Learning Structural Representations via Dynamic Object Landmarks Discovery for Sketch Recognition and Retrieval

Hua Zhang, Peng She, Yong Liu, Jianhou Gan, Xiaochun Cao, *Senior Member, IEEE*, and Hassan Foroosh, *Senior Member, IEEE*

***Abstract*— State-of-the-art methods on sketch classification and retrieval are based on deep convolutional neural network to learn representations. Although deep neural networks have the ability to model images with hierarchical representations by convolution kernels, they cannot automatically extract the structural representations of object categories in a human-perceptible way. Furthermore, sketch images usually have large-scale visual variations caused by the styles of drawing or viewpoints, which make it difficult to develop generalized representations using the fixed computational mode of convolutional kernel. In this paper, our aim is to address the problem of fixed computational mode in feature extraction process without extra supervision. We propose a novel architecture to dynamically discover the object landmarks and learn the discriminative structural representations. Our model is composed of two components: a representative landmark discovering module that localizes the key points on the object and a category-aware representation learning module that develops the category-specific features. Specifically, we develop a structure-aware offset layer to dynamically localize the representative landmarks, which is optimized based on the category labels without extra supervision. After that, a diversity branch is introduced to extract the global discriminative features for each category. Finally, we employ a multi-task loss function to develop an end-to-end trainable architecture. At testing time, we fuse all the predictions with different number of landmarks to achieve the final results. Through extensive experiments, we compare our model with several state-of-the-art methods on two challenging datasets, TU-Berlin and Sketchy, for sketch classification and retrieval, and the experimental results demonstrate the effectiveness of our proposed model.**

***Index Terms*— Sketch based image classification and retrieval, dynamic landmarks discovery, structural feature representation.**

Manuscript received June 6, 2018; revised January 26, 2019 and March 19, 2019; accepted March 24, 2019. Date of publication April 19, 2019; date of current version July 16, 2019. This work was supported in part by the National Key R&D Program of China under Grant 2016YFC0801002, in part by the National Natural Science Foundation of China under Grant 61602464, Grant U1636214, Grant U1803264, and Grant 61861166002, in part by the Beijing Natural Science Foundation under Grant KZ201910005007, and in part by the CCF-Tencent Open Research Fund. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Aydin Alatan. *(Corresponding author: Xiaochun Cao.)*

H. Zhang, P. She, Y. Liu, and X. Cao are with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: zhanghua@iie.ac.cn; julian.pshe@gmail.com; liuyong@iie.ac.cn; caoxiaochun@iie.ac.cn).

J. Gan is with the Key Laboratory of Educational Informatization for Nationalities, Yunnan Normal University, Ministry of Education, Kunming 650031, China (e-mail: ganjh@ynnu.edu.cn).

H. Foroosh is with the Department of Computer Science, University of Central Florida, Orlando, FL 32816 USA (e-mail: foroosh@cs.ucf.edu).

Digital Object Identifier 10.1109/TIP.2019.2910398

I. Introduction

FREE-hand sketch could be seen as one of the intuitive ways to reflect the goal of users without much effort, *e.g.* Sketch2photo [1]. With the development of intelligent mobile terminals, it is now possible to draw sketches that make sketch-based recognition and retrieval be a trending research topic in the fields of graphics and computer vision. Various interesting applications exist, including sketch-based image retrieval [2]–[7], human-computer interaction [8], face verification [9]–[11] and other relevant works such as [12]–[14]. Despite the significant progress of sketch classification and retrieval based on deep neural networks [10], [15]–[17] in recent years, this problem is still a challenging task due to the visual variations of sketch images.

Extensive efforts have been dedicated to solve the problem by learning discriminative representations [18]–[20], which are learned by directly treating the sketch images as the natural images to train the deep neural networks. However, sketch images show more visual variations than the natural images, which causes the learned model degenerated at test time. Although some works [21], [22] have been proposed to handle the visual variations, there are still several challenges influencing the performance of existing methods. First, the styles of sketches are diverse as shown in Fig. 1. Users draw sketch images based on their personalized preferences without any unified reference templates, which would generate a large variation of sketch appearances. Second, existing deep models would generate the ambiguous feature representations due to the large scale of visual variations. The feed-forward convolutional neural network can only capture the specific structure patterns due to the fixed computational mode of convolution kernels [23]–[27]. The fixed mode refers that we use the convolution kernel samples the input feature map at fixed locations and the region-of-interest pooling layer divides a region into fixed spatial bins. Several works have been proposed to deal with abovementioned fixed spatial computational mode of CNN by estimating the transformation model for feature alignment. For examples, STN [23], [24] aimed to achieve geometric invariance by mapping features through a global transformation. While Rocco *et al.* [26], [27] proposed a CNN architecture for estimating a geometric alignment model. However, existing models focus on estimating the

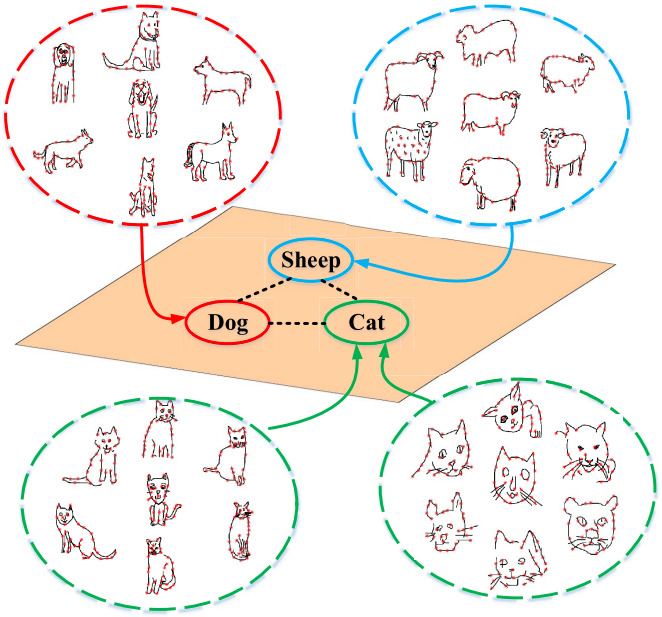


Fig. 1. Learning structural feature representations for sketch images via discovering the representative landmarks. Utilizing the discovered representative landmarks (red points) can jointly minimize intra-class discrepancy and maximize inter-category distance, and thus results in more discriminative feature representations.

