Sketch Fewer to Recognize More by Learning a Co-Regularized Sparse Representation

Yonggang Qi, *Member, IEEE*, and Yi-Zhe Song, *Senior Member, IEEE*

***Abstract*— Categorizing free-hand human sketches has profound implications in applications such as human computer interaction and image retrieval. The task is non-trivial due to the iconic nature of sketches, signified by large variances in both appearance and structure when compared with photographs. Despite recent advances made by deep learning methods, the requirement of a large training set is commonly imposed making them impractical for real-world applications where training sketches are cumbersome to obtain – sketches have to be hand-drawn one by one other than crawled freely on the Internet. In this work, we aim to delve further into the data scarcity problem of sketch-related research, by proposing a few-shot sketch classification framework. The model is based on a co-regularized embedding algorithm where common/shareable parts of learned human sketches are exploited, thereby can embed query sketch into a co-regularized sparse representation space for few-shot classification. A new dataset of 8,000 part-level annotated sketches of 100 categories is also proposed to facilitate future research. Experiment shows that our approach can achieve an 5-way one-shot classification accuracy of 85%, and 20-way one-shot at 51%.**

***Index Terms*— Sketch categorization, few-shot classification.**

I. Introduction

SKETCHES are intuitive to humans and descriptive in nature, and have been used to visually record the world since ancient times in forms like fresco and oracle bones. Research on sketches has flourished in recent years, largely due to the prevailing nature of touchscreens. In particular, tasks such as sketch recognition [1]–[3] and sketch-based image retrieval (SBIR) [4]–[8] have been studied in depth. It had become common-ground that in order to effectively solve sketch-related problems, the following unique characteristics of sketches need to be carefully accommodated for: (i) sketches are often highly abstract in representation, and (ii) sketches are hand-drawn by non-expert with different levels of artistic skills, which consequently leads to large diversity and richness of sketch appearance, even for the same sketch object with the same pose.

Early attempts [1], [2], [4] address the sketch recognition problem follow a traditional supervised learning pipeline widely adopted for object recognition. That is, first a large number (e.g. hundreds) of labeled instances are collected for each class, followed by feature extraction and finally learning a classifier. Many existing works focus on designing features specifically engineered for sketches [9]–[11] due to the unique characteristics of sketches described above.

Recently, tremendous advances on photo recognition tasks brought by deep learning found its way to sketch recognition, and achieved promising results. Yu *et al.* [3] was first to show that machines can beat humans on the task of sketch recognition by specially designing a deep neural network, i.e. Sketch-A-Net. However, it is notorious for requiring extensive and incremental training on large training set. In addition, data augmentation techniques are commonly applied as well to alleviate over-fitting.

In contrast, humans are highly capable of learning new categories with very little supervision, e.g. given a few, even just a single picture of "penguin", a child can recognize it reliably later on. This is because humans learn more efficient than machines do [12], humans typically (i) parse objects into parts and relations, (ii) create abstract of new concepts based on prior knowledge of existing categories, (iii) identify new instances by matching the learned abstract. To make machines recognize new objects more efficiently as humans do, a number of few-shot learning classification algorithms [13]–[21] have made significant progress toward this goal. Vinyals et al. [14] proposed *Matching Networks* to learn a weighted nearest-neighbor classifier within an embedding space, and notably a *episodes* tactic is used, which aims to mimic the few-shottask by sub-sampling classes as well as data points, therefore can make the training problem more faithful to the testing stage and thereby improves generalization. Snell *et al.* [13] proposed *Prototypical Networks* to learn a non-linear mapping from the very limited data in support set into an embedding space, by using a neural network to find *prototype* of a new class, hence classification can be achieved by measuring the nearest class prototype for any new input data point.

However, most of current arts are designed either for dealing with photos (e.g. *mini*ImageNet [14]) or relatively easy recognition task (e.g. Omniglot handwritten characters dataset [22], obtained over 96% accuracy on 20-way 1-shot setting). As far as we know, very few attempted to develop sketch-oriented

Manuscript received June 14, 2019; revised September 18, 2019; accepted December 31, 2019. Date of publication January 3, 2020; date of current version December 4, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61601042 and in part by the 111 Project of China under Grant B08004 and Grant B17007. This article was recommended by Associate Editor G. Hua. *(Corresponding author: Yonggang Qi.)*

Yonggang Qi is with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: qiyg@bupt.edu.cn).

Yi-Zhe Song is with the SketchX Research Laboratory, Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey, Guildford GU2 7XH, U.K. (e-mail: y.song@surrey.ac.uk).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2019.2963862

few-shot learning models. This problem is much more acute for sketches because it is more expensive to collect sketches than photos. As a result, sketches per category is often limited, and more importantly, the number of categories are scarce when compared to common photo datasets. Note that although the recent proposed very large sketch dataset QuickDraw [23] offers 50 million drawings across 345 categories, the amount of sketch is still far less being satisfactory comparing against with the photograph dataset ImageNet, which is organized according to the WordNet hierarchy to cover all possible visual concepts, offers over 20,000 "synsets" or classes with over total 14 billion images. Hence the problem of lacking training data remains for sketch when it comes to learning to recognize an new category.

In this work, we develop a novel few-shot learning framework for addressing the problem of sketch recognition given only very few labeled samples. Inspired by previous one-shot learning work [24], the framework takes advantage of knowledge gained from previously learned sketch categories to make learning unseen categories more efficient. In particular, we make use of common sketch parts/basis learned from an auxiliary set labeled by human and utilize them in a sparse coding based learning framework. Our underlying hypothesis is that common parts exist among sketches from distinct object categories (e.g. wings of 'bird' and 'airplane', the legs of 'cat' and 'dog').

The common parts can then be learned to form a set of sparse basis from the auxiliary set and used as transferable knowledge to help learn a classification model for the target classes. We importantly introduce a novel co-regularized sparse coding algorithm where the sparse coding models for both the auxiliary set and target set are learned jointly. The objective is to make sure that the resulting sparse representation agrees as much as possible between the two sets. More specifically, we propose a co-regularized embedding (CRE) algorithm for few-shot learning of sketch categories, there are two main stages. (i) Basis discovery: given very few example sketches from never-before-seen categories, i.e. target set, sharable basis (e.g. bird wings for replacing wings of airplane) are discovered from learned categories, i.e. auxiliary set, for each novel sketch category in target set. (ii) Embedding space mapping: considering which categories an query sketch most relevant to, the sketch is mapped into an embedding space via the proposed co-regularized sparse coding algorithm, which enforces the resulting sparse representation to use as much as possible the relevant basis discovered during the first stage (e.g. suppose we know the given query sketch is an 'apple', it would likely to be encoded by the basis of 'tomato' and 'peach' in the auxiliary set because of their similar looking). To perform categorization, we employ a simple sparse representation classifier (SRC) [25]. Intuitively, the first step of our algorithm aims to select the sharable/common basis for each novel category, followed by the second step that encourage query sketches to be reconstructed with chosen basis. This is not only to handle the over-complete dictionary problem [26] but also leads to smaller reconstruction error to facilitate SRC classification.

The contributions of this work can be summarized as follows:

(i) A novel few-shot learning framework based on co-regularized sparse coding algorithm is developed, which can be well generalized to one-shot learning of sketch categories. To the best of our knowledge, this is the first work on few-shot learning for sketch recognition.

(ii) We contribute a novel sketch dataset *SketchPart* which is built upon TU-Berlin sketch dataset [1]. *SketchPart* is the first dataset of sketch parts, where consists of 8000 sketches across 100 categories with 13655 parts are manually labeled.

(iii) We demonstrate that our approach is more efficient and can achieve improved few-shot recognition performance comparing against with several baseline methods on the proposed *SketchPart* dataset.

II. Related Works

A. Sketch Recognition

There exist plenty of works on sketch recognition with "non-deep" architecture [1], [10], most of which employ a bag-of-visual-words (BoVW) representation coupled with local features. Some of the features can be commonly found in the vision literature [4], while others are specifically engineered for sketches [9], [11]. Of all "non-deep" features tested, it was shown that Histogram of Oriented Gradient (HOG) based features are among the most effective ones [2], [4]. Despite being useful, unstructured local features are often incapable of capturing the relatively high degree of intra-class variance and inter-class ambiguity associated with human sketches. Yi *et al.* [10] tackled this problem by proposing a novel mid-level sketch representation in the form of a star-graph that encapsulates local features to encode holistic object structure.

Recently, Yu *et al.* [3] proposed Sketch-A-Net that exploits sequential ordering information in sketches to capture multiple levels of abstraction by using deep convolutional neural network, thereby can beat humans and offer the state-of-the-art performance to date on TU-Berlin dataset [1]. Nonetheless, existing approaches often assume the availability of a large number of training sketch data, which seriously limits their scalability to new categories.

To utilize the powerful representation learning ability of deep model, Sketch-A-Net is applied as feature descriptor. Based on this, it is able to further explore the task of few-shot learning on sketch recognition in a more subtle manner. Hence this work is to show how this subtle (parts) clue works for few-shot sketch recognition, and such subtle clue works well even with a "non-deep" routine, i.e. a sparse coding based approach.

