterized for big data

TABLE 9
Comparison of different deep network architectures for sketch-oriented tasks.

Arc.	Input	Model Space	Motivations	Advantages	Disadvantages
CNNs	picture (full sketch)	spatial (Euclidean)	imitate human reception field regard sketch as binary pixel matrix	good performance in global-level tasks perceive full sketch (no information lost) can stack deeper layers allow a variety of perception granularities	weak in stroke-level tasks fail in temporal tasks relatively more parameters vulnerable to overly fit
RNNs	stroke vector (key points)	spatial&temporal	imitate human sketching process	recurrently capture stroke temporal patterns relatively fewer parameters	weak in stacking deeper layers weak in long stroke sequences perception granularity is fixed
GNNs + Transformers	stroke vector (key points)	spatial&temporal	represent sketch as graph (key point $\rightarrow$ node, stroke $\rightarrow$ node, *etc*)	higher flexibility in network design relatively fewer parameters can handle long stroke sequences	weak in stacking deeper layers abstract network structure
TCNs	stroke vector (key points)	spatial&temporal	imitate temporal reception field (key point $\rightarrow$ word, stroke $\rightarrow$ sentence)	concise network architecture relatively fewer parameters allow a variety of perception granularities	weak in stacking deeper layers fail to recurrently perceive strokes

deep learning and traditional methods on sketch in Table 8; and we compare different network architectures for sketch oriented tasks in Table 9.

5.1.2 Data and Annotations

Annotation Uni-modal sketch datasets have begun to provide some annotation such as grouping [152] and segmentation [25]. However more fine-grained annotation of this type is necessary, wit sketch attributes in particular being lacking. As discussed in Section 4.2.3, existing multi-modal sketch benchmark annotation is primarily in terms of *pairings* (*e.g.*, sketch-photo, sketch-3D). However, fine-grained/local annotations (such as stroke-contour, parts, and attributes) would enable richer cross-modal alignment models to be learned.

Meta-Data and Fairness Unlike photos, sketches are uniquely influenced by the demographics, perception, memory, and drawing style of the artist; as well as the conditions under which they are drawn (*i.e.*, time-limited or not). Most existing datasets do not take care to acquire balanced samples of users across background, age, gender, *etc*; or record such meta-data about their participants. However, such sampling and meta-data are necessary for studying changes in drawing style with these covariates, as well as for ensuring that sketch-based applications work well for users of different backgrounds. Furthermore, existing sketch datasets are mainly created as bespoke efforts by researchers or casual participants in online games with time-pressure. These may lead to sketches that are either excessively well dawn, or too poorly drawn. Data collected under a variety of drawing conditions, and meta-data about those conditions, would help sketch research in future.

5.1.3 Architectures and Sketch-Specific Design

Architectures As discussed in Section 4.3 and Table 6, there are a variety of network architectures that can be applied to sketch, and these lead to a range of performances. These results suggest that performance will continue to advance as better architectures within each family are developed. The best architecture for sketch perception is still an open question.

Sketch-Specific Design Another open question is to what extent sketch-specific designs are important vs. generic computer vision architectures and learning algorithms. Clearly sketch has unique challenges (sequential time-series nature, sparsity, abstraction, artist style, *etc*) that can be better taken into account with sketch-specific designs. However the broader vision and learning community can bring greater effort to bear on developing more advanced general purpose models. Therefore it remains to be seen in which sketch applications sketch-specific designs can take a decisive lead over generic architectures and algorithms. For example, sketch-specific designs may be more important in fine-grained tasks such as segmentation, grouping, and FG-SBIR compared to the simplest coarse-grained object categorization task.

Association with Other Sketches Some CNN based models designed for other kinds of sketches (*e.g.*, well-drawn line drawings [16], [177], [287], [329], cartoons [16], [177]) can be applied to free-hand sketch. In particular, it has been verified by relevant literature that CNN models oriented at high-quality line drawings can be successful for free-hand sketch tasks, including vectorization [155], [164], [330], [331], shading [332], inking [333], *etc*. It is interesting to evaluate which methods designed for other sketches can or cannot

be applied to free-hand sketches. This will depend on to what extent architectures designed for well-drawn sketches (Figure 2) become over-specialized to that type of data, or can generalize to the more abstract, diverse, and sparse free-hand sketches (Figure 1).

5.2 Potential Research Directions

In this section, we outline some potential research directions that we believe are promising in future, from both the perspectives of potential application and underpinning research value.

5.2.1 Potential Application-Oriented Research

Scene Sketches Scene-level sketch oriented deep learning is still under-studied. Recently, several seminal works (*e.g.*, SketchyCOCO [17], SceneSketcher [78]) have opened up promising directions for scene-level sketch research. They not only contribute large-scale scene-level sketch datasets but also propose novel research topics, *e.g.*, scene sketch based image generation [17], fine-grained scene sketch based image retrieval [78]. These novel topics are useful for the practical applications, *e.g.*, sketch based scene design. It remains to be seen up to complexity of scenes are technically feasible - and practically appealing for users - to retrieve using sketches.

3D Sketches Collecting 3D sketches [334] is now easier thanks to new data collection equipment. This could support many interesting 3D sketch related research topics, *e.g.*, combining virtual reality (VR) [335] and augmented reality (AR) [10], [11], [336]. 3D sketch research will help to bring sketch-based human-computer interaction from the 2D plane of the touch-screens to 3D spaces, enabling more immersive experience.

Diverse Sketch Subjects Existing sketch datasets and applications mainly focus on sketches depicting objects. However, in practical applications users may be interested in machines understanding more diverse sketched concepts, *e.g.*, sheets, curves, histograms [337], maps [338], engineering sketches [309], and user interface (UI) prototype drawings [339]. Sketch also can be studied together with handwritten characters.

Sketch Color and Pressure Existing free-hand sketches are collected by common touch-screen devices, *e.g.*, phone or tablet. Here the position of strokes is the main feature, with color and texture not being widely collected. Thus, existing sketch analysis has mostly focused on black or grayscale sketches. However, sketches can already be colored, and recent devices can increasingly sense pressure along strokes. Upgrading models to exploit color, texture, and pressure properties of sketches remains as outstanding work.

Sketch Beautification How to beautify sketch [340] in various sketch-related HCI applications is interesting and challenging. Sketch beautification will not only refine user experience but also improve interaction efficiency in the sketch-based design applications, *e.g.*, modifying a non-professional sketch to a professional sketch [128].

Sketch-Based Design People often use very simple or even scrawled sketches at the beginning of a design to brainstorm and generate inspiration. Thus, there are some very useful sketch-based designing applications that can help people to design 2D or 3D products, *e.g.*, sketch-to-comic [341]–[344], sketch-to-shadow [34], [332], and sketch-to-normal [345], [346]. We believe that in future these techniques can be continually improved further by future deep learning techniques. These sketch-based designing methods can improve the efficiency in HCI.

Efficient Models For sketch-based perception and user-interfaces in mobile devices such as phones, tablets, or AR/VR gear, processing should be real-time and light-weight enough to run on embedded battery powered devices. How to compress sketch models and improve their efficiency is an important question for future work.

5.2.2 Potential Theoretical Research

Sketch-oriented deep learning models have achieved good performance within well curated datasets. However, how well they can generalise to uncontrolled real-world conditions remains to be seen.

Diversity, Style, and Robustness The impact of dataset shift [347] has only begun to be studied in sketches. More uniquely, sketch is particularly subjective in terms of influence by the user's drawing style, culture and potentially demographics. Robustness to these stylistic aspects is an under-studied area of sketch research. Similarly, taking the native format of sketch, existing adversarial attack studies could be extended to sketch domain by studying adversarial *strokes* or *waypoints*, rather than pixel perturbations.

Sketch as a Robustness Test Sketch images differ dramatically from photos, yet are easily recognized by humans. Sketch images such as ImageNet-Sketch [41], SketchTransfer [66], and PACS [73] can thus be used to benchmark the robustness and generalization ability of more general image-oriented deep learning models. Going beyond recognition, similar robustness evaluations could be performed on other instance-level retrieval problems such as person Re-ID.