globally-varying geometric fields, thus leading to limited performance in dealing with locally-varying geometric variations. Thus, existing models cannot be directly transferred to the novel style of sketches, especially for those not seen before. Last but not least, the receptive field of each convolutional kernel is localization, which is developed with a regular sampling grid on feature maps. This probably induces some loss in the feature discriminability. Several related methods [28]–[31] have proposed to use multiple size of receptive fields to improve the discrimination of representations.

To overcome the above-mentioned limitations and learn the robust feature representation model, we introduce a more efficient strategy based on the following observations: For the sketch images, there exist the representative structures for distinct categories, which make the representations robust for the visual variations. Furthermore, not all the whole sketch image can provide the category-specific features, only a set of sparse regions contain the useful appearances for recognition. Therefore, to deal with the limitations, we should consider these two aspects at the same time.

This paper aims to simultaneously address the problem of discovering representative landmarks and learning the discriminative feature representations for sketch images. In particular, we treat representative landmark discovery as an intermediate step for feature extraction. In the meanwhile, we also introduce additional regularization to capture the global feature representation and to prevent the representative landmarks from encoding irrelevant information. To that end, we develop a novel weakly-supervised deep architecture named landmarks-aware network as presented in Fig. 2. Our model is composed of two modules, representative landmarks discovering module and category-aware representation learning module. Given the training samples, we first feed them into the front-end CNN (ResNet-101 [32]), and then the last convolution layer of the network is extracted for landmarks localization. To discover the representative landmarks, we first sampling along the edges of sketch images, and then develop an offset layer at the top of extracted layers as shown in Fig. 2 (c). This module consists of landmark localization and pooling operations, which are optimized with category labels without extra supervision. Although the discovered landmarks can capture the structure-invariant feature representations, it may confuse the structures between categories because of the visual variations. To improve the feature discrimination, we propose to learn the category-specific features via introducing the diversity regularization term as shown in Fig. 2 (d). In this module, we generate the class-aware feature maps with 3×3 convolution kernels. In the step of optimization, besides the softmax loss based on the category labels, we also require the filters and their responses to be orthogonal. Furthermore, these two components are implemented with feed-forward networks, therefore the network is an end-to-end trainable architecture. To train our proposed network, we only need the category label of the inputs. At test time as shown in Fig. 2(e), all the prediction scores are fused to achieve the final results. Finally, extensive experiments are conducted for the task of sketch classification and sketch based image retrieval on two challenging datasets. In particular, we evaluate sketch classification on TU-Berlin dataset [33], which only contains sketch images with weakly supervised category label annotations. While on the task of sketch based image retrieval, we develop the experiment on Sketchy dataset [34], which is composed of real and sketch images with the category label annotations. Experimental results show that our proposed method can significantly improve the performance of sketch classification and sketch based image retrieval, which demonstrates the effectiveness and efficiently of our method.

The contributions of our proposed method can be summarized as follows: (i) We first propose a novel sketch representation learning method via dynamical landmarks discovery to address the problem of sketch visual variations. which is an end-to-end trainable architecture. (ii) To achieve the discriminative representation, we introduce the diversity regularization term to learn the category-specific feature representations. (iii) Our method boosts the benchmark of sketch classification, achieving the state-of-the-art accuracy in terms of classification evaluation metrics.

II. Related Work

In this section, we review the most recent related work on sketch classification and retrieval. The generative and discriminative power of deep features have been used to build deep generative shape models [35], [36]. In [36], the authors propose to use Deep Belief Nets (DBN) to generate the hand-written digits, which has achieved notable performance. While Eslami *et al.* [35] develop an object shape model method by using the Boltzmann machine. In [37], a neural network is trained to generate shape descriptors that lie close to a vector representation of the shape class for sketch-based