B. Few-Shot Classification

Few-shot classification [13], [14], [19] is to train a classifier to be efficiently capable of identifying never-before-seen classes with only very few examples of each of these classes. Although showing convincing performances, most current best models require complex deep neural network architectures [14], [15], [19], such as recurrent neural networks (RNNs), and often require immense iterations [13] to fine-tune the large amount of parameters. Sketch recognition is a typical few-shot learning problem, which often suffers from lacking of training data for learning new sketch categories. However, there has very few work in the literature for dealing with this limitation. Inspired by the observations about how humans learn new concepts, we illustrate that by exploiting sketch parts, a simpler framework based on spare coding is proposed in this work that can offer compelling performances on few-shot sketch recognition.

C. Sparse Coding

For transferring part-based knowledge of sketches, a co-regularized sparse coding model is developed. Sparse coding has been widely used in image classification [27], [28], face recognition [29], visual tracking [30] and many other computer vision areas. However, there are very limited previous work on how to use of sparse coding for sketch recognition, despite the fact that sparse coding is intrinsically appropriate for mining sharable parts from human segmented sketches.

Therefore, we propose a novel co-regularized sparse coding model and apply it to the new problem of few-shot sketch recognition. An early version of this work is [31], in this work there are three significant extensions comparing with the early version: (i) Technically the framework in this work is for few-shot recognition, which is more flexible that can be easily generalized to deal with one-shot case, while the early version is only capable for one-shot learning. Specifically, a very simple while efficient “stackable” basis discovery strategy is used, i.e. the parts basis given by every single sample sketch are gathered together as one union for a never-before-seen category. In addition, a dictionary learning step is applied to further make basis set more representative as well as to reduce the redundancy. (ii) There is a super extension about the dataset, we explicitly define over 140 semantic labels of sketch part and manually label each part out at stroke-level for every sketch. All these information are not available in the early version. Also the number of sketch category in the dataset is extended to 100. (iii) Apart from comparing against with “non-deep” methods, three “deep” models are compared. This significantly illustrate the effectiveness of using subtle clues for few-shot learning even just take a “non-deep” routine.

III. SketchPart Dataset: A Large Manually Labeled Sketch Part Dataset

Our approach is based on learning transferable knowledge by exploiting sketch parts of learned categories. Hence, in this section we begin by describing the construction of a sketch parts dataset, in which semantic parts of each sketch are manually labeled. A similar dataset was proposed by Schneider and Tuytelaars [32] for sketch segmentation, while there are only 6 categories with 480 sketches in total. We extend the dataset to 100 categories where 8000 sketches are involved.

A. Data Collection

In this work, one hundred commonly used sketch object categories (80 sketches per category, and 8,000 sketches totally) are selected from TU-Berlin sketch dataset [1]. And the strokes of each sketch will be assigned with a semantic label, which is predefined carefully. Then 10 annotators are recruited to assign a label to each stroke of sketch objects by using a MATLAB tool.

B. Semantic Part Label

The concepts of parts are predefined for each specific sketch category, annotators are then asked to manually assign strokes of each sketch with semantic part labels. In total, there are $13,655$ parts labeled on the given over 140 different semantic part labels. A subset of annotated sketches are illustrated in Fig. 1, where semantic parts are color-coded. We can see that objects are parsed into smaller components. This is the first dataset explicitly demonstrate how people draw parts, where especially contains many subtle concepts (e.g. Flame, sound, water and light) which do not studied in object-level sketch dataset. Fig. 4 shows some different instances how people draw sound wave of megaphone.

C. Statistical Analysis

In order to further inspect our proposed SketchPart dataset in statistic, three distributions are provided on (i) number of predefined part labels per category, (ii) number of annotated parts per category in average and (iii) the variance between the above two, i.e., how often the predefined part labels would be used for annotation per category, as shown in Fig. 2(a), Fig. 2(b), and Fig. 2(c) respectively. Firstly, we can observe from Fig. 2(a) that most categories have more than two part labels, while there are still seven categories, including ‘arm’, ‘banana’, ‘boomerang’, ‘t-shirt’, ‘socks’, ‘scissors’ and ‘paper-clip’, have only one part due to they typically perform as a whole which no more subtle parts could be defined (see Fig. 3). Secondly, it can be found from Fig. 2(b) that most categories (about 57%) have more than 2 part annotations per sketch in average. There are about 34 categories that have less than 2 parts per sketch, this happens when a sketch with less parts drawn than the predefined labels in the corresponding class. Lastly, Fig. 2(c) is to further inspect the inconsistency between the annotated parts and predefined labels. We can see that most cases drop into the scope in less than one, which indicates a good consistency between the actual annotations and the available part labels per sketch. However, there exists rare categories which suffer severe inconsistency because the much larger variance within the class.

In overall, our proposed SketchPart dataset exhibit reasonable part partitions regard to the number and its variety of predefined semantic part labels, and the amount of annotations per sketch in average. Apart from sketch recognition, this dataset could serve as a perfect benchmark for sketch semantic segmentation, although which is not our focus in this work.

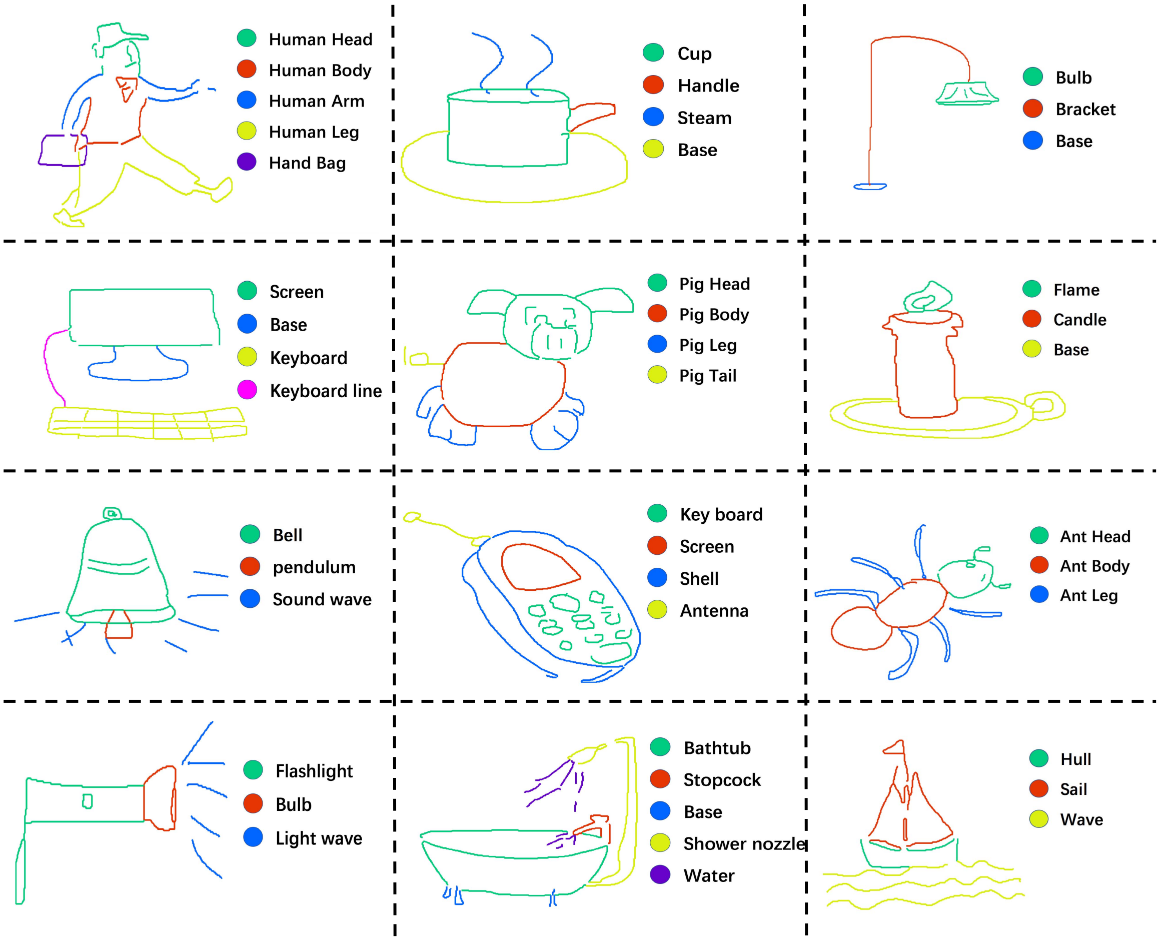


Fig. 1. Example sketches from the human-labeled sketch dataset. All the strokes in each sketch are manually labeled into groups of semantic parts.