Data-Efficient Sketch Models Major process in deep learning for sketch research has been driven by increasingly large sketch datasets (Table 2). However, ultimately these datasets are harder to scale than corresponding photo datasets due to the need for manual sketching. Therefore data-efficient approaches to all the main sketch-analysis tasks of interest from recognition to SBIR are of major importance going forward. Whether this is best achieved by few-shot learning, self-supervised learning, or cross-modality knowledge transfer from photo domain remains to be seen.

6 Conclusion

This survey reviewed the landscape of contemporary deep-learning based sketch research. We introduced the unique aspects of sketch in terms of sketch-specific challenges and diverse potential representations, and analyzed both existing datasets and existing methods in terms of a rich ecosystem of uni-modal and multi-modal sketch analysis tasks. We discussed open problems and under-studied research directions throughout. We hope this survey will help new researchers and practitioners get up to speed, provide a convenient reference for sketch experts, and encourage future progress in this exciting field.

REFERENCES

[1] R. Hu and J. Collomosse, "A performance evaluation of gradient field hog descriptor for sketch based image retrieval," *CVIU*, 2013.

[2] F. Wang, L. Kang, and Y. Li, "Sketch-based 3d shape retrieval using convolutional neural networks," in *CVPR*, 2015.

[3] Q. Yu, Y. Yang, F. Liu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Sketch-a-net: A deep neural network that beats humans," *IJCV*, 2017.

[4] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C.-C. Loy, "Sketch me that shoe," in *CVPR*, 2016.

[5] D. Ha and D. Eck, "A neural representation of sketch drawings," in *ICLR*, 2018.

[6] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *TOG*, 2012.

[7] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: learning to retrieve badly drawn bunnies," *TOG*, 2016.

[8] F. Huang, J. F. Canny, and J. Nichols, "Swire: Sketch-based user interface retrieval," in *CHI*, 2019.

[9] S. Suleri, V. P. Sermuga Pandian, S. Shishkovets, and M. Jarke, "Eve: A sketch-based software prototyping workbench," in *CHI*, 2019.

[10] K. C. Kwan and H. Fu, "Mobi3dsketch: 3d sketching in mobile ar," in *CHI*, 2019.

[11] D. Gasques, J. G. Johnson, T. Sharkey, and N. Weibel, "What you sketch is what you get: Quick and easy augmented reality prototyping with pintar," in *CHI*, 2019.

[12] A. Kotani and S. Tellex, "Teaching robots to draw," in *ICRA*, 2019.

[13] J. E. Fan, M. Dinculescu, and D. Ha, "collabdraw: an environment for collaborative sketching with an artificial agent," in *ACM SIGCHI Conference on Creativity and Cognition*, 2019.

[14] I. E. Sutherland, "Sketchpad a man-machine graphical communication system," *Simulation*, 1964.

[15] C. F. Herot, "Graphical input through machine recognition of sketches," *TOG*, 1976.

[16] E. Simo-Serra, S. Iizuka, K. Sasaki, and H. Ishikawa, "Learning to simplify: fully convolutional networks for rough sketch cleanup," *TOG*, 2016.

[17] C. Gao, Q. Liu, Q. Xu, L. Wang, J. Liu, and C. Zou, "Sketchycoco: Image generation from freehand scene sketches," in *CVPR*, 2020.

[18] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Sketch-a-net that beats humans," in *BMVC*, 2015.

[19] P. Xu, Y. Huang, T. Yuan, K. Pang, Y.-Z. Song, T. Xiang, T. M. Hospedales, Z. Ma, and J. Guo, "Sketchmate: Deep hashing for million-scale human sketch retrieval," in *CVPR*, 2018.

[20] C. Hu, D. Li, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Sketch-a-classifier: Sketch-based photo classifier generation," in *CVPR*, 2018.

[21] U. Riaz Muhammad, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Learning deep sketch abstraction," in *CVPR*, 2018.

[22] T. Portenier, Q. Hu, A. Szabo, S. A. Bigdeli, P. Favaro, and M. Zwicker, "Faceshop: Deep sketch-based face image editing," *TOG*, 2018.

[23] P. Xu, C. K. Joshi, and X. Bresson, "Multigraph transformer for free-hand sketch recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[24] L. Yang, J. Zhuang, H. Fu, K. Zhou, and Y. Zheng, "Sketchgcn: Semantic sketch segmentation with graph convolutional networks," *arXiv preprint arXiv:2003.00678*, 2020.

[25] Y. Qi and Z.-H. Tan, "Sketchsegnet+: An end-to-end learning of rnn for multi-class sketch semantic segmentation," *Access*, 2019.

[26] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *ICCV*, 2017.

[27] S. Ouyang, T. M. Hospedales, Y.-Z. Song, and X. Li, "Forgetmenot: Memory-aware forensic facial sketch matching," in *CVPR*, 2016.

[28] C. Hu, D. Li, Y.-Z. Song, and T. M. Hospedales, "Now you see me: Deep face hallucination for unviewed sketches." in *BMVC*, 2017.

[29] S. Nagpal, M. Singh, R. Singh, M. Vatsa, A. Noore, and A. Majumdar, "Face sketch matching via coupled deep transform learning," in *ICCV*, 2017.

[30] D.-P. Fan, S. Zhang, Y.-H. Wu, Y. Liu, M.-M. Cheng, B. Ren, P. L. Rosin, and R. Ji, "Scoot: A perceptual metric for facial sketches," in *ICCV*, 2019.

[31] L. Pang, Y. Wang, Y.-Z. Song, T. Huang, and Y. Tian, "Cross-domain adversarial feature learning for sketch re-identification," in *MM*, 2018.

[32] M. Huang, J. Lin, N. Chen, W. An, and W. Zhu, "Reversed sketch: A scalable and comparable shape representation," *PR*, 2018.

[33] S. Kazuma, S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Learning to Restore Deteriorated Line Drawing," *The Visual Computer*, 2018.

[34] Q. Zheng, Z. Li, and A. Bargteil, "Learning to shadow hand-drawn sketches," in *CVPR*, 2020.

[35] J. Lee, E. Kim, Y. Lee, D. Kim, J. Chang, and J. Choo, "Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence," in *CVPR*, 2020.

[36] S. Gui, Y. Zhu, X. Qin, and X. Ling, "Learning multi-level domain invariant features for sketch re-identification," *Neurocomputing*, 2020.

[37] C. Yan, D. Vanderhaeghe, and Y. Gingold, "A benchmark for rough sketch cleanup," *TOG*, 2020.

[38] L. Zhang, C. Li, E. Simo-Serra, Y. Ji, T.-T. Wong, and C. Liu, "User-Guided Line Art Flat Filling with Split Filling Mechanism," in *CVPR*, 2021.

[39] S.-B. Chen, P.-C. Wang, B. Luo, C. H. Ding, and J. Zhang, "Sragan: Generating colour landscape photograph from sketch," in *IJCNN*, 2019.

[40] M. R. Amer, S. Yousefi, R. Raich, and S. Todorovic, "Monocular extraction of 2.1 d sketch using constrained convex optimization," *IJCV*, 2015.

[41] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," in *NeurIPS*, 2019.

[42] K. Chen, I. Rabkina, M. D. McLure, and K. D. Forbus, "Human-like sketch object recognition via analogical learning," in *AAAI*, 2019.

[43] X. Han, K. Hou, D. Du, Y. Qiu, Y. Yu, K. Zhou, and S. Cui, "Caricatureshop: Personalized and photorealistic caricature sketching," *arXiv preprint arXiv:1807.09064*, 2018.

[44] L. Zhao, F. Han, X. Peng, X. Zhang, M. Kapadia, V. Pavlovic, and D. N. Metaxas, "Cartoonish sketch-based face editing in videos using identity deformation transfer," *Computers Graphics*, 2019.

[45] M. Yuan and E. Simo-Serra, "Line Art Colorization with Concatenated Spatial Attention," in *CVPR Workshops*, 2021.

[46] J. Delanoy, M. Aubry, P. Isola, A. A. Efros, and A. Bousseau, "3d sketching using multi-view deep volumetric prediction," *CGIT*, 2018.

[47] B. Li, Y. Lu, A. Godil, T. Schreck, B. Bustos, A. Ferreira, T. Furuya, M. J. Fonseca, H. Johan, T. Matsuda *et al.*, "A comparison of methods for sketch-based 3d shape retrieval," *CVIU*, 2014.