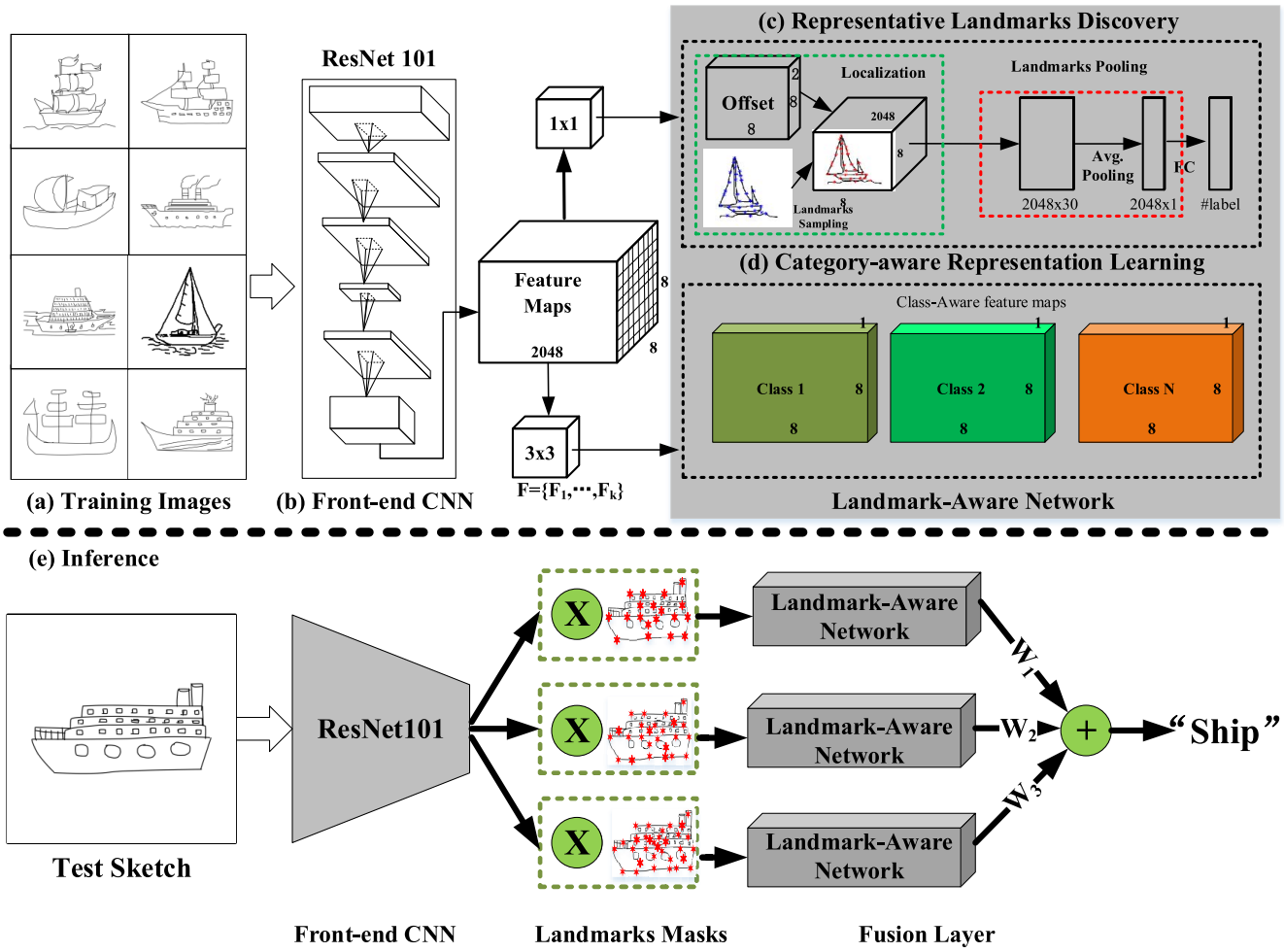


Fig. 2. Our proposed object landmarks discovery architecture for sketch classification and retrieval. First, images are fed into the ResNet101, and then we extract the last convolutional feature maps whose size is 2048×8×8. After that, we add a 1 × 1 convolution kernel to discover the landmarks and learn the discriminative feature representation (top-right branch). While a set of 3 × 3 convolution kernel is used to improve the discrimination of representation with diversity loss function (bottom-right branch).

shape retrieval via embedding the word information. Tolias and Chum [17] introduce the asymmetric feature maps to evaluate multiple kernels between the query and database entries on sketch based retrieval. This method could provide the query localization in the retrieved images. Although these methods achieve the discriminative feature representation, they do not consider the large visual variations of sketch images especially for new styles of test sketches.

There has also been some work on retrieving 3D shapes with sketches. Different from the traditional sketch-based image retrieval, a large visual discrepancy between sketches and 3D shapes limits the performance of retrieval. Xie *et al.* [19] propose to project the 3D shapes into 2D, and then compute the Wasserstein barycenters of multiple projections to form a barycentric representation.Wu *et al.* [38] propose a novel shape representation named 3D shapeNets, which focuses on modeling the 3D shapes. In this method [38], the authors propose to recognize the object from the depth images and then construct a 3D shapenets to capture the structure of 3D shapes. Traditional methods on retrieving 3D models from sketches are based on two stage view selection and matching. For instance, Wang *et al.* [39] propose a two siamese convolutional neural networks to compute the cross-domain similarities. In [14], a novel deep correlated holistic metric learning (DCHML) method is introduced to mitigate the discrepancy between sketch and 3D shape domains, which learns two deep nonlinear transformations to map features from both domains into a new feature space. Our proposed method can solve the problem of variations by dynamically selecting the representative landmarks.

To solve the problem of limited training samples, Zhang *et al.* [22] propose a novel deep convolutional neural network termed as SketchNet which incorporates the web images as the reference for discriminative feature learning. Yu et al. [40] propose two data augmentation strategies for exploiting the discriminative sketch structures. Similarly, Li *et al.* [16] propose a novel binary coding method, named deep sketch hashing for sketch based image retrieval. The learned Deep Sketch Hashing (DSH) codes can both effectively solve the geometric distortion between different domains and efficiently reduce the distance computation. Zhang *et al.* [41] propose a Generative

Domain-migration Hashing (GDH) approach, which can preserve the domain-invariant information between sketches and real images. GDH aims to learn the hash mapping by using an adversarial loss and the cycle consistency loss. Different from DSH [16] and GDH [41], our proposed method focuses on learning a robust model for sketch variations, which can be used for sketch classification and retrieval.

For sketch based image retrieval, metric learning is usually introduced to bridge the domain gap. Radenovic *et al.* [42] introduce the deep shape matching, which first extracts the edge map by edge detector from real images, and then the edge maps are feed to a neural network for extracting the global feature representations. Different from [42], our method conduct the feature representations by introducing the landmarks and diversity feature map, which makes our model robust to sketch visual variations. In [43], the authors proposed an attention-ensemble framework to encourage diversity in feature embeddings. While in [44], they aim to improve the robustness of feature embeddings by exploiting the independence within ensembles. To exploit the diversity, the authors introduce online boosting to build our metric ensemble and different loss functions. Different from [43] and [44], our model is proposed to conduct the diversity feature maps by requiring the filters and maps to be orthogonality instead of learning the metric mappings.