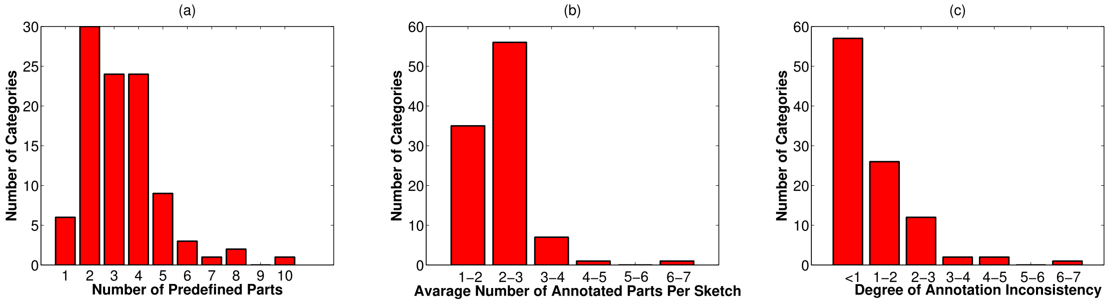


Fig. 2. Statistics of SketchPart dataset. **(a)** shows distribution on sketch categories over how many part labels are predefined per class. **(b)** indicates the number of annotated parts per sketch in average. **(c)** shows the degree of inconsistency over the average amount of annotated parts and the number of predefined parts.

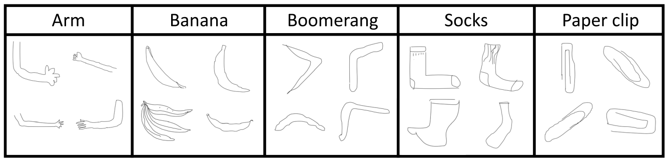


Fig. 3. Example sketch objects that can hardly be divided into parts.

IV. Co-Regularized Embedding Model

A. Problem Statement and Notations

For our problem of few-shot sketch recognition, we split the dataset into three sets: (i) A source set[1] which is composed of annotated parts of learned sketches. (ii) A support set which only consists of K labeled example sketches that without parts annotations (parts are unknown for a new sketch category before learning) for each of never-before-seen sketch categories. (iii) A target set is composed of a few new query sketches with the same class labels in support set. Assume the number of never-before-seen sketch categories is C, then the task is set to a C-way K-shot classification problem.

Fig. 5 shows the overview of our approach. Given a set of learned sketch with their corresponding parts available, we aim to learn basis from these parts to help with few-shot classification later on, the parts in the source set can be denoted as a matrix $\boldsymbol{A} \in \mathbb{R}^{d \times N}$, which representing N sketch parts in a d-dimensional embedding space. The target set is denoted as $\boldsymbol{B} \in \mathbb{R}^{d \times M}$ representing $M = C * K$

[1] Without using whole sketch images, we only use parts to form the source set.

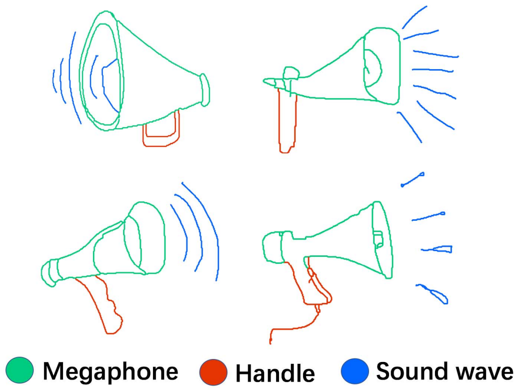


Fig. 4. Components of megaphone. For each predefined semantic part concept, there are diverse instances. We can see how humans draw sound wave.

examples in the same embedding space in the support set, in particular $\boldsymbol{B} = \{\boldsymbol{B_1}, \boldsymbol{B_2}, \ldots, \boldsymbol{B_j}, \ldots, \boldsymbol{B_C}\}$ where $\boldsymbol{B_j} = \{\boldsymbol{b}_j^1, \boldsymbol{b}_j^2, \ldots, \boldsymbol{b}_j^k, \ldots, \boldsymbol{b}_j^K\}$ represent the K examples for the j-th new category, and vector $\boldsymbol{b}_j^k \in \mathbb{R}^d$ is a example in this class. Therefore, given a query sketch y, two different sparse representations, i.e. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, can be obtained based on $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. Notably, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are learned jointly hence can perform for classification more efficient, which will be detailed in the following.

B. Modeling

1) Basis Discovery: The problem of sharable basis discovery is casted into a sparse coding problem. For a target/unseen category, sharable basis is discovered from source set by finding the non-zero entries of the sparse representation of its corresponding K example sketches. The intuition is that, the selected parts in source set which are able to perfectly reconstruct the example sketches, should be also qualified to reconstruct other sketches in the same target category. However, since sketch parts are typically strokes with large appearance variance which might lead to large reconstruction error in practice. In other words, the selected sharable parts from original source set are hardly optimal. To deal with it, we motivate to find a set of bases $\hat{\boldsymbol{A}} \in \mathbb{R}^{d \times n}$ by adopting an efficient dictionary learning strategy on the source set $\boldsymbol{A}$:

$$\begin{aligned} &\text{minimize}_{\hat{\boldsymbol{A}},\boldsymbol{S}} \|\boldsymbol{A} - \hat{\boldsymbol{A}}\boldsymbol{S}\|_F^2 \\ &s.t. \quad \sum_{i=1}^{d} \hat{A}_{i,j}^2 \leqslant c, \quad \forall j = 1, \ldots, n \end{aligned} \qquad (1)$$

where $\hat{A}_{i,j}$ is the i-th row and j-th column element of matrix $\hat{\boldsymbol{A}}$, $\boldsymbol{S}$ is the coefficient matrix and c is a constant. The goal is to represent parts in source set approximately as a weighted linear combination of a small number of bases specifying by n. The optimization problem can be addressed by iteratively optimizing with respect to $\hat{\boldsymbol{A}}$ and $\boldsymbol{S}$ while holding the other fixed. For learning the bases $\hat{\boldsymbol{A}} \in \mathbb{R}^{d \times n}$, the problem is a least squares problem with quadratic constrains which can be addressed very efficiently by a Lagrange dual approach [33]. Therefore, given the examples in support set $\boldsymbol{B_j} = \{\boldsymbol{b}_j^1, \boldsymbol{b}_j^2, \ldots, \boldsymbol{b}_j^k, \ldots, \boldsymbol{b}_j^K\}$ for target category j, a sparse representation $\boldsymbol{v}_j^k$ for $\boldsymbol{b}_j^k$ can be obtained by:

$$\min_{\boldsymbol{v}_j^k} \frac{1}{2}\|\boldsymbol{b}_j^k - \hat{\boldsymbol{A}}\boldsymbol{v}_j^k\|_2^2 + \sigma \left\|\boldsymbol{v}_j^k\right\|_1, \quad s.t. \quad v_{ji}^k \geqslant 0 \qquad (2)$$

where v_{ji}^k is the i-th entry of $\boldsymbol{v}_j^k$, hence $\boldsymbol{v_j} = \bigoplus_k \boldsymbol{v}_j^k$ ($k = \{1, 2 \ldots, K\}$) is given by element-wise summing up all the vectors $\boldsymbol{v}_j^k$, thus $\boldsymbol{v_j}$ is a n-dimensional vector whose non-zero entries indicate the relevance between the sketch parts basis in $\hat{\boldsymbol{A}}$ and the j-th target category. Intuitively, with the increasing number of observed example sketches, more basis for a target category can be discovered. And entries in $\boldsymbol{v_j}$ can be considered as the probabilities of the sketch parts basis being relevant to the j-th target category, if a normalization is further imposed.

2) Embedding Space Mapping by Co-Regularization: Here we describe how to utilize these discovered parts basis to encode query by our proposed co-regularized sparse representation, hence can benefit few-shot classification.

Our algorithm is developed based on the hypothesis that, a query should be well reconstructed by part basis discovered by the few-shot examples if belong to the same class. Otherwise, the reconstruction should be much worse if using improper part basis given by examples in any different class in the support set. On the other hand, our method is inspired by the process of how humans learn rich concepts (never-before-seen objects) from very limited data: humans normally firstly parse learned object into parts then generate new concepts from related concepts/parts. For example, one can generate/learn a novel object one-wheeled motorcycle by parsing and combining scooter and wheelbarrow as shown in Fig. 6.