[48] N. Prajapati and G. Prajapti, "Sketch based image retrieval system for the web-a survey," *IJCSIT*, 2015.

[49] M. Indu and K. Kavitha, "Survey on sketch based image retrieval methods," in *ICCPCT*, 2016.

[50] Y. Li and W. Li, "A survey of sketch-based image retrieval," *Machine Vision and Applications*, 2018.

[51] X. Zhang, X. Li, Y. Liu, and F. Feng, "A survey on freehand sketch recognition and retrieval," *Image and Vision Computing*, 2019.

[52] M. Schrapel, F. Herzog, S. Ryll, and M. Rohs, "Watch my painting: The back of the hand as a drawing space for smartwatches," *CHI*, 2020.

[53] M. Dvorožňák, D. Sýkora, C. Curtis, B. Curless, O. Sorkine-Hornung, and D. Salesin, "Monster Mash: A single-view approach to casual 3d modeling and animation," *ACM Transactions on Graphics*, 2020.

[54] R. K. Sarvadevabhatla, S. Surya, T. Mittal, and V. B. Radhakrishnan, "Pictionary-style word-guessing on hand-drawn object sketches: dataset, analysis and deep network models," *TPAMI*, 2020.

[55] J. Song, K. Pang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Learning to sketch with shortcut cycle consistency," in *CVPR*, 2018.

[56] A. Das, Y. Yang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Cloud2curve: Generation and vectorization of parametric sketches," in *CVPR*, 2021.

[57] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[58] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *CVPR*, 2017.

[59] K. Li, K. Pang, J. Song, Y.-Z. Song, T. Xiang, T. M. Hospedales, and H. Zhang, "Universal sketch perceptual grouping," in *ECCV*, 2018.

[60] J. Choi, H. Cho, J. Song, and S. M. Yoon, "Sketchhelper: Real-time stroke guidance for freehand sketch retrieval," *TMM*, 2019.

[61] F. Wang, S. Lin, H. Wu, H. Li, R. Wang, X. Luo, and X. He, "Sp-fusionnet: Sketch segmentation using multi-modal data fusion," in *ICME*, 2019.

[62] R. K. Sarvadevabhatla, S. Suresh, and R. V. Babu, "Object category understanding via eye fixations on freehand sketches," *TIP*, 2017.

[63] "Aaron koblin sheep dataset," https://github.com/hardmaru/sketch-rnn-datasets/tree/master/aaron_sheep.

[64] C. Tirkaz, B. Yanikoglu, and T. M. Sezgin, "Sketched symbol recognition with auto-completion," *PR*, 2012.

[65] S. Dey, P. Riba, A. Dutta, J. Llados, and Y.-Z. Song, "Doodle to search: Practical zero-shot sketch-based image retrieval," in *CVPR*, 2019.

[66] A. Lamb, S. Ozair, V. Verma, and D. Ha, "Sketchtransfer: A new dataset for exploring detail-invariance and the abstractions learned by deep networks," in *WACV*, 2020.

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

[68] T. Jiang, G.-S. Xia, and Q. Lu, "Sketch-based aerial image retrieval," in *ICIP*, 2017.

[69] T.-B. Jiang, G.-S. Xia, Q.-K. Lu, and W.-M. Shen, "Retrieving aerial scene images with learned deep image-sketch features," *JCST*, 2017.

[70] X. Wang, X. Duan, and X. Bai, "Deep sketch feature for cross-domain image retrieval," *Neurocomputing*, 2016.

[71] M. Eitz, R. Richter, T. Boubekeur, K. Hildebrand, and M. Alexa, "Sketch-based shape retrieval," *TOG*, 2012.

[72] B. Li, Y. Lu, C. Li, A. Godil, T. Schreck, M. Aono, M. Burtscher, H. Fu, T. Furuya, H. Johan *et al.*, "Shrec'14 track: Extended large scale sketch-based 3d shape retrieval," in *Eurographics workshop on 3D object retrieval*, 2014.

[73] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *ICCV*, 2017.

[74] C. Xiao, C. Wang, L. Zhang, and L. Zhang, "Sketch-based image retrieval via shape words," in *ICMR*, 2015.

[75] L. Castrejon, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba, "Learning aligned cross-modal representations from weakly aligned data," in *CVPR*, 2016.

[76] C. Zou, Q. Yu, R. Du, H. Mo, Y.-Z. Song, T. Xiang, C. Gao, B. Chen, and H. Zhang, "Sketchyscene: Richly-annotated scene sketches," in *ECCV*, 2018.

[77] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, "Moment matching for multi-source domain adaptation," in *ICCV*, 2019.

[78] F. Liu, C. Zou, X. Deng, R. Zuo, Y.-K. Lai, C. Ma, Y.-J. Liu, and H. Wang, "Scenesketcher: Fine-grained image retrieval with scene sketches," in *ECCV*, 2020.

[79] T. Kato, T. Kurita, N. Otsu, and K. Hirata, "A sketch retrieval method for full color image database-query by visual example," in *ICPR*, 1992.

[80] Y. Li, T. M. Hospedales, Y.-Z. Song, and S. Gong, "Fine-grained sketch-based image retrieval by matching deformable part models," in *BMVC*, 2014.

[81] T. de Vries, I. Misra, C. Wang, and L. van der Maaten, "Does object recognition work for everyone?" in *CVPR Workshops*, 2019.

[82] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019, http://www.fairmlbook.org.

[83] Y. Matsui, T. Shiratori, and K. Aizawa, "Drawfromdrawings: 2d drawing assistance via stroke interpolation with a sketch database," *TVCG*, 2016.

[84] K. D. Forbus, B. Garnier, B. Tikoff, W. Marko, M. Usher, and M. McLure, "Sketch worksheets in stem classrooms: Two deployments," in *AAAI*, 2018.

[85] Y. Ye, Y. Lu, and H. Jiang, "Human's scene sketch understanding," in *ICMR*, 2016.

[86] Y. Xie, P. Xu, and Z. Ma, "Deep zero-shot learning for scene sketch," *arXiv preprint arXiv:1905.04510*, 2019.

[87] O. Seddati, S. Dupont, and S. Mahmoudi, "Deepsketch: deep convolutional neural networks for sketch recognition and similarity search," in *CBMI*, 2015.

[88] Y. Zhang, Y. Zhang, and X. Qian, "Deep neural networks for free-hand sketch recognition," in *Pacific Rim Conference on Multimedia*, 2016.

[89] J. Guo, C. Wang, E. Roman-Rangel, H. Chao, and Y. Rui, "Building hierarchical representations for oracle character and sketch recognition," *TIP*, 2016.

[90] P. Ballester and R. M. Araujo, "On the performance of googlenet and alexnet applied to sketches," in *AAAI*, 2016.

[91] O. Seddati, S. Dupont, and S. Mahmoudi, "Deepsketch 2: Deep convolutional neural networks for partial sketch recognition," in *CBMI*, 2016.

[92] H. Zhang, P. She, Y. Liu, J. Gan, X. Cao, and H. Foroosh, "Learning structural representations via dynamic object landmarks discovery for sketch recognition and retrieval," *TIP*, 2019.

[93] O. Seddati, S. Dupont, and S. Mahmoudi, "Deepsketch2image: deep convolutional neural networks for partial sketch recognition and image retrieval," in *MM*, 2016.

[94] K. Zhang, W. Luo, L. Ma, and H. Li, "Cousin network guided sketch recognition via latent attribute warehouse," in *AAAI*, 2019.

[95] X. Zhang, Y. Huang, Q. Zou, Y. Pei, R. Zhang, and S. Wang, "A hybrid convolutional neural network for sketch recognition," *PRL*, 2020.

[96] G. Jain, S. Chopra, S. Chopra, and A. S. Parihar, "Transsketchnet: Attention-based sketch recognition using transformers," in *European Conference on Artificial Intelligence*, 2020.

[97] J. Jiao, Y. Cao, M. Lau, and R. Lau, "Tactile sketch saliency," in *MM*, 2020.

[98] Q. Jia, X. Fan, M. Yu, Y. Liu, D. Wang, and L. J. Latecki, "Coupling deep textural and shape features for sketch recognition," in *MM*, 2020.