There are also some related work on landmark learning [45]–[47] and deformable region of interest (ROI) pooling [48], [49]. In [45], the authors propose an end-to-end neural network for discriminative feature points learning with three loss functions. Furthermore, they propose a part orthogonality constraint in the step of learning feature representation, which is similar with our diversity regularization. The main difference with our work is that our method can dynamically localize the key points while their method focuses on the regions which would be sensitive for the object variations. Dai *et al.* [48] propose to learn the deformable convolution kernel for object detection and semantic segmentations. Since typically sketch images contain large areas of white spaces, their method cannot capture enough texture information to compute offsets. Our proposed method, however, focuses on the edges of the sketch, which would be helpful to extract the landmarks.

III. Method

In this section, we introduce our landmark-aware network for sketch based image classification and retrieval. Our method consists of two modules: representative landmarks discovery and category-aware representation learning. Different from traditional methods on extracting the landmarks [46], [47], our framework first uses a uniform sampling method to extract the candidate key points along the edges of sketch images. For each candidate key point, we compute the offset by the proposed offset layer to refine the locations for discovering the representative landmarks. After the landmarks are determined, we use a novel landmark pooling operation to obtain the feature representations of sketch images. Furthermore, to further improve the discrimination of representations, we introduce a diversity regularization [45] to learn the sketch category-specific features. At test time, a fusion layer is proposed to compute the final predictions.

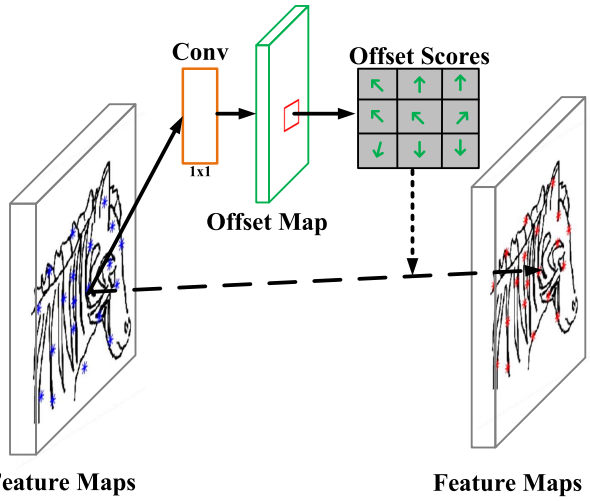


Fig. 3. Illustration of representative landmarks discovery via offset learning. We firstly uniformly sample the key points along the object edge (blue points in the left feature map), and then we refine the location of key points by adding the computed offsets (red points in the right feature map).

As shown in Fig. 2 (b), the output of the last convolutional layer is a tensor $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$, where H and W denote the height and width of the feature map and C indicates the number of channels. Thus, we define the representation at each spatial location (u, v) as $d_{uv} = (x_{uv1}, \ldots, x_{uvC}) \in \mathbb{R}^C$. The responses of the last convolution layer are 3D, while our proposed landmarks localization modules operate on the 2D spatial domain. Thus, we do the same operation on different channel dimensions. For notation clarity, we describe our method in 2D and it is straightforward to extend to 3D.

A. Representative Landmarks Discovery

Landmarks Localization: Given the observed sketch image $\mathbf{I}$, the objective is to estimate the representative landmarks and achieve the discriminative feature representations. To do that, we propose a novel approach for landmark localization, which is composed of two steps. First, a sketch image is represented by a discrete set of points sampled from the contours on the object, denoted as $P = \{p_1, \ldots, p_U\}$, where U is the total number of candidate key points. Then, given the number of landmarks N, we extract the candidate points with roughly uniform spacing, which is denoted as $l^o = \{l^o_1, \ldots, l^o_N\}$, where N is the number of extracted key points as shown in Fig. 3. Since these candidate points l^o typically will not correspond to the representative landmarks, we need to refine the localization of each candidate landmark for discovering the representative landmarks. Here, we should note that the sampling order is not critical to obtain good initial landmarks. This is because we treat all the candidate landmarks equally, which are uniformly distributed along the edges of the object. To address the rotation of objects, we propose to augment the training set by rotating the training samples.

Instead of computing the offsets in the image space, we propose to seek the offsets on the feature maps, which need to

project the landmarks from the image space to the feature maps. To that end, we first compute the ratios between the original image and feature maps along the width and height directions. Then, we project these locations to the feature maps by multiplying the ratios along two directions. To guarantee the projected locations are integers, we use the operation of ceil on the projected locations, and the localizations of landmarks on the feature maps are denoted as $l = \{l_1, \ldots, l_N\}$. This is based on the observation that there exists a linear mapping between the original image and feature maps, which has been demonstrated in several works [32], [46], [47].

After projecting the landmarks on the last convolutional feature maps $\mathbf{x}$, we apply a 1×1 convolution layer on $\mathbf{x}$ to compute the offsets for each landmark as shown in Fig. 3. The size of the offset map is $H \times W \times 2N$, where the channel dimensions are the 2D offsets along the width and height directions for all the landmarks. Then, we define the offsets of candidate landmarks as Δl, and the offset for each candidate is $\Delta l = \{\Delta l_1, \ldots, \Delta l_N\}$. Based on the novel localizations for the landmarks, we need to extract the feature representations of the computed representative landmarks.

Since the offsets would be fractional, the bilinear interpolation [23] operation is introduced to extract the feature representations for each representative landmark:

$$\mathbf{x}(l_i + \Delta l_i) = \sum_q G(l_i + \Delta l_i, q) \cdot x(q), \tag{1}$$

where q represents all integral spatial locations in the feature map $\mathbf{x}$ and $G(\cdot, \cdot)$ denotes the bilinear interpolation kernel. l_i is the i^{th} candidate point. Since we compute the feature maps on the 2D plane, $G(\cdot, \cdot)$ is a two-dimensional function, which can be divided into two one-dimensional kernels:

$$G(l, q) = g(l_x, q_x) \cdot g(l_y, q_y), \tag{2}$$

$$g(l_x, q_x) = max(0, 1 - |l_x - q_x|), \tag{3}$$

$$g(l_y, q_y) = max(0, 1 - |l_y - q_y|), \tag{4}$$

where l and q are the location on the corresponding feature maps. We observe that the non-zero locations are very sparse, which is useful for computing $G(l, q)$. Finally, we achieve the feature representation for each discovery landmark. To update the parameters of 1×1 convolution layer, we employ the standard backpropagation through the bilinear operations. Some qualitative results with different number of landmarks are shown in Fig. 5.