Formally, given an unknown sketch $\boldsymbol{y}$, the discovered basis $\hat{\boldsymbol{A}}$ from source set, and the examples in support set (target dictionary) $\boldsymbol{B}$. We firstly measure the coarse relevance between $\boldsymbol{y}$ and each of the target categories, just like the process of picking scooter and wheelbarrow as the references or related objects. Then to utilize the corresponding part basis to reconstruct query $\boldsymbol{y}$, comparing to combine the relevant parts on scooter and wheelbarrow as a one-wheeled motorcycle. The formal step can be achieved by obtaining a sparse representation $\boldsymbol{\beta}$ according to the target dictionary $\boldsymbol{B}$:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\beta}\|_2^2 + \gamma \left\|\boldsymbol{\beta}\right\|_1, \quad s.t. \quad \beta_j^k \geqslant 0 \qquad (3)$$

where $\boldsymbol{\beta} = \{\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \ldots, \boldsymbol{\beta_j}, \ldots, \boldsymbol{\beta_C}\}$, and $\boldsymbol{\beta_j} = \{\beta_j^1, \ldots, \beta_j^k, \ldots, \beta_j^K\}$. We further reduce the dimension of $\boldsymbol{\beta}$ from $C \times K$ to C by summing up all the entries of vector $\boldsymbol{\beta_j}$ into a scalar $\beta_j = \beta_j^1 + \beta_j^2 + \cdots + \beta_j^K$, which indicates the coarse relevance between $\boldsymbol{y}$ and the j-th category in target set. Therefore, $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_j, \ldots, \beta_C\}$. Secondly, based on the relevance, the co-regularized sparse representation $\boldsymbol{\alpha}$ is obtained by the guidance of $\boldsymbol{\beta}$, to pick up the basis from the reference categories to represent $\boldsymbol{y}$. Formally, $\boldsymbol{\alpha}$ can be

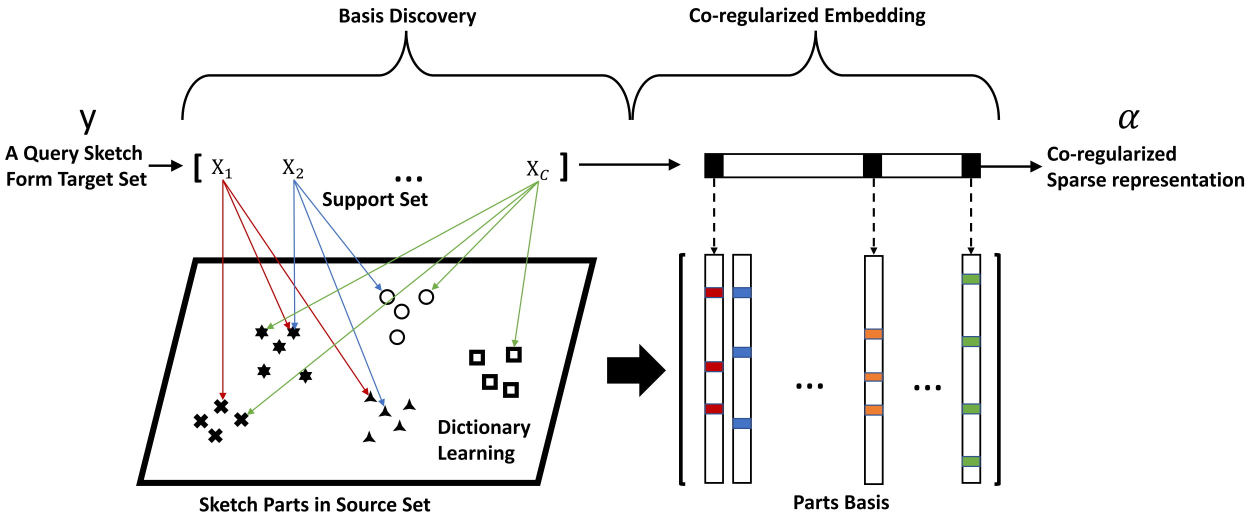


Fig. 5. Framework Overview. Parts in source sketch categories are exploited to construct a parts basis. Hence given a support set which contains just a few samples of never-before-seen categories, their most qualified basis can be discovered by sparse coding. And given an input query sketch, according to it's possible relevance to the samples in support set, our co-regularized embedding algorithm produces a sparse representation, thus to do classification by sparse reconstruction classifier. The co-regularized vector is used to fire the relevant parts basis (the black block indicates to use the corresponding basis set).

obtained by:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\boldsymbol{y}-\hat{\boldsymbol{A}}\boldsymbol{\alpha}\|_2^2+\sigma\,\|\boldsymbol{\alpha}\|_1-\frac{\lambda}{C}\langle \boldsymbol{V}^T\boldsymbol{\alpha},\boldsymbol{\beta}\rangle$$
$$s.t.\quad \alpha_i \geqslant 0, \beta_j \geqslant 0 \qquad (4)$$

where $\boldsymbol{V}=\{\boldsymbol{v_1},\boldsymbol{v_2},\dots,\boldsymbol{v_j},\dots,\boldsymbol{v_C}\}$ is constructed by Eq. (2), $\boldsymbol{v_j}$ indicates the related part basis to the j-th class in target set, and $\langle \boldsymbol{V}^T\boldsymbol{\alpha},\boldsymbol{\beta}\rangle=\sum_{j=1}^{C}(\langle \boldsymbol{v_j},\boldsymbol{\alpha}\rangle\times\beta_j)$. According to the role of $\boldsymbol{v_j}$ in Eq. 2, $\langle \boldsymbol{v_j},\boldsymbol{\alpha}\rangle$ indicates how strong the resulting sparse representation α is linked to the j-th target category, and the penalty $\langle \boldsymbol{V}^T\boldsymbol{\alpha},\boldsymbol{\beta}\rangle$ is to encourage the learning of $\boldsymbol{\alpha}$ such that the response on the entries relevant to the j-th target category (non-zero entries of $\boldsymbol{v_j}$) should agree with the response on the corresponding entries in $\boldsymbol{\beta}$. In other words, this is to constraint the sparse representation of $\boldsymbol{y}$, i.e. $\boldsymbol{\alpha}$, to encode with those basis that confirmed by samples in support set. For example, to represent "one-wheeled motorcycle", our approach guides to use basis of "scooter" and "wheelbarrow", because of their strong appearance relevance. The co-regularization process is the reason why we call our method co-regularized embedding. We also name the novel penalty $\langle \boldsymbol{V}^T\boldsymbol{\alpha},\boldsymbol{\beta}\rangle$ as 'co-regularization term', which controls the strength of the co-regularization, therefore the amount of knowledge transferred between the source and target sets.

We address the optimization of Criterion Eq. 4 in the next section by reformulating it as a quadratic program (QP) and further derive an equivalent linear complementary problem (LCP), such that an efficient principle pivoting algorithm can be used to solve the problem.

C. Optimization Algorithm

Here we describe the details on the optimization algorithm of our co-regularized embedding (CRE) model. To simplify the notations in Eq. 4, we set $g(\boldsymbol{\alpha})=\langle \boldsymbol{V}^T\boldsymbol{\alpha},\boldsymbol{\beta}\rangle$. We then re-formulate the problem in Eq. 4 as the following quadratic program[2]:

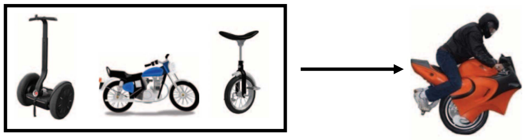


Fig. 6. Human can learn new concept from limited data by parsing learned objects and combining related ones to generate new object [12].

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}}\boldsymbol{\alpha}+(\sigma\boldsymbol{1}-\hat{\boldsymbol{A}}^T\boldsymbol{y})^T\boldsymbol{\alpha}-\frac{\lambda}{C}g(\boldsymbol{\alpha})$$
$$s.t.\quad \alpha_i \geqslant 0 \qquad (5)$$

Since $\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}}$ is a positive semidefinite matrix, this quadratic program in Eq. 5 is convex, where Karush-Kuhn-Tucker optimal conditions constitute the following monotone linear complementary problem [29]:

$$\boldsymbol{\delta}=\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}}\boldsymbol{\alpha}-\hat{\boldsymbol{A}}^T\boldsymbol{y}+\sigma\boldsymbol{1}-\frac{\lambda}{C}g'(\boldsymbol{\alpha})$$
$$s.t.\ \boldsymbol{\delta}\geqslant 0,\boldsymbol{\alpha}\geqslant 0,\boldsymbol{\alpha}^T\boldsymbol{\delta}=0 \qquad (6)$$

Here $g'(\boldsymbol{\alpha})\in\mathbb{R}^n$ is given by the differential of $g(\boldsymbol{\alpha})$ over $\boldsymbol{\alpha}$, and the i-th entry $g'(\boldsymbol{\alpha})_i=\beta_1 v_{1i}+\beta_2 v_{2i}+\dots+\beta_C v_{Ci}$. In our problem, the matrix $\hat{\boldsymbol{A}}^T\hat{\boldsymbol{A}}$ is always positive definite, so the convex problem in Eq. 5 and the monotone LCP in Eq. 6 thus have unique solutions for each vector $\boldsymbol{y}$.