[99] R. K. Sarvadevabhatla, J. Kundu *et al.*, "Enabling my robot to play pictionary: Recurrent neural networks for sketch recognition," in *MM*, 2016.

[100] P. Xu, Y. Huang, T. Yuan, T. Xiang, T. M. Hospedales, Y.-Z. Song, and L. Wang, "On learning semantic representations for million-scale free-hand sketches," *arXiv preprint arXiv:2007.04101*, 2020.

[101] Q. Jia, M. Yu, X. Fan, and H. Li, "Sequential dual deep learning with shape and texture features for sketch recognition," *arXiv preprint arXiv:1708.02716*, 2017.

[102] J.-Y. He, X. Wu, Y.-G. Jiang, B. Zhao, and Q. Peng, "Sketch recognition with deep visual-sequential fusion model," in *MM*, 2017.

[103] A. Prabhu, V. Batchu, S. A. Munagala, R. Gajawada, and A. Namboodiri, "Distribution-aware binarization of neural networks for sketch recognition," in *WACV*, 2018.

[104] L. Li, C. Zou, Y. Zheng, Q. Su, H. Fu, and C.-L. Tai, "Sketch-r2cnn: An attentive network for vector sketch recognition," *arXiv preprint arXiv:1811.08170*, 2018.

[105] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012.

[106] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *ECCV*, 2012.

[107] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[108] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[109] A. v. d. Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

[110] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[111] P. Xu, Z. Song, Q. Yin, Y.-Z. Song, and L. Wang, "Deep self-supervised representation learning for free-hand sketch," *arXiv preprint arXiv:2002.00867*, 2020.

[112] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[113] A. Mishra and A. K. Singh, "Deep embedding using bayesian risk minimization with application to sketch recognition," in *ACCV*, 2018.

[114] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, 1995.

[115] A. K. Bhunia, P. N. Chowdhury, Y. Yang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Vectorization and rasterization: Self-supervised learning for sketch and handwriting," in *CVPR*, 2021.

[116] C. Pan, J. Huang, J. Gong, and C. Chen, "Teach machine to learn: hand-drawn multi-symbol sketch recognition in one-shot," *Applied Intelligence*, 2020.

[117] F. Liu, X. Deng, Y.-K. Lai, Y.-J. Liu, C. Ma, and H. Wang, "Sketchgan: Joint sketch completion and recognition with generative adversarial network," in *CVPR*, 2019.

[118] F. Wang and Y. Li, "Spatial matching of sketches without point correspondence," in *ICIP*, 2015.

[119] A. Creswell and A. A. Bharath, "Adversarial training for sketch retrieval," in *ECCV*, 2016.

[120] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.

[121] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014.

[122] S. Balasubramanian, V. N. Balasubramanian *et al.*, "Teaching gans to sketch in vector format," *arXiv preprint arXiv:1904.03620*, 2019.

[123] L. S. F. Ribeiro, T. Bui, J. Collomosse, and M. Ponti, "Sketchformer: Transformer-based representation for sketched structure," in *CVPR*, 2020.

[124] H. Lin, Y. Fu, X. Xue, and Y.-G. Jiang, "Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt," in *CVPR*, 2020.

[125] T. Zhou, C. Fang, Z. Wang, J. Yang, B. Kim, Z. Chen, J. Brandt, and D. Terzopoulos, "Learning to doodle with deep q-networks and demonstrated strokes," in *BMVC*, 2018.

[126] N. Jaques, J. McCleary, J. Engel, D. Ha, F. Bertsch, R. Picard, and D. Eck, "Learning via social awareness: Improving a deep generative sketching model with facial feedback," *arXiv preprint arXiv:1802.04877*, 2018.

[127] K. Sasaki and T. Ogata, "Adaptive drawing behavior by visuomotor learning using recurrent neural networks," *IEEE Transactions on Cognitive and Developmental Systems*, 2019.

[128] J. Li, N. Gao, T. Shen, W. Zhang, T. Mei, and H. Ren, "Sketchman: Learning to create professional sketches," in *MM*, 2020.

[129] S. Ge, V. Goswami, L. Zitnick, and D. Parikh, "Creative sketch generation," in *ICLR*, 2021.

[130] A. Das, Y. Yang, T. Hospedales, T. Xiang, and Y.-Z. Song, "Béziersketch: A generative model for scalable vector sketches," in *ECCV*, 2020.

[131] A. K. Bhunia, A. Das, U. R. Muhammad, Y. Yang, T. M. Hospedales, T. Xiang, Y. Gryaditskaya, and Y.-Z. Song, "Pixelor: A competitive sketching ai agent. so you think you can beat me?" *ACM Transactions on Graphics*, vol. 39, no. 6, 2020.

[132] G. Hinton and V. Nair, "Inferring motor programs from images of handwritten digits," in *NeurIPS*, 2005.

[133] Y. Li, Y.-Z. Song, T. M. Hospedales, and S. Gong, "Free-hand sketch synthesis with deformable stroke models," *IJCV*, 2017.

[134] A. Jenal, N. Savinov, T. Sattler, and G. Chaurasia, "Rnn-based generative model for fine-grained sketching," *arXiv preprint arXiv:1901.03991*, 2019.

[135] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[136] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *TSP*, 1997.

[137] N. Cao, X. Yan, Y. Shi, and C. Chen, "Ai-sketcher: A deep generative model for producing high-quality sketches," in *AAAI*, 2019.

[138] N. Zheng, Y. Jiang, and D. Huang, "Strokenet: A neural painting environment," in *ICLR*, 2019.

[139] Y. Ganin, T. Kulkarni, I. Babuschkin, S. M. A. Eslami, and O. Vinyals, "Synthesizing programs for images using reinforced adversarial learning," in *ICML*, 2018.

[140] J. F. Mellor, E. Park, Y. Ganin, I. Babuschkin, T. Kulkarni, D. Rosenbaum, A. Ballard, T. Weber, O. Vinyals, and S. Eslami, "Unsupervised doodling and painting with improved spiral," *arXiv preprint arXiv:1910.01007*, 2019.

[141] K. Ellis, D. Ritchie, A. Solar-Lezama, and J. B. Tenenbaum, "Learning to infer graphics programs from hand-drawn images," in *NeurIPS*, 2018.

[142] V. Egiazarian, O. Voynov, A. Artemov, D. Volkhonskiy, A. Safin, M. Taktasheva, D. Zorin, and E. Burnaev, "Deep vectorization of technical drawings," *arXiv preprint arXiv:2003.05471*, 2020.

[143] S. Wieluch and F. Schwenker, "Strokecoder: Path-based image generation from single examples using transformers," *arXiv preprint arXiv:2003.11958*, 2020.

[144] J. P. Collomosse, G. McNeill, and L. Watts, "Free-hand sketch grouping for video retrieval," in *ICPR*, 2008.

[145] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh, and R. von der Heydt, "A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization." *Psychological bulletin*, 2012.

[146] K. Koffka, *Principles of Gestalt psychology*. Routledge, 2013.

[147] Z. Sun, C. Wang, L. Zhang, and L. Zhang, "Free hand-drawn sketch segmentation," in *ECCV*, 2012.

[148] Y. Qi, J. Guo, Y.-Z. Song, T. Xiang, H. Zhang, and Z.-H. Tan, "Im2sketch: Sketch generation by unconflicted perceptual grouping," *Neurocomputing*, 2015.

[149] Y. Qi, Y.-Z. Song, T. Xiang, H. Zhang, T. Hospedales, Y. Li, and J. Guo, "Making better use of edges via perceptual grouping," in *CVPR*, 2015.

[150] X. Liu, T.-T. Wong, and P.-A. Heng, "Closure-aware sketch simplification," *TOG*, 2015.

[151] X. Wang, X. Chen, and Z. Zha, "Sketchpointnet: A compact network for robust sketch recognition," in *ICIP*, 2018.

[152] K. Li, K. Pang, Y.-Z. Song, T. Xiang, T. M. Hospedales, and H. Zhang, "Toward deep universal sketch perceptual grouper," *TIP*, 2019.

[153] Z. Huang, H. Fu, and R. W. Lau, "Data-driven segmentation and labeling of freehand sketches," *TOG*, 2014.

[154] R. G. Schneider and T. Tuytelaars, "Example-based sketch segmentation and labeling using crfs," *TOG*, 2016.