Dynamic Landmark Pooling: To integrate the detected landmarks into the process of feature learning, we propose a dynamic landmark pooling based on the localization of landmarks. Existing methods use the region of interest (RoI) pooling operation [49]–[52] to fuse the locationsâĽ™ information and convert the feature maps to fixed length vector representations. Although RoI pooling has achieved impressive results on object detection, it would not be suitable for sketches, due to large areas of white space and structure variations, which may lead the pooled representation to be ambiguous and sensitive to visual variations.

We introduce the dynamic landmark pooling as shown in Fig. 4 that pools the features based on the detected landmarks. Based on representative landmarks, we extract all the features to develop a novel feature map $\hat{\mathbf{x}} = \{\mathbf{d}_{l1}, \ldots, \mathbf{d}_{lN}\} \in \mathbb{R}^{1 \times N \times C}$, where l_i is the location of the i^{th} landmark. After that, we pool the landmarks score maps $\hat{\mathbf{x}}$ into a vector representation via average pooling $\mathbf{f} \in \mathbb{R}^{1 \times C}$.

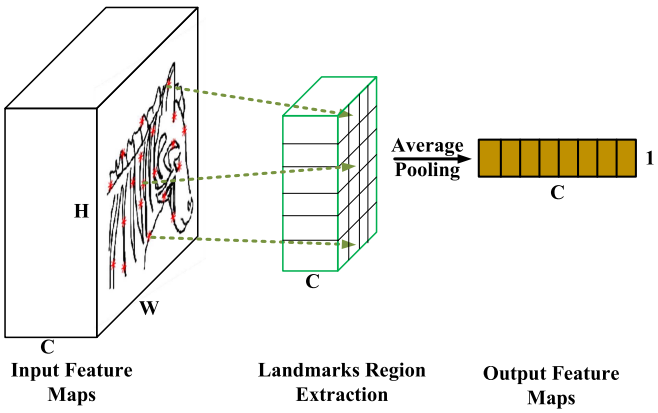


Fig. 4. Illustration of dynamic landmark pooling. The size of input feature map is $H \times W \times C$ and the locations of representative landmarks are denoted as red points. Then, the output feature map is $1 \times N \times C$, which is developed based on pooling the features on N detected points. Finally, we use average pooling on the output maps to obtain the final sketch representation whose size is $1 \times C$.

To optimize representative landmarks discovery module, we construct the SoftMax loss function with the category labels of sketch images.

$$\begin{aligned} L_{cls}(I^i, y^i, W_{fc}) &= -log P(y^i = k | I^i, W_{fc}) \\ &= -log \frac{e^{-f^k(I^i, W_{fc})}}{\sum_{l=1}^{K} e^{-f^l(I^i, W_{fc})}} \end{aligned} \tag{5}$$

where I^i represents the i^{th} input image, y^i is its label, and K is the number of sketch categories. k is the category label of current input image and W_{fc} denotes the weights in the fully connected layers, which is used to map the extracted deep features to the labels.

B. Category-Aware Representation Learning

The representations extracted from the representative landmark discovery module are high-capacity descriptors and robust to sketch local visual variations. However, some categories of sketches share similar structures, e.g. sheep, cat, and dog. The learned representations may be ambiguous due to these similar structures. To further improve the discrimination of feature representations, we introduce the global category-specific representations. To that end, a set of category-aware filters are added on top of the last convolutional layer.

Specifically, we design a set of category-aware filters $F = \{F_1, \ldots, F_K\}$ on top of the last convolutional layer as shown in Fig. 2 (d). To achieve the unique features of each category, we require the filters F_k to be active on diverse regions. We implement the diversity regularization by encouraging orthogonality of the filters' responses and enforcing the filters to be orthogonal. We introduce the diversity loss function L_{div}^f and the regularization L_{div}^r, which represent the loss of filter