Now we describe how a complementary solution can be obtained. Let F and G be two subsets of {1,…,N} such that $F\cup G=\{1,\dots,N\}$ and $F\cap G=\varnothing$. Then consider the following partition of the matrix $\hat{\boldsymbol{A}}$: $\hat{\boldsymbol{A}}=[\hat{\boldsymbol{A}}_F,\hat{\boldsymbol{A}}_G]$, where $\hat{\boldsymbol{A}}_F\in\mathbb{R}^{d\times|F|}$, $\hat{\boldsymbol{A}}_G\in\mathbb{R}^{d\times|G|}$, and $|F|$ and $|G|$ are the numbers of F and G, irespectively. Based on the partition

[2] Quite similar optimization problem and solution can be found in [29], the major difference is that we have an additional co-regularized term required to be solved together during optimization.

we reformulate Eq. 6 as follows:

$$\begin{bmatrix} \boldsymbol{\delta}_F \\ \boldsymbol{\delta}_G \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{A}}_F^T \hat{\boldsymbol{A}}_F & \hat{\boldsymbol{A}}_F^T \hat{\boldsymbol{A}}_G \\ \hat{\boldsymbol{A}}_G^T \hat{\boldsymbol{A}}_F & \hat{\boldsymbol{A}}_G^T \hat{\boldsymbol{A}} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_F \\ \boldsymbol{\alpha}_G \end{bmatrix} - \begin{bmatrix} \hat{\boldsymbol{A}}_F^T \boldsymbol{y} \\ \hat{\boldsymbol{A}}_G^T \boldsymbol{y} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\sigma}_F - \frac{\gamma}{C} g'(\boldsymbol{\alpha})_F \\ \boldsymbol{\sigma}_G - \frac{\gamma}{C} g'(\boldsymbol{\alpha})_G \end{bmatrix} \quad (7)$$

where $\boldsymbol{\alpha}_F, \boldsymbol{\delta}_F, \boldsymbol{\sigma}_F, g'(\boldsymbol{\alpha})_F \in \mathbb{R}^{|F|}$, $\boldsymbol{\alpha}_G, \boldsymbol{\delta}_G, \boldsymbol{\sigma}_G, g'(\boldsymbol{\alpha})_G \in \mathbb{R}^{|G|}$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha_F}, \boldsymbol{\alpha_G})$, and $\boldsymbol{\delta} = (\boldsymbol{\delta}_F, \boldsymbol{\delta}_G)$. A complementary basic solution is obtained by setting $\boldsymbol{\alpha}_G = 0$ and $\boldsymbol{\delta}_F = 0$ in Eq. 7, and we can compute the values of the basic variables $\boldsymbol{\alpha}_F$ and $\boldsymbol{\delta}_G$ by :

$$\min_{\boldsymbol{\alpha}_F \in \mathbb{R}^{|F|}} \frac{1}{2} \| \hat{\boldsymbol{A}}_F \boldsymbol{\alpha}_F - \boldsymbol{y} \|_2^2 + \sigma \sum_{i \in F} \alpha_i - \frac{\lambda}{C} < g'(\boldsymbol{\alpha})_F, \boldsymbol{\alpha}_F > \quad (8)$$

$$\boldsymbol{\delta}_G = \hat{\boldsymbol{A}}_G^T (\hat{\boldsymbol{A}}_F^T \boldsymbol{\alpha}_F - \boldsymbol{y}) + \boldsymbol{\sigma}_G - \frac{\lambda}{C} g'(\boldsymbol{\alpha})_G \quad (9)$$

Finally the optimal solution is given by setting $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_F, 0)$ and $\boldsymbol{\delta} = (0, \boldsymbol{\delta}_G)$. Please refer to [34] for more details.

D. Classifier

As in Eq. 4, we aim to reconstruct a test sketch y using basis/parts in source dictionary as well as possible, and parts belonging to the same class of $\boldsymbol{y}$ shall be expected to contribute the most during reconstruction. This is the basic idea of sparse representation classifier (SRC).Therefore, we design a class specific reconstruction classifier similar to the sparse classifier proposed by [25].

More specifically, for each target class j, let $\chi_j : \mathbb{R}^n \rightarrow \mathbb{R}^{n_j}$ be a function which selects the coefficients belonging to class j, i.e. $\chi_j(\boldsymbol{\alpha}) \in \mathbb{R}^{n_j}$ is a vector whose entries are the entries in $\boldsymbol{\alpha}$ corresponding to class j. Thus the unknown sketch $\boldsymbol{y}$ could be reconstructed as $\hat{\boldsymbol{y}}_j^{\boldsymbol{\alpha}} = \hat{\boldsymbol{A}}_j \chi_j(\boldsymbol{\alpha})$, which is reconstructed by only using the coefficients associated with class j. Similarly, the sparse representation $\boldsymbol{\beta}$ can be used for classification as well, where $\boldsymbol{y}$ can be reconstructed as $\hat{\boldsymbol{y}}_j^{\boldsymbol{\beta}} = \hat{\boldsymbol{A}}_j \chi_j(\boldsymbol{\beta})$.

To this end, $\boldsymbol{y}$ can be classified by assigning it to the class j corresponding to the minimal euclidean distance (Eu) between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}_j = \{\hat{\boldsymbol{y}}_j^{\boldsymbol{\alpha}}, \hat{\boldsymbol{y}}_j^{\boldsymbol{\beta}}\}$, which has shown to be suitable for many pattern recognition problems for matching patterns represented as features [35]. Therefore, our classifier is defined as:

$$\min_j r_j(\boldsymbol{y}, \hat{\boldsymbol{y}}_j) = \theta Eu(\boldsymbol{y}, \hat{\boldsymbol{y}}_j^{\boldsymbol{\alpha}}) + (1 - \theta) Eu(\boldsymbol{y}, \hat{\boldsymbol{y}}_j^{\boldsymbol{\beta}}) \quad (10)$$

where $\theta \in [0, 1]$ is a weight determines how much the final decision relies on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the optimal value of θ can be obtained by grid search in our case. By obtaining the minimal distance of $r_j(\boldsymbol{y}, \hat{\boldsymbol{y}}_j)$, the corresponding class label j is the output prediction.

V. Experiments

We evaluate the proposed co-regularized embedding (CRE) few-shot learning algorithm under a sketch recognition framework, and offer comparisons against six baseline methods, including a naive template matching (TM) method, a traditional learning method of bag-of-words feature trained support vector machine (Bow+SVM), sparse coding based sparse representation classifier (SC+SRC) and three state-of-the-art few-shot classification algorithms, which are all deep-based models: Prototype Networks [13], Matching Networks [14] and NTM meta-learning [15].

A. Experimental Settings

Data Split: – Among the 100 sketch categories labeled by annotators, 80 categories are randomly selected as the source, i.e. learned knowledge, and the rest 20 categories reserved as never-before-seen data for target. Then the target dataset is further split into a support set and a testing set, where a few (or only one for one-shot setting) sample sketches are randomly selected as support set and the rest for querying for each category.

We evaluate on six few-shot learning settings: *5-way 1-shot* where five categories are randomly selected from the target categories and only one sample sketch is provided for each, the resting 79 sketches are queries. *5-way 3-shot* and *5-way 5-shot* differ in the number of provided samples of each target categories. Similarly, *20-way 1-shot*, *20-way 3-shot* and *20-way 5-shot* test on 20 target categories with various amount of sample shots are offered.

Features: – Sketch-A-Net [3] performs even better than human on the sketch recognition task, hence it is be the best deep-feature extractor for sketch recognition to date. We re-train this network with all the sketches in source categories, and is employed to encode both parts of sketches and sketches themselves. We abandoned the information of stroke order during training Sketch-A-Net, Since the stroke orders are commonly absent when providing a pure sketch image, which is a much more common situation. Hence adopting an stroke-order-free version of the model makes our algorithm more applicable in most cases.

More specifically, $80 \times 80 = 6400$ sketches in the source set are used for training Sketch-A-Net, where we follow exactly the same network architecture proposed in [3], i.e., there are five convolutional layers followed by rectifier (ReLU) units, and the 1st, 2nd and 5th layers are equipped by max pooling (Maxpool). And a softmax loss is applied for training the network, in which Joint Bayesian Fusion strategy is not used in our case.

Given the trained Sketch-A-Net, it can be used to extract deep feature of both whole sketch ($\boldsymbol{y}$) and semantic part[3] (i.e., a column vector of matrix $\boldsymbol{A}$). More specifically, for any input after scaling it into a size of 256×256, following [3] a multi-view cropping strategy is applied firstly, which will crop each input 10 times (4 corner, 1 centre and flipped). Hence a forward pass of all 10 cropped inputs will result in a $250 \times 10 = 2500$ dimensional vector as its original feature, hence can be used to obtain a co-regularized sparse representation.

And importantly because of the large variability and redundancy of the original parts data mentioned in section IV-B.1, we further apply a efficient dictionary learning

[3] Although training on sketches, it turns out the network still can be applied for representing sketch parts.

algorithm to form a set of 512 common part basis. A similar practice is also used in [36] to obtain tokens.

B. Baseline Methods

1) Template Matching (TM): This is a naive while classical strategy for one/few-shot recognition. In our case, we use the given K samples of sketches per target class as template, where representation is given by the same Sketch-A-Net model. A class label is assigned by measuring the distance between the query sketch and each template. Specifically, the predicted class will be assigned with the corresponding class label of the template with minimal matching distance.