[155] B. Kim, O. Wang, A. C. Öztireli, and M. Gross, "Semantic segmentation for line drawing vectorization using neural networks," in *Computer Graphics Forum*, 2018.

[156] K. Kaiyrbekov and M. Sezgin, "Stroke-based sketched symbol reconstruction and segmentation," *arXiv preprint arXiv:1901.03427*, 2019.

[157] X. Wu, Y. Qi, J. Liu, and J. Yang, "Sketchsegnet: A rnn model for labeling sketch strokes," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2018.

[158] L. Li, H. Fu, and C.-L. Tai, "Fast sketch segmentation and labeling with deep learning," *IEEE computer graphics and applications*, 2018.

[159] L. Yi, V. G. Kim, D. Ceylan, I. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, L. Guibas *et al.*, "A scalable active framework for region annotation in 3d shape collections," *TOG*, 2016.

[160] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *TPAMI*, 2017.

[161] C. Wang, B. Yang, and Y. Liao, "Unsupervised image segmentation using convolutional autoencoder with total variation regularization as preprocessing," in *ICASSP*, 2017.

[162] L. Donati, S. Cesano, and A. Prati, "An accurate system for fashion hand-drawn sketches vectorization," in *ICCV Workshops*, 2017.

[163] J. Chen, M. Du, X. Qin, and Y. Miao, "An improved topology extraction approach for vectorization of sketchy line drawings," *The Visual Computer*, 2018.

[164] Y. Guo, Z. Zhang, C. Han, W. Hu, C. Li, and T.-T. Wong, "Deep line drawing vectorization via line subdivision and topology reconstruction," in *Computer Graphics Forum*, 2019.

[165] L. Donati, S. Cesano, and A. Prati, "A complete hand-drawn sketch vectorization framework," *Multimedia Tools and Applications*, 2019.

[166] T. Stanko, M. Bessmeltsev, D. Bommes, and A. Bousseau, "Integer-grid sketch simplification and vectorization," in *Computer Graphics Forum*, 2020.

[167] A. D. Parakkat, M.-P. R. Cani, and K. Singh, "Color by numbers: Interactive structuring and vectorization of sketch imagery," in *CHI*, 2021.

[168] R. K. Sarvadevabhatla, I. Dwivedi, A. Biswas, S. Manocha *et al.*, "Sketchparse: Towards rich descriptions for poorly drawn sketches using multi-task hierarchical deep networks," in *MM*, 2017.

[169] J. Jiang, R. Wang, S. Lin, and F. Wang, "Sfsegnet: Parse freehand sketches using deep fully convolutional networks," in *IJCNN*, 2019.

[170] K. Mukherjee, R. X. Hawkins, and J. E. Fan, "Communicating semantic part information in drawings," in *Annual Conference of the Cognitive Science Society*, 2019.

[171] Y. Zheng, H. Yao, and X. Sun, "Deep semantic parsing of freehand sketches with homogeneous transformation, soft-weighted loss, and staged learning," *arXiv preprint arXiv:1910.06023*, 2019.

[172] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[173] X. Hilaire and K. Tombre, "Robust and accurate vectorization of line drawings," *TPAMI*, 2006.

[174] Y. Chien, W.-C. Lin, T.-S. Huang, and J.-H. Chuang, "Line drawing simplification by stroke translation and combination," in *ICGIP*, 2014.

[175] T. Ogawa, Y. Matsui, T. Yamasaki, and K. Aizawa, "Sketch simplification by classifying strokes," in *ICPR*, 2016.

[176] P. Barla, J. Thollot, and F. X. Sillion, "Geometric clustering for line drawing simplification," in *TOG*, 2005.

[177] E. Simo-Serra, S. Iizuka, and H. Ishikawa, "Mastering sketching: adversarial augmentation for structured prediction," *TOG*, 2018.

[178] X. Xu, M. Xie, P. Miao, W. Qu, W. Xiao, H. Zhang, X. Liu, and T.-T. Wong, "Perceptual-aware sketch simplification based on integrated vgg layers," *TVCG*, 2019.

[179] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[180] R. K. Sarvadevabhatla *et al.*, "Eye of the dragon: Exploring discriminatively minimalist sketch-based abstractions for object categories," in *MM*, 2015.

[181] U. R. Muhammad, Y. Yang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Goal-driven sequential data abstraction," in *ICCV*, 2019.

[182] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *IJCV*, 2015.

[183] R. K. Sarvadevabhatla and R. V. Babu, "Freehand sketch recognition using deep features," *arXiv preprint arXiv:1502.00254*, 2015.

[184] B. Graham, "Spatially-sparse convolutional neural networks," *arXiv preprint arXiv:1409.6070*, 2014.

[185] Y. Zheng, H. Yao, X. Sun, S. Zhang, S. Zhao, and F. Porikli, "Sketch-specific data augmentation for freehand sketch recognition," *arXiv preprint arXiv:1910.06038*, 2019.

[186] R. Liu, Q. Yu, and S. Yu, "An unpaired sketch-to-photo translation model," *arXiv preprint arXiv:1909.08313*, 2019.

[187] Y.-P. Tan, S. R. Kulkarni, and P. J. Ramadge, "A framework for measuring video similarity and its application to video query by example," in *ICIP*, 1999.

[188] Y. Matsui, "Challenge for manga processing: Sketch-based manga retrieval," in *MM*, 2015.

[189] Y. Li, T. M. Hospedales, Y.-Z. Song, and S. Gong, "Fine-grained sketch-based image retrieval by matching deformable part models," in *BMVC*, 2014.

[190] P. Xu, K. Li, Z. Ma, Y.-Z. Song, L. Wang, and J. Guo, "Cross-modal subspace learning for sketch-based image retrieval: A comparative study," in *IC-NIDC*, 2016.

[191] P. Xu, Q. Yin, Y. Huang, Y.-Z. Song, Z. Ma, L. Wang, T. Xiang, W. B. Kleijn, and J. Guo, "Cross-modal subspace learning for fine-grained sketch-based image retrieval," *Neurocomputing*, 2018.

[192] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "Crossatnet-a novel cross-attention based framework for sketch-based image retrieval," *Image and Vision Computing*, 2020.

[193] F. Yang, Y. Wu, Z. Wang, X. Li, S. Sakti, and S. Nakamura, "Instance-level heterogeneous domain adaptation for limited-labeled sketch-to-photo retrieval," *TMM*, 2020.

[194] A. Fuentes and J. M. Saavedra, "Sketch-qnet: A quadruplet convnet for color sketch-based image retrieval," in *CVPR Workshops*, 2021.

[195] P. Torres and J. M. Saavedra, "Compact and effective representations for sketch-based image retrieval," in *CVPR Workshops*, 2021.

[196] T. Bui, L. Ribeiro, M. Ponti, and J. Collomosse, "Deep manifold alignment for mid-grain sketch based image retrieval," in *ACCV*, 2018.

[197] L. Zheng, Y. Yang, and Q. Tian, "Sift meets cnn: A decade survey of instance retrieval," *TPAMI*, 2017.

[198] S.-i. Kondo, M. Toyoura, and X. 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*, 2014.

[199] S. D. Bhattacharjee, J. Yuan, W. Hong, and X. Ruan, "Query adaptive instance search using object sketches," in *MM*, 2016.

[200] J. Lei, K. Zheng, H. Zhang, X. Cao, N. Ling, and Y. Hou, "Sketch based image retrieval via image-aided cross domain learning," in *ICIP*, 2017.

[201] S. D. Bhattacharjee, J. Yuan, Y. Huang, J. Meng, and L. Duan, "Query adaptive multiview object instance search and localization using sketches," *TMM*, 2018.

[202] J. Canny, "A computational approach to edge detection," *TPAMI*, 1986.

[203] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *ECCV*, 2014.

[204] S. Xie and Z. Tu, "Holistically-nested edge detection," in *ICCV*, 2015.

[205] S. Chopra, R. Hadsell, Y. LeCun *et al.*, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR*, 2005.

[206] P. Xu, Q. Yin, Y. Qi, Y.-Z. Song, Z. Ma, L. Wang, and J. Guo, "Instance-level coupled subspace learning for fine-grained sketch-based image retrieval," in *ECCV Workshops*, 2016.