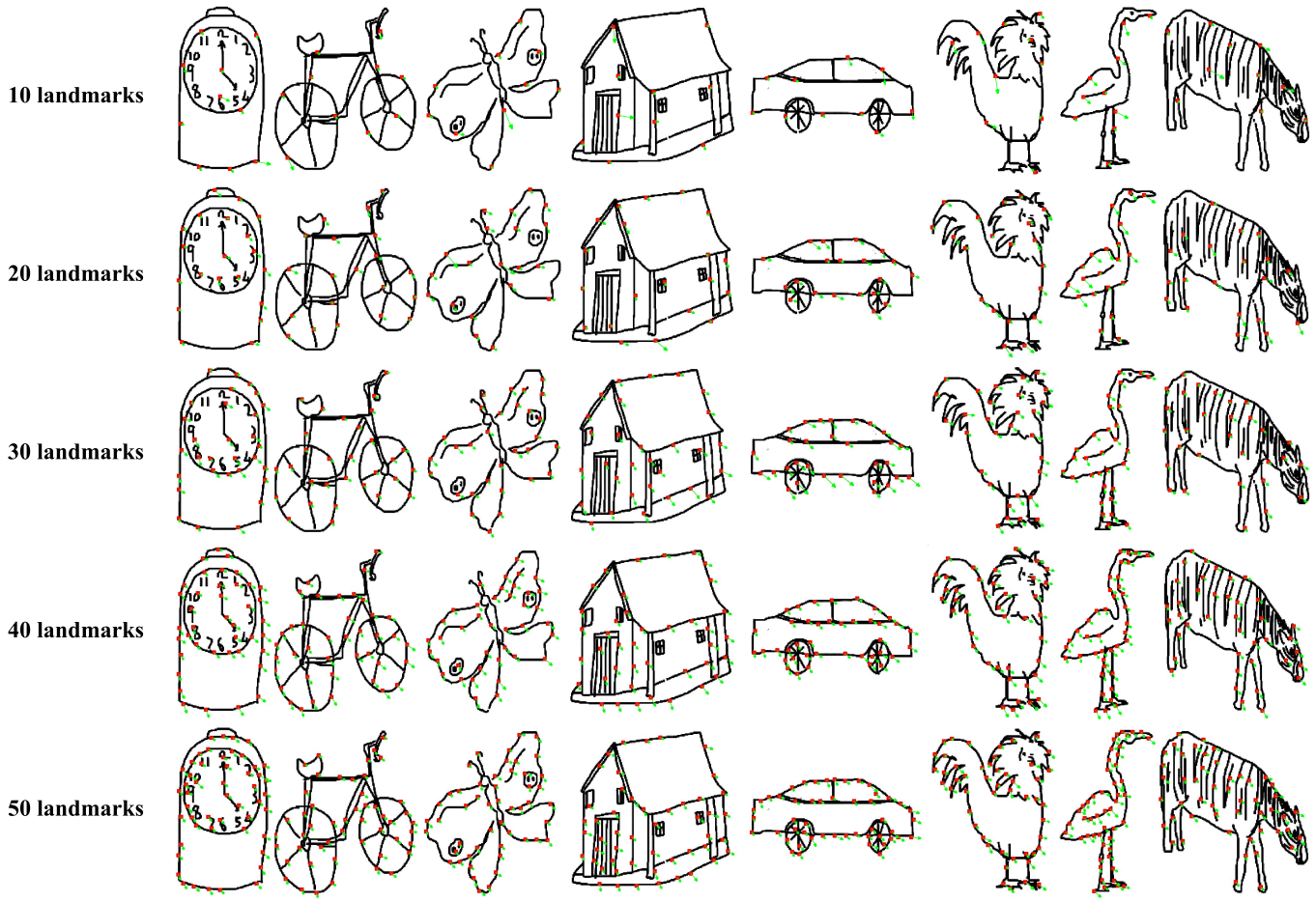


Fig. 5. Illustration of representative landmarks on sketch images. Red points of the sketch indicate the uniform sampling points, and green arrows are the indicators of the computed offsets. We display different number of landmarks on distinct categories. We can observe that the representative landmarks are localized on similar locations for the same sketch with different number of landmarks. While distinct categories are focus on different regions, which could be observed along the row directions. Please zoom in for the details and best view in color.

and their responses, respectively.

$$L_{div}^{f} = \sum_{i \neq j} | \sum_{p} \frac{\langle F_i^p, F_j^p \rangle}{||F_i^p||_F ||F_j^p||_F} |, \tag{6}$$

where F_i^p denotes the filter weighting score in the position p and $||\cdot||_F$ is the F-norm. i and j are the indices for different filters. To further improve the discrimination of filters, we add the diversity on the feature maps $\psi_k^I = \psi(F_k * \varphi(I))$, where $\varphi(\cdot)$ is the function to extract the representations. Then, the regularization could be defined as:

$$L_{div}^{r} = \sum_{i \neq j} | \frac{\langle \psi_i^I, \psi_j^I \rangle}{||\psi_i^I||_F ||\psi_j^I||_F} |, \tag{7}$$

where ψ_i^I and ψ_j^I denote the feature maps generated by the i^{th} and j^{th} filters. With this diversity regularization, the filters should be sparsely distributed on the sketch, especially on the latent semantic parts.

In Fig. 6, we provide the visualizations of the learned category-aware filters on different categories. To achieve the heatmaps, we first extract the outputs of category-aware filters, and then normalize the scores of outputs with softmax operation. After that we upsample the score of output to the original input size. It can be seen that: (i) Across all the category, the category-aware filters tend to be associated with the specific parts of the object, which is complicated and distinct visual patterns. For example, the shell of sea turtle, the body of bird and the trigger of revolver. (ii) The category-aware filters seen to align well across different poses of sketch images, e.g. the leg of sheep and the head of cows.

For most sketch images, uniformly sampling the landmarks along edges are insufficient to represent visual contents. Because there may exist repeated patterns and regional amplifications, which degenerate the discrimination of learned representations. While based on the introduced diversity regularization, our model can automatically encourage the representative landmarks to be distributed at distinct local regions. Moreover, the regularizer can also prevent the network from overfitting to specific details of individual sketch images.

It should be noted that our proposed diversity regularization is conceptually similar to [45] in that it is enforced the representation of different landmarks to be orthogonal. While [45]