2) Bag-of-Words-Based Support Vector Machine (BoW+SVM): Follow the approach in [1], based on a Sketch-A-Net formed bag-of-words (BoW) as features to train SVM as classifiers, which is the most popular strategy for sketch recognition. In particular, all the sketches in source set are employed, hence totally $80 \times 80 = 6400$ sketches are utilized to form a 500 visual words vocabulary by K-means. Therefore, a 500 dimensional histogram of visual words can be constructed to represent a sketch, thus be able to train SVM classifiers, where the training data are the given sample sketches of each never-before-seen sketch categories and the testing data are the rest of the sketches in those never-before-seen classes.

3) Spare Coding Based Sparse Representation Classifier (SC+SRC): where all the samples from the support set are constructed as a dictionary, and the standard sparse coding (SC) algorithm is employed to produce sparse representation for an unknown sketch, followed by a sparse representation classifier (SRC) to assign a class label. And this is actually equivalent to set $\theta = 0$ in Eq. (10) in our proposed framework. This baseline method is used to measure the effectiveness of the proposed co-regularized term by comparison.

4) Prototypical Networks (Prototype Nets): where utilizes all the training data (6400 totally) from 80 source categories to learn a non-linear mapping of the input into an embedding space using a neural network. A new class's prototype is the mean of its support set in the embedding space. As a validation set is required for training the model of prototypical networks [13], and since we want to keep the same number of sketch classes (i.e., 80 classes) as training set for all the competitors for fairer comparison, 50 extra categories of sketches from TU-Berlin sketch dataset [1] are introduced as validation set for training prototypical networks. That is, there are more data, i.e. 130 categories (80 categories formed as training set and extra 50 categories as validation set), used for learning prototypical networks than our proposed model.

5) Matching Networks (Matching Nets): where learns a weighted nearest-neighbor classifier by using all the 6400 sketches training data set and the same validation data as in Prototype Nets. Importantly, the adopted sampled mini-batches method can mimic the few-shot classification task by showing only very low shots per category, just like how it will be tested when given a few samples of a new class. In addition, we follow the same augmentation strategy that can get a richer training data set with random rotations by multiples of 90 degrees for each sketch. The same to Prototype Nets, an additional 50 classes are used for validation both for 5-way and 20-way classification.

6) Sketch-A-Net (SAN) Serves as Backbone Network: for both Prototype Nets and Matching Nets, the backbone network is replaced by Sketch-A-Net, which results in two variant baseline methods: **Prototype Nets (SAN)** and **Matching Nets (SAN)**. This aims to testify if improvements could be achieved when varying the backbone network architecture by using a sketch-oriented one, i.e., Sketch-A-Net. Specifically, to facilitate a better learning of Prototype Nets and Matching Nets, an enforced data augmentation is used, i.e., a random rotation with degree ranged in $(-6,6)$ is performed for each input, then cropping it 10 times similar to [3]. In particular, as gradient vanishing was found for training Matching Nets and gradient exploding happened for learning Prototype Nets, if using the original architecture design of Sketch-A-Net, batch normalization (BN) is utilized before applying activation function ReLU for each conv layer in Sketch-A-Net. Again, the same data split is used to baseline methods, Prototype Nets and Matching Nets.

7) NTM Meta-Learning (NTM META): is a memory-augmented neural network (MANN [15]) that is capable of slowly learn an abstract for obtaining useful representations of raw sketch data via gradient descent, and is able to rapidly bind never-before-seen information given a few samples by an external memory module. Specifically, all the same 6400 sketches from source set are used for training, where the same Sketch-A-Net feature extractor is applied for representation of sketches.

C. Results and Discussions

We run our few-shot sketch recognition experiment 10 times by randomly sampling 80 source and 20 target categories each time. The average recognition accuracy is reported in Table I and Fig.7 and Fig.8 present example 5-way and 20-way classification results.

As can be seen from Table I, there are no major surprises in results: 5-way task is more easier than 20-way, and more samples help all models on recognizing sketches of never-before-seen categories. Specifically, our proposed CRE method basically outperforms all the baselines, and can achieve an recognition accuracy about 85% with only one sample is given for each of the 5 new sketch categories (5-way one-shot), and achieves over 90% given 2/4 more samples (5-way 3/5-shot). 20-way is obvious harder than 5-way, while our approach still can offer an around of 50% classification accuracy, which outperforms over all the other baseline methods with recognition rate at 50.82% (one-shot), 62.01% (three-shot) and 70.47% (five-shot) respectively. TM performs the worst that only obtain a 5-way one-shot accuracy of nearly 30%, and severely drops to 13.85% at 20-way one-shot. Matching Nets achieves the best comparing with the other two state-of-the-art few-shot learning algorithms (Prototype Nets and NTM META), while these methods is inferior to CRE especially when looking at setting of 20-way. It is worth noting that since there is no readily proposed

TABLE I

COMPARISON ON FEW-SHOT CLASSIFICATION RESULTS. "-" : NOT REPORTED

	5-way acc.			20-way acc.		
Model	1-shot	3-shot	5-shot	1-shot	3-shot	5-shot
TM	29.62%	38.85%	43.14%	13.86%	18.25%	20.75%
BoW+SVM	**85.06%**	87.27%	83.47%	48.42%	61.88%	63.53%
NTM META[15]	29.84%	56.80%	64.94%	-	-	-
Prototype Nets[13]	53.13%	64.41%	68.17%	26.28%	36.87%	41.49%
Prototype Nets (SAN)	38.07%	53.34%	56.13%	14.29%	25.41%	30.16%
Matching Nets[14]	57.38%	66.25%	69.26%	34.15%	41.37%	45.29%
Matching Nets (SAN)	36.72%	43.92%	48.56%	-	-	-
SC+SRC	82.03%	89.87%	91.46%	49.81%	60.39%	69.53%
CRE	84.81%	**92.21%**	**92.53%**	**50.82%**	**62.01%**	**70.47%**

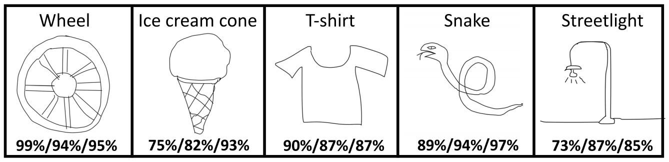


Fig. 7. Example 5-way sketch recognition. xx%/xx%/xx% denotes the achieved one/three/five-shot classification accuracy on five randomly selected sketch categories.

approach to train deep-based models by also using parts while can predict on whole sketches to our best knowledge, the comparisons conducted here are not completely fair that deep-based models train on whole sketches. However, on the other hand, it shows how parts could significantly benefit few-shot learning for sketch recognition. And to develop a part-based few-shot deep learning algorithms for sketch object recognition would be interesting which will be our future work. Interestingly, BoW+SVM and SC+SRC provide better results than Matching Nets, Prototype Nets and NTM META, BoW+SVM even achieves roughly equal to CRE for 5-way one-shot classification (85.06% vs 84.81%). In overall, our parts-powered model achieve a better performance over all the other competitors that work without sketch parts, which validates the importance of introducing sketch parts for few-shot learning of sketch categories. At last, it is interesting to find that, inferior performances are achieved when replacing the backbone network by Sketch-A-Net for both Prototype Nets and Matching Nets. The reason is that, very limited data is available for each meta learning epoch under the settings of few-shot learning, which might be not adaptable for the original design of network architecture of Sketch-A-Net. To apply Sketch-A-Net to Prototype Nets and Matching Nets, a heavy engineering or careful design of network architecture is probably demanded for performance boosting.

Fig.7 and Fig.8 shows an example category-level 5-way and 20-way classification results respectively. For most of the categories, there is a steady performance improvement when given more samples. Especially for some hard cases, there is an obvious improvement obtained, for example "banana" achieves a ten-fold improvement (from 3% to 32%), "Ear" offers four-fold improvement and "Megaphone" obtains a three-fold improvement when given 5 samples rather than one.

However, there are some categories with low recognition rate given 5 samples, such as "syringe"(24%) and "sword"(31%). In addition, rare cases occur that recognition rate doesn't gain from more samples and even perform worse, such as "Rifle" and "Snake". The reason is that sometimes the additional samples can not provide enough discriminative ability and might cause confusions as well. Fig.9 demonstrates the classification confusion matrix, where diagonal entries tell how many sketches are correctly classified, and non-diagonal entries shows the amount of sketches that are mis-classified and which categories were predicted to. For example, Fig.9(a) tells that most of the "syringes" are classified as "pen", and this error still exist when given 2 or 4 more samples as shown in Fig.9(b) and Fig.9(c). But in overall, the confusion entries (positive non-diagonal entries) become far less when more samples are used, which illustrates the classification ability of our approach.