[207] Y. Qi, Y.-Z. Song, H. Zhang, and J. Liu, "Sketch-based image retrieval via siamese convolutional neural network," in *ICIP*, 2016.

[208] J. Song, Y.-Z. Song, T. Xiang, T. M. Hospedales, and X. Ruan, "Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval." in *BMVC*, 2016.

[209] Q. Yu, X. Chang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "The devil is in the middle: Exploiting mid-level representations for cross-domain instance matching," *arXiv preprint arXiv:1711.08106*, 2017.

[210] Y. Yan, X. Wang, X. Yang, X. Bai, and W. Liu, "Joint classification loss and histogram loss for sketch-based image retrieval," in *ICIG*, 2017.

[211] J. Collomosse, T. Bui, M. J. Wilber, C. Fang, and H. Jin, "Sketching with style: Visual search with sketches and aesthetic context," in *ICCV*, 2017.

[212] J. Song, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Fine-grained image retrieval: the text/sketch input dilemma," in *BMVC*, 2017.

[213] F. Huang, Y. Cheng, C. Jin, Y. Zhang, and T. Zhang, "Deep multimodal embedding model for fine-grained sketch-based image retrieval," in *SIGIR*, 2017.

[214] S. Dey, A. Dutta, S. K. Ghosh, E. Valveny, J. Lladós, and U. Pal, "Learning cross-modal deep embeddings for multi-object image retrieval using text and sketch," in *ICPR*, 2018.

[215] Y. Wang, F. Huang, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval," *PR*, 2019.

[216] K. Pang, K. Li, Y. Yang, H. Zhang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Generalising fine-grained sketch-based image retrieval," in *CVPR*, 2019.

[217] T. Dutta and S. Biswas, "s-sbir: Style augmented sketch based image retrieval," in *WACV*, 2020.

[218] A. K. Bhunia, Y. Yang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Sketch less for more: On-the-fly fine-grained sketch based image retrieval," in *CVPR*, 2020.

[219] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML*, 2013.

[220] F. Huang, C. Jin, Y. Zhang, and T. Zhang, "Towards sketch-based image retrieval with deep cross-modal correlation learning," in *ICME*, 2017.

[221] D. Xu, X. Alameda-Pineda, J. Song, E. Ricci, and N. Sebe, "Cross-paced representation learning with partial curricula for sketch-based image retrieval," *TIP*, 2018.

[222] C. Li, Y. Zhou, and J. Yang, "Sketch-based image retrieval via a semi-heterogeneous cross-domain network," in *ICME Workshops*, 2019.

[223] J. Collomosse, T. Bui, and H. Jin, "Livesketch: Query perturbations for guided sketch-based visual search," in *CVPR*, 2019.

[224] O. Seddati, S. Dupont, and S. Mahmoudi, "Quadruplet networks for sketch-based image retrieval," in *ICMR*, 2017.

[225] T. Bui, L. Ribeiro, M. Ponti, and J. Collomosse, "Generalisation and sharing in triplet convnets for sketch based visual search," *arXiv preprint arXiv:1611.05301*, 2016.

[226] H. Lin, Y. Fu, P. Lu, S. Gong, X. Xue, and Y.-G. Jiang, "Tc-net for isbir: Triplet classification network for instance-level sketch based image retrieval," in *MM*, 2019.

[227] L. Guo, J. Liu, Y. Wang, Z. Luo, W. Wen, and H. Lu, "Sketch-based image retrieval using generative adversarial networks," in *MM*, 2017.

[228] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[229] K. Pang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Cross-domain generative learning for fine-grained sketch-based image retrieval." in *BMVC*, 2017.

[230] F. Huang, C. Jin, Y. Zhang, K. Weng, T. Zhang, and W. Fan, "Sketch-based image retrieval with deep visual semantic descriptor," *PR*, 2018.

[231] T. Bui, L. Ribeiro, M. Ponti, and J. Collomosse, "Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network," *CVIU*, 2017.

[232] J. Lei, Y. Song, B. Peng, Z. Ma, L. Shao, and Y.-Z. Song, "Semi-heterogeneous three-way joint embedding network for sketch-based image retrieval," *TCSVT*, 2019.

[233] H. Zhang, C. Zhang, and M. Wu, "Sketch-based cross-domain image retrieval via heterogeneous network," in *VCIP*, 2017.

[234] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, "Deep sketch hashing: Fast free-hand sketch-based image retrieval," in *CVPR*, 2017.

[235] J. Zhang, F. Shen, L. Liu, F. Zhu, M. Yu, L. Shao, H. Tao Shen, and L. Van Gool, "Generative domain-migration hashing for sketch-to-image retrieval," in *ECCV*, 2018.

[236] G. Tolias and O. Chum, "Asymmetric feature maps with application to sketch based retrieval," in *CVPR*, 2017.

[237] K. Lin, H.-F. Yang, J.-H. Hsiao, and C.-S. Chen, "Deep learning of binary hash codes for fast image retrieval," in *CVPR Workshops*, 2015.

[238] J. Wang, T. Zhang, N. Sebe, H. T. Shen *et al.*, "A survey on learning to hash," *TPAMI*, 2017.

[239] K. Pang, Y. Yang, T. Hospedales, T. Xiang, and Y.-Z. Song, "Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval," in *CVPR*, 2020.

[240] F. Radenovic, G. Tolias, and O. Chum, "Deep shape matching," in *ECCV*, 2018.

[241] A. Sablayrolles, M. Douze, N. Usunier, and H. Jégou, "How should we evaluate supervised hashing?" in *ICASSP*, 2017.

[242] P. Lu, G. Huang, Y. Fu, G. Guo, and H. Lin, "Learning large euclidean margin for sketch-based image retrieval," *arXiv preprint arXiv:1812.04275*, 2018.

[243] W. Thong, P. Mettes, and C. G. Snoek, "Open cross-domain visual search," *arXiv preprint arXiv:1911.08621*, 2019.

[244] T. Dutta and S. Biswas, "Style-guided zero-shot sketch-based image retrieval," in *BMVC*, 2019.

[245] A. Dutta and Z. Akata, "Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval," in *CVPR*, 2019.

[246] S. Kiran Yelamarthi, S. Krishna Reddy, A. Mishra, and A. Mittal, "A zero-shot framework for sketch based image retrieval," in *ECCV*, 2018.

[247] J. Li, Z. Ling, L. Niu, and L. Zhang, "Bi-directional domain translation for zero-shot sketch-based image retrieval," *arXiv preprint arXiv:1911.13251*, 2019.

[248] A. Pandey, A. Mishra, V. Kumar Verma, and A. Mittal, "Adversarial joint-distribution learning for novel class sketch-based image retrieval," in *ICCV Workshops*, 2019.

[249] V. Kumar Verma, A. Mishra, A. Mishra, and P. Rai, "Generative model for zero-shot sketch-based image retrieval," in *CVPR Workshops*, 2019.

[250] Q. Liu, L. Xie, H. Wang, and A. L. Yuille, "Semantic-aware knowledge preservation for zero-shot sketch-based image retrieval," in *ICCV*, 2019.

[251] A. Pandey, A. Mishra, V. K. Verma, A. Mittal, and H. Murthy, "Stacked adversarial network for zero-shot sketch based image retrieval," in *WACV*, 2020.

[252] X. Xu, C. Deng, M. Yang, and H. Wang, "Progressive domain-independent feature decomposition network for zero-shot sketch-based image retrieval," *arXiv preprint arXiv:2003.09869*, 2020.

[253] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "A simplified framework for zero-shot cross-modal sketch data retrieval," in *CVPR Workshops*, 2020.

[254] Z. Zhang, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Zero-shot sketch-based image retrieval via graph convolution network," in *AAAI*, 2020.

[255] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "A zero-shot sketch-based inter-modal object retrieval scheme for remote sensing images," *arXiv preprint arXiv:2008.05225*, 2020.

[256] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Transductive multi-view zero-shot learning," *TPAMI*, 2015.

[257] Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, L. Sigal, and S. Gong, "Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content," *IEEE Signal Processing Magazine*, 2018.

[258] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *TIST*, 2019.

[259] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NeurIPS*, 2013.

[260] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *TPAMI*, 2013.

[261] Y. Shen, L. Liu, F. Shen, and L. Shao, "Zero-shot sketch-image hashing," in *CVPR*, 2018.