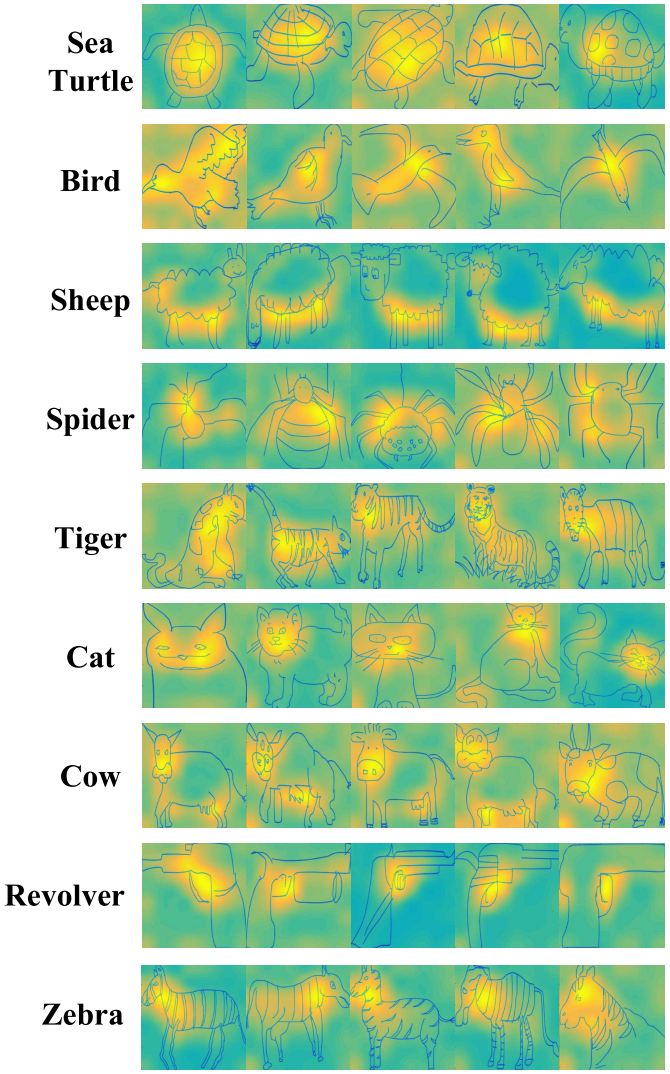


Fig. 6. Visualization of category-aware filters on sketch images. The highlight region of each image indicates that it plays an important effect on determining the category labels. From the visualization, we can observe that not all the pixels are contributed equally for category prediction.

utilizes only positive samples in learning the diversity filters, our method considers both positive and negative samples to enhance representation ability of filters.

To avoid the trivial solution, we further introduce the SoftMax loss function L_{div}^{c} on top of the category-aware responses. Finally, this module is optimized based on the traditional back-propagation with the category label as the supervision. Our proposed diversity regularization indirectly enforces the filters to localize the unique regions of the sketch categories. It is straightforward to show the sparse responses on sketch images, and capturing of salient structures for the category. This observation is also strongly supported by the results of our experiments.

C. Optimization of the Whole Network

To develop an end-to-end deep architecture termed as landmark-aware ConvNet as shown in Fig. 7, we introduce a multi-task framework to update the parameters.

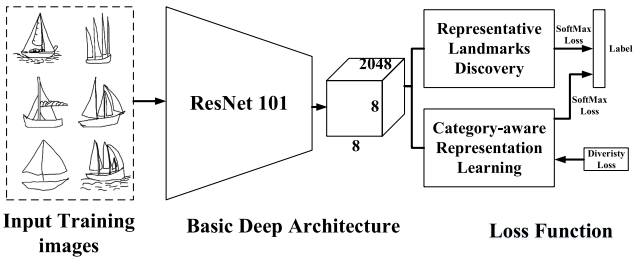


Fig. 7. Basic architecture of our proposed landmark-aware ConvNets. We employ the ResNet101 as the basic deep structure that ignores the fully connected layers. For the loss function, a multi-task framework is utilized to fuse all the loss functions.

The proposed deep network is optimized with stochastic gradient descent (SGD) by minimizing the sum of the proposed losses. The ResNet-101 network pretrained on the ILSVRC12 is used to initialize the front-end CNN of our model. Note that, we do not fix any layer of ResNet101 as in the traditional methods. Instead, we retrain the whole deep neural network whose parameters are updated with the total loss L_{total}.

$$L_{total} = L_{cls} + \beta L_{div} \tag{8}$$

where L_{cls} indicates the softmax loss on the branch of landmarks discovery and $L_{div} = L_{div}^{f} + L_{div}^{r} + L_{div}^{c}$ is the diversity loss on the branch of diversity feature as shown in Fig.7. β is the trade-off parameter.

In the inference step as shown in Fig. 2 (e), a 3-channel sketch image is fed into our proposed neural network with different number of landmarks. Considering the differences between sketch classification and sketch-based retrieval, we develop two different strategies of inferencing. For sketch classification, we just use the predictions of each network, where all the predictions extracted from different landmark-aware networks are fused to achieve the final result. We introduce a fusion layer to give different landmarks distinct weightings, which is computed by using the validation set.

For sketch retrieval, our aim is to compute the similarities between a query sketch and real images. Therefore, we extract the feature representations $\mathbf{f} \in \mathbb{R}^{1\times C}$ from the representative landmarks discovery module and the representation $\mathbf{f}^{s} \in \mathbb{R}^{1\times 64}$ for each category from the category-aware representation learning by max-pooling operations. After that, we concatenate these two types of representations to develop the final representations. Finally, we compute the cosine distances between the query sketch image and real images, and then real images are ranked based on their similarities.

IV. Experiments

To demonstrate the effectiveness of our proposed landmarks aware models for sketch classification and retrieval, we performed experiments on two publicly available datasets: TU-Berlin [33] and Sketchy [34]. In the following, we first describe the experimental setting and evaluation metric, an then present the experimental results and analysis.