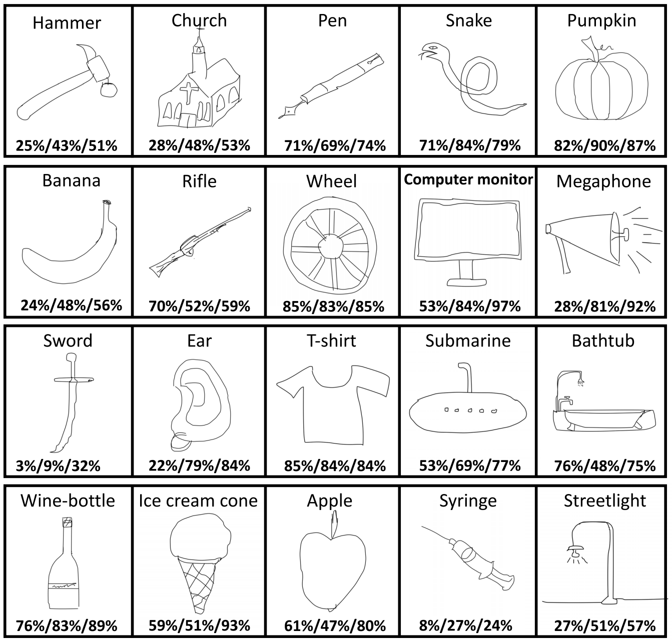
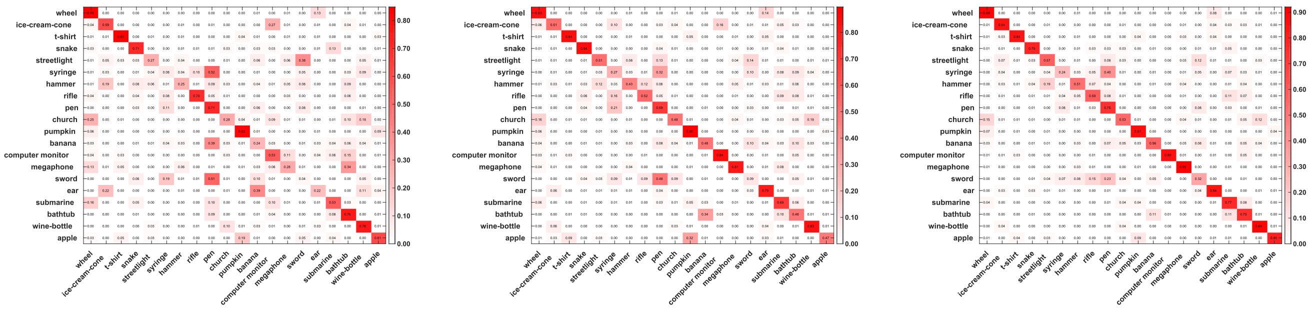


Fig. 8. Example 20-way sketch recognition results.

D. Further Analysis

1) Effect of Co-Regularization: The co-regularized term, $g(\alpha) =< V^T\alpha, \beta >$ in Eq. (4), is an important penalty in the proposed CRE framework. In particular, its weight λ controls how strong this constraint is to enforce the use of the corresponding basis to encode a sketch, e.g. in the case of small λ, a sketch would be encoded with the optimal parts by searching all the words in source part dictionary which leads to precise reconstruction. In contrast, in the case of



(a) 20-way one-shot (b) 20-way three-shot (c) 20-way five-shot

Fig. 9. Confusion matrix of recognition accuracy. Diagonal entries indicate classification accuracy for each class. Non-diagonal entries stands for how many sketches was incorrectly classified, and which categories they were classified to.

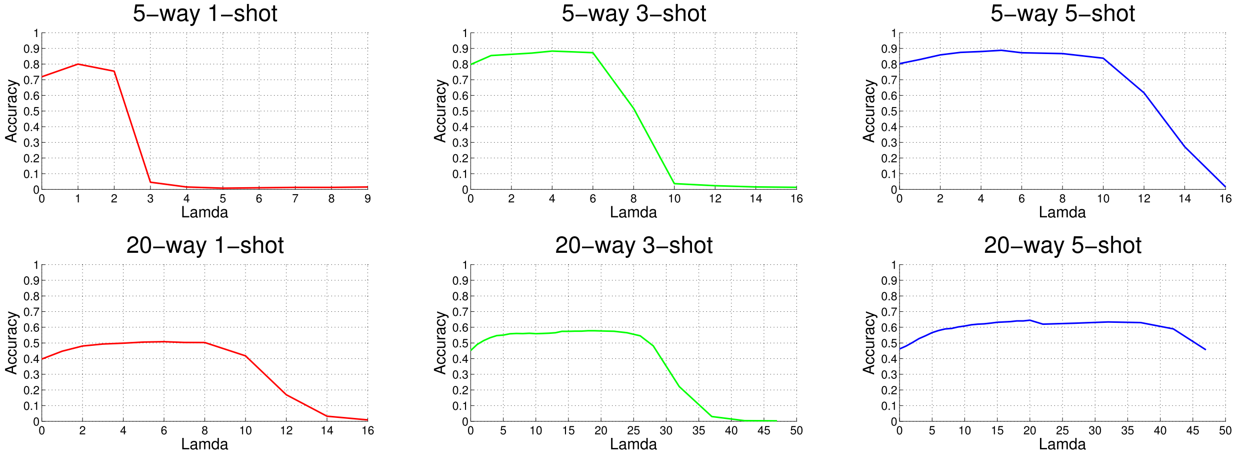


Fig. 10. Effect of co-regularized term. We increase the value of λ with the other parameters fixed, and evaluate the changes of recognition rate. For clarity, we only provide the recognition rates by using source set produced sparse representation α by setting $\theta = 1$ in Eq.(10). In this case, it may not achieve the best performances.

large λ, it would be encoded by a subset of predetermined parts, i.e., the parts of relevant categories given by Eq. (2) and Eq. (3).

Fig. 10 illustrates how overall the recognition rate changes while increasing the value of λ, with the other parameters fixed. It can be clearly observed from all the cases that there is a steep climb in recognition rate before the optimal value of λ is reached. Notably, the classification performance improvement lasts "more longer" in case of 20-way than it in 5-way when increasing λ, while this is actually because of the normalization factor C in Eq. (4) which is different in two cases. Afterwards, performance drops steadily while approaching 0%. Such a reduction of recognition rate reflects the trade-off between co-regularized term and the regression term, $\|y - \hat{A}\alpha\|_2^2$, in Eq. (4). That is, too large a weight on co-regularized term will make the regression problem ill-conditioned that consequently impacts the overall classification accuracy. Note that it becomes the standard sparse coding problem when removing the guidance term in Eq. (4), which is also equivalent to setting $\lambda = 0$.

2) How Sample Sketch Matters: Experimental results illustrated that the number of the given sample sketches plays a key role in few-shot sketch classification. However, very few previous work shed light on studying the influence of the quality of the given samples. In other words, the question we attempt to answer here is, what kind of sample sketches are the most efficient ones for learning the new category.

To gain some insights into how different sample sketches influence the classification accuracy, we randomly choose different sets of sample sketches, and measure the recognition results. In particular, to better understand how different sample sketch matters, we visualize the layout of sketches of streetlight by applying t-distributed stochastic neighbor embedding (t-SNE) with their features (i.e. the 2500 dimensional vector given by Sketch-A-Net), thus to measure their corresponding 5-way one-shot recognition accuracy given different one-shot samples. Interestingly, we can discover from Fig.11 that the most representative sketches are these laid in the middle space (green boxes), which are the normal or more general ones, hence benefit one-shot learning. In contrast, the ones laid on the boarder are some very unique ones which lead to inferior performances (red boxes).

3) Influence of Bases Number: We apply a dictionary learning approach to discover the common part bases, hence be able to reduce the redundancy of parts source set as well. To identify how the amount of bases number influences recognition accuracy, we vary the settings of bases number to 256, 512 and 768 with other parameters fixed. And we measure the 20-way 1/3/5-shot sketch recognition accuracy, see Table II. It turns out that the final recognition accuracy is

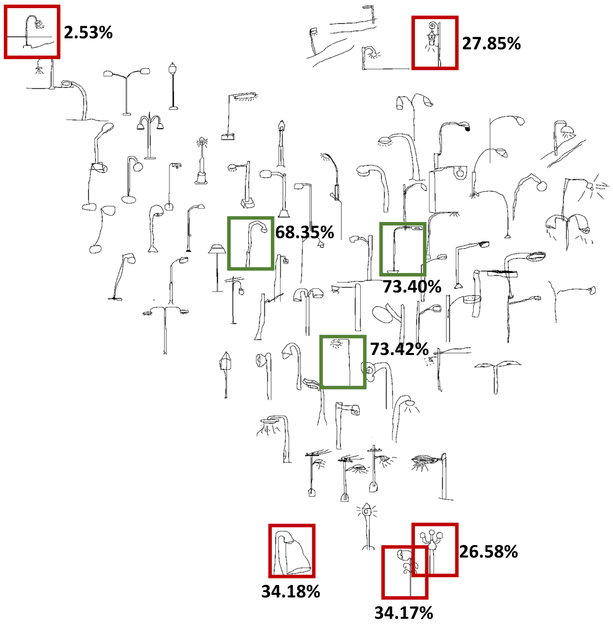


Fig. 11. Layout of streetlight, and its 5-way one-shot classification rate by using different selected samples. The layout of human free-hand drawn streetlight is made by applying t-SNE, which is a dimensionality reduction technique where t-SNE essentially offers a mapping of distances in high-dimensional space to distances in low-dimensional space such that the 2D layout still preserve overall global distances. Streetlights laid on boarder (in red boxes) achieve worse one-shot classification accuracy than the ones laid in middle (in green boxes).