[262] W. Chen and J. Hays, "Sketchygan: Towards diverse and realistic sketch to image synthesis," in *CVPR*, 2018.

[263] Y. Jo and J. Park, "Sc-fegan: Face editing generative adversarial network with user's sketch and color," in *ICCV*, 2019.

[264] S. Yang, Z. Wang, J. Liu, and Z. Guo, "Deep plastic surgery: Robust and controllable image editing with human-drawn sketches," *arXiv preprint arXiv:2001.02890*, 2020.

[265] W. Xia, Y. Yang, and J.-H. Xue, "Cali-sketch: Stroke calibration and completion for high-quality face image generation from poorly-drawn sketches," *arXiv preprint arXiv:1911.00426*, 2019.

[266] A. Ghosh, R. Zhang, P. K. Dokania, O. Wang, A. A. Efros, P. H. S. Torr, and E. Shechtman, "Interactive sketch & fill: Multiclass sketch-to-image translation," in *ICCV*, 2019.

[267] Z. Li, C. Deng, E. Yang, and D. Tao, "Staged sketch-to-image synthesis via semi-supervised generative adversarial networks," *TMM*, 2020.

[268] B. Liu, K. Song, and A. Elgammal, "Sketch-to-art: Synthesizing stylized art images from sketches," *arXiv preprint arXiv:2002.12888*, 2020.

[269] Y. Lu, S. Wu, Y.-W. Tai, and C.-K. Tang, "Image generation from sketch constraint using contextual gan," in *ECCV*, 2018.

[270] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, and J. Hays, "Texturegan: Controlling deep image synthesis with texture patches," in *CVPR*, 2018.

[271] M. Li, Z. Lin, R. Mech, E. Yumer, and D. Ramanan, "Photo-sketching: Inferring contour drawings from images," in *WACV*, 2019.

[272] M. Kampelmuhler and A. Pinz, "Synthesizing human-like sketches from natural images using a conditional convolutional decoder," in *WACV*, 2020.

[273] S.-Y. Chen, W. Su, L. Gao, S. Xia, and H. Fu, "Deepfacedrawing: deep generation of face images from sketches," *TOG*, 2020.

[274] J. Huang, J. Liao, Z. Tan, and S. Kwong, "Multi-density sketch-to-image translation network," *arXiv preprint arXiv:2006.10649*, 2020.

[275] Y. Zhang, G. Su, Y. Qi, and J. Yang, "Unpaired image-to-sketch translation network for sketch synthesis," in *VCIP*, 2019.

[276] J. Li, S. Liu, and M. Cao, "Line artist: A multiple style sketch to painting synthesis scheme," *arXiv preprint arXiv:1803.06647*, 2018.

[277] M. Li, A. Sheffer, E. Grinspun, and N. Vining, "Foldsketch: Enriching garments with physically reproducible folds," *TOG*, 2018.

[278] T. Y. Wang, D. Ceylan, J. Popovic, and N. J. Mitra, "Learning a shared shape space for multimodal garment design," *arXiv preprint arXiv:1806.11335*, 2018.

[279] J. Collomosse, T. Bui, M. J. Wilber, C. Fang, and H. Jin, "Sketching with style: Visual search with sketches and aesthetic context," in *ICCV*, 2017.

[280] C. Zou, H. Mo, R. Du, X. Wu, C. Gao, and H. Fu, "Lucss: Language-based user-customized colourization of scene sketches," *arXiv preprint arXiv:1808.10544*, 2018.

[281] L. Zhang, C. Li, T.-T. Wong, Y. Ji, and C. Liu, "Two-stage sketch colorization," *TOG*, 2018.

[282] C. Zou, H. Mo, C. Gao, R. Du, and H. Fu, "Language-based colorization of scene sketches," *TOG*, 2019.

[283] X. Soria, E. Riba, and A. D. Sappa, "Dense extreme inception network: Towards a robust cnn model for edge detection," *arXiv preprint arXiv:1909.01955*, 2019.

[284] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017.

[285] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[286] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[287] K. Sasaki, S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Joint gap detection and inpainting of line drawings," in *CVPR*, 2017.

[288] H. Chen, M. V. Giuffrida, P. Doerner, and S. A. Tsaftaris, "Blind inpainting of large-scale masks of thin structures with adversarial and reinforcement learning," *arXiv preprint arXiv:1912.02470*, 2019.

[289] H. Chen, M. V. Giuffrida, S. A. Tsaftaris, and P. Doerner, "Root gap correction with a deep inpainting model," in *BMVC*, 2018.

[290] H. Chen, M. Valerio Giuffrida, P. Doerner, and S. A. Tsaftaris, "Adversarial large-scale root gap inpainting," in *CVPR Workshops*, 2019.

[291] T. Shao, W. Xu, K. Yin, J. Wang, K. Zhou, and B. Guo, "Discriminative sketch-based 3d model retrieval via robust shape matching," in *Computer Graphics Forum*, 2011.

[292] B. Li, Y. Lu, and J. Shen, "A semantic tree-based approach for sketch-based 3d model retrieval," in *ICPR*, 2016.

[293] D. DeCarlo, A. Finkelstein, S. Rusinkiewicz, and A. Santella, "Suggestive contours for conveying shape," *TOG*, 2003.

[294] H. Li, H. Wu, X. He, S. Lin, R. Wang, and X. Luo, "Multi-view pairwise relationship learning for sketch based 3d shape retrieval," in *ICME*, 2017.

[295] F. Zhu, J. Xie, and Y. Fang, "Learning cross-domain neural networks for sketch-based 3d shape retrieval," in *AAAI*, 2016.

[296] Y. Ye, B. Li, and Y. Lu, "3d sketch-based 3d model retrieval with convolutional neural network," in *ICPR*, 2016.

[297] J. Chen and Y. Fang, "Deep cross-modality adaptation via semantics preserving adversarial learning for sketch-based 3d shape retrieval," in *ECCV*, 2018.

[298] J. Chen, J. Qin, L. Liu, F. Zhu, F. Shen, J. Xie, and L. Shao, "Deep sketch-shape hashing with segmented 3d stochastic viewing," in *CVPR*, 2019.

[299] A. Qi, Y. Gryaditskaya, J. Song, Y. Yang, Y. Qi, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Toward fine-grained sketch-based 3d shape retrieval," *IEEE Transactions on Image Processing*, vol. 30, pp. 8595–8606, 2021.

[300] A. Qi, Y.-Z. Song, and T. Xiang, "Semantic embedding for sketch-based 3d shape retrieval." in *BMVC*, 2018.

[301] S. Kuwabara, R. Ohbuchi, and T. Furuya, "Query by partially-drawn sketches for 3d shape retrieval," in *2019 International Conference on Cyberworlds*, 2019.

[302] G. Dai, J. Xie, F. Zhu, and Y. Fang, "Deep correlated metric learning for sketch-based 3d shape retrieval," in *AAAI*, 2017.

[303] G. Dai, J. Xie, and Y. Fang, "Deep correlated holistic metric learning for sketch-based 3d shape retrieval," *TIP*, 2018.

[304] J. Xie, G. Dai, F. Zhu, and Y. Fang, "Learning barycentric representations of 3d shapes for sketch-based 3d shape retrieval," in *CVPR*, 2017.

[305] V. I. Bogachev and A. V. Kolesnikov, "The monge-kantorovich problem: achievements, connections, and perspectives," *Russian Mathematical Surveys*, 2012.

[306] L. Wang, C. Qian, J. Wang, and Y. Fang, "Unsupervised learning of 3d model reconstruction from hand-drawn sketches," in *MM*, 2018.

[307] S.-H. Zhang, Y.-C. Guo, and Q.-W. Gu, "Sketch2model: View-aware 3d modeling from single free-hand sketches," in *CVPR*, 2021.

[308] Y. Shen, C. Zhang, H. Fu, K. Zhou, and Y. Zheng, "Deepsketchhair: Deep sketch-based 3d hair modeling," *arXiv preprint arXiv:1908.07198*, 2019.

[309] K. D. Willis, P. K. Jayaraman, J. G. Lambourne, H. Chu, and Y. Pu, "Engineering sketch generation for computer-aided design," in *CVPR Workshops*, 2021.