TABLE II

COMPARISON RESULTS ON 20-WAY RECOGNITION RATE BY VARYING BASES NUMBER

	20-way acc.		
Bases number	1-shot	3-shot	5-shot
256	49.62%	60.58%	69.73%
512	50.82%	62.01%	70.47%
768	50.57%	61.56%	70.40%

not very sensitive to the bases number, and the optimal choice is 512 in our case.

VI. CONCLUSION

We proposed to deal with the problem of few-shot classification of sketch categories via a novel co-regularized sparse coding framework. We contribute a novel part level human sketch dataset of 8,000 sketches over 100 categories, and demonstrated how shared sketch parts basis can be used for quick learning of never-before-seen sketch categories. That is, based on the learned parts bank, the sharable parts/basis can be discovered by sparse coding and employed for representing new sketch data, which a co-regularization algorithm is developed to enforce the resulting sparse representation agree with the relevant sample sketches in the support set. The experimental results on the proposed dataset shows an obvious improvement of few-shot sketch classification over baseline methods even involving three DNN-based approach, which demonstrates the usefulness of parts.

VII. FUTURE WORK

Although a deep feature based co-regularized sparse representation shows compelling results on few-shot learning of sketch categories and can even beat deep models, we are still interested to find out if an end-to-end network can be trained to achive better results. As the first study on sketch-oriented few-shot learning, we chose not to adopt a deep approach so to gain more insights.

In addition, the *SketchParts* dataset provides opportunities to learn how humans sketch basic sketch elements, which should be very helpful on learning new categories. For example, are there more subtle sketch tokens exist to be able to represent new sketch more efficiently, hence to further benefit few-shot learning? Moreover, there are some interesting phenomenons, such as water, sound and light can be drawn with very similar strokes, how to learn such prior knowledge and thus use it for few-shot learning remains open.

REFERENCES

[1] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–10, Jul. 2012.

[2] R. Hu and J. Collomosse, "A performance evaluation of gradient field HOG descriptor for sketch based image retrieval," *Comput. Vis. Image Understand.*, vol. 117, no. 7, pp. 790–806, Jul. 2013.

[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," *Int. J. Comput. Vis.*, vol. 122, no. 3, pp. 411–425, May 2017.

[4] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," *IEEE Trans. Visual. Comput. Graph.*, vol. 17, no. 11, pp. 1624–1636, Nov. 2011.

[5] 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 *Proc. ICCV*, 2017.

[6] K. Li, K. Pang, Y.-Z. Song, T. M. Hospedales, T. Xiang, and H. Zhang, "Synergistic instance-level subspace alignment for fine-grained sketch-based image retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5908–5921, Dec. 2017.

[7] P. Xu *et al.*, "Sketchmate: Deep hashing for million-scale human sketch retrieval," in *Proc. CVPR*, 2018.

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

[9] R. Hu, M. Barnard, and J. Collomosse, "Gradient field descriptor for sketch based retrieval and localization," in *Proc. ICIP*, 2010, pp. 1025–1028.

[10] Y. Li, Y.-Z. Song, and S. Gong, "Sketch recognition by ensemble matching of structured features," in *Proc. BMVC*, 2013.

[11] X. Cao, H. Zhang, S. Liu, X. Guo, and L. Lin, "Sym-fish: A symmetry-aware flip invariant sketch histogram shape descriptor," in *Proc. ICCV*, 2013, pp. 313–320.

[12] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, Dec. 2015.

[13] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. NIPS*, 2017, pp. 4080–4090.

[14] O. Vinyals *et al.*, "Matching networks for one shot learning," in *Proc. NIPS*, 2016, pp. 3630–3638.

[15] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. ICML*, 2016, pp. 1842–1850.

[16] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. ICML*, 2017.

[17] L. Kaiser, O. Nachum, A. Roy, and S. Bengio, "Learning to remember rare events," in *Proc. ICRL*, 2017.

[18] Y. Yu, Z. Ji, Y. Fu, J. Guo, Y. Pang, and Z. M. Zhang, "Stacked semantics-guided attention model for fine-grained zero-shot learning," in *Proc. NeurIPS*, 2018.

[19] B. N. Oreshkin, P. R. López, and A. Lacoste, "TADAM: Task dependent adaptive metric for improved few-shot learning," in *Proc. NeurIPS*, 2018.

[20] S. Liu, M. Long, J. Wang, and M. I. Jordan, "Generalized zero-shot learning with deep calibration network," in *Proc. NeurIPS*, 2018.

[21] B. Zhao, X. Sun, Y. Fu, Y. Yao, and Y. Wang, "Msplit LBI: Realizing feature selection and dense estimation simultaneously in few-shot and zero-shot learning," in *Proc. ICML*, 2018.

[22] B. Lake, R. Salakhutdinov, J. Gross, and J. Tenenbaum, “One shot learning of simple visual concepts,” in *Proc. Annu. Meeting Cognit. Sci. Soc.*, 2011.

[23] D. Ha and D. Eck, “A neural representation of sketch drawings,” 2017, *arXiv:1704.03477*. [Online]. Available: https://arxiv.org/abs/1704.03477

[24] L. Fei-Fei, R. Fergus, and P. Perona, “One-shot learning of object categories,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.

[25] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[26] S. Gao, I. W.-H. Tsang, and L.-T. Chia, “Laplacian sparse coding, hypergraph laplacian sparse coding, and applications,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 92–104, Jan. 2013.

[27] M. J. Gangeh, A. Ghodsi, and M. S. Kamel, “Kernelized supervised dictionary learning,” *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4753–4767, Oct. 2013.

[28] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong, “Locality-constrained linear coding for image classification,” in *Proc. CVPR*, 2010, pp. 3360–3367.

[29] R. He, W.-S. Zheng, and B.-G. Hu, “Maximum correntropy criterion for robust face recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1561–1576, Aug. 2011.

[30] S. Zhang, H. Yao, X. Sun, and X. Lu, “Sparse coding based visual tracking: Review and experimental comparison,” *Pattern Recognit.*, vol. 46, no. 7, pp. 1772–1788, Jul. 2013.

[31] Y. Qi, W.-S. Zheng, T. Xiang, Y.-Z. Song, H. Zhang, and J. Guo, “One-shot learning of sketch categories with co-regularized sparse coding,” in *Proc. ISVC*, 2014.

[32] R. G. Schneider and T. Tuytelaars, “Example-based sketch segmentation and labeling using CRFs,” *ACM Trans. Graph. (TOG)*, vol. 35, p. 151, Jul. 2016.

[33] H. Lee, A. Battle, R. Raina, and A. Y. Ng, “Efficient sparse coding algorithms,” in *Proc. NIPS*, 2007.

[34] L. F. Portugal, J. J. Júdice, and L. N. Vicente, “A comparison of block pivoting and interior-point algorithms for linear least squares problems with nonnegative variables,” *Math. Comp.*, vol. 63, no. 208, p. 625, 1994.

[35] K. Grauman and T. Darrell, “Fast contour matching using approximate earth mover’s distance,” in *Proc. CVPR*, vol. 1, 2004, pp. 220–227.

[36] J. J. Lim, C. L. Zitnick, and P. Dollár, “Sketch tokens: A learned mid-level representation for contour and object detection,” in *Proc. CVPR*, 2013, pp. 3158–3165.

Yonggang Qi (Member, IEEE) received the Ph.D. degree in signal processing from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2015, the Ph.D. degree from SketchX Research Laboratory, Queen Mary University of London (QMUL), and the Ph.D. degree from Aalborg University, Denmark, in 2013. He has served as a Visiting Researcher with Sun Yat-sen University, China, in 2014. He is an Assistant Professor (lecturer) with the BUPT. His research interests include perceptual grouping, sketch-based tasks evolving, sketch-based image retrieval (SBIR), sketch recognition, sketch generation, and language-based sketch understanding.

Yi-Zhe Song (Senior Member, IEEE) received the bachelor’s degree from the University of Bath in 2003, the M.Sc. degree from the University of Cambridge in 2004, and the Ph.D. degree in computer vision and machine learning from the University of Bath in 2008. He is a Reader of computer vision and machine learning with the Centre for Vision Speech and Signal Processing (CVSSP), the largest academic research centre for artificial intelligence with approximately 200 researchers, University of Surrey, U.K. Previously, he was a Senior Lecturer with the Queen Mary University of London and a Research and Teaching Fellow with the University of Bath. He is a fellow of the Higher Education Academy. He is a Full Member of the review college of the Engineering and Physical Sciences Research Council (EPSRC), a main agency for funding research in engineering and the physical sciences. He received the Best Dissertation Award from the University of Cambridge. He serves as an expert reviewer for the Czech National Science Foundation.