[310] H. Huang, E. Kalogerakis, E. Yumer, and R. Mech, "Shape synthesis from sketches via procedural models and convolutional networks," *TVCG*, 2016.

[311] X. Han, C. Gao, and Y. Yu, "Deepsketch2face: a deep learning based sketching system for 3d face and caricature modeling," *TOG*, 2017.

[312] M. Ye, S. Zhou, and H. Fu, "Deepshapesketch: Generating hand drawing sketches from 3d objects," in *IJCNN*, 2019.

[313] J. P. Collomosse, G. McNeill, and Y. Qian, "Storyboard sketches for content based video retrieval," in *ICCV*, 2009.

[314] P. Xu, K. Liu, T. Xiang, T. M. Hospedales, Z. Ma, J. Guo, and Y.-Z. Song, "Fine-grained instance-level sketch-based video retrieval," *arXiv preprint arXiv:2002.09461*, 2020.

[315] S. Wu, H. Su, S. Zheng, H. Yang, and Q. Zhou, "Motion sketch based crowd video retrieval via motion structure coding," in *ICIP*, 2016.

[316] S. Wu, H. Yang, S. Zheng, H. Su, Q. Zhou, and X. Lu, "Motion sketch based crowd video retrieval," *Multimedia Tools and Applications*, 2017.

[317] F. Huang and J. F. Canny, "Sketchforme: Composing sketched scenes from text descriptions for interactive applications," *arXiv preprint arXiv:1904.04399*, 2019.

[318] F. Huang, E. Schoop, D. Ha, and J. Canny, "Scones: towards conversational authoring of sketches," in *International Conference on Intelligent User Interfaces*, 2020.

[319] C. Hu, D. Li, Y. Yang, T. M. Hospedales, and Y.-Z. Song, "Sketch-a-segmenter: Sketch-based photo segmenter generation," *IEEE Transactions on Image Processing*, vol. 29, pp. 9470–9481, 2020.

[320] R. K. Sarvadevabhatla, S. Surya, T. Mittal, and R. V. Babu, "Game of sketches: Deep recurrent models of pictionary-style word guessing," in *AAAI*, 2018.

[321] K. Pang, D. Li, J. Song, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Deep factorised inverse-sketching," in *ECCV*, 2018.

[322] A. Tripathi, R. R. Dani, A. Mishra, and A. Chakraborty, "Sketch-guided object localization in natural images," *arXiv preprint arXiv:2008.06551*, 2020.

[323] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[324] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[325] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018.

[326] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[327] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," in *ICLR*, 2018.

[328] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[329] J.-D. Favreau, F. Lafarge, and A. Bousseau, "Fidelity vs. simplicity: a global approach to line drawing vectorization," *TOG*, 2016.

[330] M. Bessmeltsev and J. Solomon, "Vectorization of line drawings via polyvector fields," *TOG*, 2019.

[331] H. Mo, E. Simo-Serra, C. Gao, C. Zou, and R. Wang, "General Virtual Sketching Framework for Vector Line Art," *TOG*, 2021.

[332] R. B. Venkataramaiyer, A. Joshi, S. Narang, and V. P. Namboodiri, "Shad3s: A model to sketch, shade and shadow," in *WACV*, 2021.

[333] E. Simo-Serra, S. Iizuka, and H. Ishikawa, "Real-Time Data-Driven Interactive Rough Sketch Inking," *TOG*, 2018.

[334] P. Xu, H. Fu, Y. Zheng, K. Singh, H. Huang, and C.-L. Tai, "Model-guided 3d sketching," *TVCG*, 2018.

[335] B. Jackson and D. F. Keefe, "Lift-off: Using reference imagery and freehand sketching to create 3d models in vr," *TVCG*, 2016.

[336] D. Giunchi, D. Degraen, A. Steed *et al.*, "Mixing realities for sketch retrieval in virtual reality," *arXiv preprint arXiv:1910.11637*, 2019.

[337] J. C. Roberts, C. Headleand, and P. D. Ritsos, "Sketching designs using the five design-sheet methodology," *TVCG*, 2015.

[338] F. Boniardi, A. Valada, W. Burgard, and G. D. Tipaldi, "Autonomous indoor robot navigation using a sketch interface for drawing maps and routes," in *ICRA*, 2016.

[339] V. Jain, P. Agrawal, S. Banga, R. Kapoor, and S. Gulyani, "Sketch2code: Transformation of sketches to ui in real-time using deep neural network," *arXiv preprint arXiv:1910.08930*, 2019.

[340] B. Paulson and T. Hammond, "Paleosketch: accurate primitive sketch recognition and beautification," in *Proceedings of the 13th international conference on Intelligent user interfaces*, 2008, pp. 1–10.

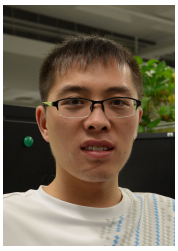
[341] D. Sỳkora, J. Dingliana, and S. Collins, "Lazybrush: Flexible painting tool for hand-drawn cartoons," in *Computer Graphics Forum*, 2009.

[342] D. Sỳkora, M. Ben-Chen, M. Čadík, B. Whited, and M. Simmons, "Textoons: practical texture mapping for hand-drawn cartoon animations," in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Non-Photorealistic Animation and Rendering*, 2011.

[343] Y. Liu, Z. Qin, T. Wan, and Z. Luo, "Auto-painter: Cartoon image generation from sketch by using conditional wasserstein generative adversarial networks," *Neurocomputing*, 2018.
[344] H. Kim, H. Y. Jhoo, E. Park, and S. Yoo, "Tag2pix: Line art colorization using text tag with secat and changing loss," in *ICCV*, 2019.
[345] M. Hudon, M. Grogan, A. Smolic *et al.*, "Deep normal estimation for automatic shading of hand-drawn characters," in *ECCV Workshops*, 2018.
[346] W. Su, D. Du, X. Yang, S. Zhou, and H. Fu, "Interactive sketch-based normal map generation with deep neural networks," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 2018.
[347] K. T. Yesilbek and M. Sezgin, "On training sketch recognizers for new domains," in *CVPR Workshops*, 2021.

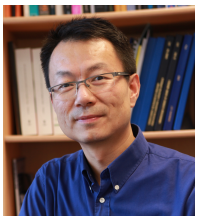
Peng Xu is currently a postdoctoral research assistant in Department of Engineering Science, University of Oxford. Previously he was a postdoctoral research fellow in School of Computer Science and Engineering, Nanyang Technological University, Singapore. He received his PhD degree from Beijing University of Posts and Telecommunications, China. His research interests include fine-grained sketch analysis, multimodal computing, and computer vision.

Timothy M. Hospedales is a Professor within IPAB in the School of Informatics at the University of Edinburgh. His research focuses on machine learning, particularly life-long transfer learning, with both probabilistic and deep learning approaches. He has studied at a variety application areas including computer vision, robotics, natural language processing and beyond.



Qiyue Yin received PhD degree in Pattern Recognition and Intelligent Systems from the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2017. Now he serves as an Associate Professor in CASIA. His major research interests include pattern recognition, deep learning and artificial intelligence on games.

Yi-Zhe Song is a Professor of Computer Vision and Machine Learning at the Centre for Vision Speech and Signal Processing (CVSSP), UK's largest academic research centre for Artificial Intelligence with approx. 200 researchers. Previously, he was a Senior Lecturer at the Queen Mary University of London, and a Research and Teaching Fellow at the University of Bath.



Tao Xiang received the BS degree from Xi'an Jiaotong University, Xi'an, China, in 1995, and the PhD degree from the National University of Singapore, in 2001. He is a Professor of Computer Vision and Machine Learning and Distinguished Chair at Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. His research interests include computer vision, pattern recognition and machine learning.

Liang Wang received both the BEng and MEng degrees from Anhui University in 1997 and 2000, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2004. From 2004 to 2010, he was a research assistant at Imperial College London, United Kingdom, and Monash University, Australia, a research fellow at the University of Melbourne, Australia, and a lecturer at the University of Bath, United Kingdom, respectively. Currently, he is a full professor of the Hundred Talents Program at the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published in highly ranked international journals such as TPAMI and TIP, and leading international conferences such as CVPR, ICCV, and ICDM. He is an IEEE Fellow, and an IAPR Fellow.