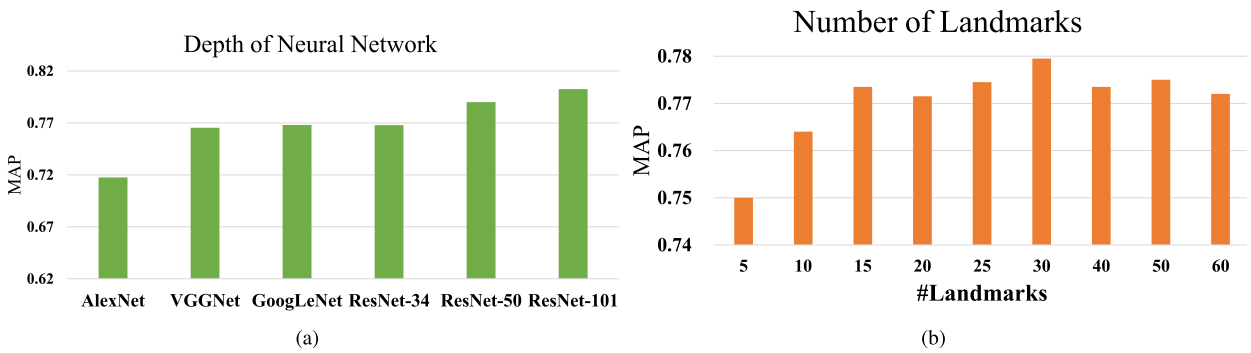


Fig. 8. Component analysis on TU-Berlin dataset. Experimental results of sketch classification based on: (a) different depth of deep neural networks, (b) different number of landmarks.

A. Experimental Setup and Evaluation Metrics

TU-Berlin Dataset[1] [33] is a challenging benchmark for sketch classification and recognition. It is composed of 250 object categories, which cover mostly daily objects in the life, *e.g.* train, cup. There are 80 sketch images for each category, which are collected without any common references. After the dataset was constructed, there was a phase where humans tried to recognize the sketches and the mean recognition accuracy over all the 250 categories is 73%.

Sketchy[2] [34] is proposed for fine-grained sketch based image retrieval, which is composed of $75,471$ hand-drawn sketches of 12,500 objects (images) from 125 categories. This database provides the finegrained associations between particular photos and sketches.

Following previous works [22], [40], [53] we adopt the similar evaluation metrics to quantitatively assess the performance of our model on sketch classification and retrieval. Specifically, on sketch classification, we use the Average Precision (**AP**) to evaluate the performance of classification on each sketch category. And then Mean Average Precision (**MAP**) is used to compare with the existing methods. To evaluate the performance on sketch based image retrieval, we report the **top-k accuracy** where k is the number of retrieved real images. Given a query sketch image, our method first computes the distance between sketch and real images whose identity matches that of the textual query class.

B. Implementation Details

We implemented the proposed deep model using the popular PyTorch framework on a single NVIDIA TitanX GPU with 12G memory. The ResNet-101 network pretrained on the ILSVRC12 is used to initialize the front-end CNN of our model. To improve the robustness of our model, we augment the training samples. All the sketch images are cropped with the object centered at the images with roughly the same scale, and resized the image patch to the size of 256×256. Then, we randomly rotate ($\pm 60°$) and flip the images. We used a learning rate of 10^{-2}, and a momentum of 0.0005. For testing, we follow the standard flowchart to crop the image patches with the given rough positions. All the experimental results on different tasks are produced from the original and flipped images. We train the model on the TU-Berlin dataset for 120 epochs and on the Sketchy dataset for 80 epochs.

[1]http://cybertron.cg.tu-berlin.de/eitz/projects/classifysketch/
[2]http://sketchy.eye.gatech.edu/

C. Ablation Studies

We conduct the ablation study on the TU-Berlin benchmark and report the classification results produced by the landmark discovery module. Following the related work [22], the dataset is first divided into the training and testing sets. The training sketch images for each class are randomly determined, and the remaining images are used as the test set. Moreover, 20% training data are randomly selected to form the validation set.

The effect of front-end networks: Our proposed landmarks discovery neural network is a generic framework, so that different kinds of deep architectures can be used as the front-end network. We evaluate the effect of different types of neural network for sketch classifications. To do that, we select six types of different deep neural networks, i.e. AlexNet [54], VGGNet [55], GoogLeNet [56], ResNet-34 [32], ResNet-50 [32], and ResNet-101 [32]. The number of training samples is 72, which is randomly selected from the dataset. The experimental results are shown in Fig.8 (a). We can observe that a deeper CNN architecture produces more accurate predictions, and the ResNet-101 achieves the best performance among all the front-end neural networks. Thus, considering the performance and computation efficiency, ResNet-101 is selected as the basic deep architecture.

The effect of number of landmarks: Next, we conduct the experiments to validate the effectiveness of landmarks. We randomly collected 72 sketch images for each category as the training set, the remaining 8 sketches were used as the test set. Since the size of last convolutional layer of ResNet-101 is $2048 \times 8 \times 8$, the maximum number of landmarks would be 64, which indicates the whole image is being used for feature representation. We set the number of landmarks as $\{5, 10, 15, 20, 25, 30, 40, 50, 60\}$ and display the experimental results in Fig.8 (b). With the number of landmarks growing, the performance does not always improve, and the best performance is achieved with 30 landmarks. This observation can confirm that not all pixels are

TABLE I

EVALUATION OF THE TRADEOFF PARAMETER OF THE LOSS FUNCTION ON TU-BERLIN DATASET FOR SKETCH CLASSIFICATION TASK

Param.	β (w/o)	β = 0	β = 1e-4	β = 1e-3
MAP	75.97%	74.44%	77.35%	**77.95%**
Param.	β = 1e-2	β = 1e-1	β = 1	β = 10
MAP	77.65%	76.45%	59%	47.35%

contributing equally to sketch classification and the most important region should be emphasized. With the right number of landmarks, our model is robust to the sketch visual variations. The experimental results also demonstrate that multiple landmarks have the ability to describe the whole sketch images, however, only landmarks are also sensitive for the sketch visual variations. So, we need to extra feature representations to encode complementary information. Though necessary, the extra features should encode the holistic structure information that can overwhelm the local similarity reflected by the landmarks. To obtain these features, we introduce the class-aware branch to learn the discriminative feature representations.

The effect of diversity regularization: Finally, we conduct an evaluation on the tradeoff parameter β of Eq. 8. To build the experiments, we select the number of landmarks as 30, and then randomly choose 64 sketch images as the training set. After that we select 8 images from the training samples to develop the validation set, which is used to validate the performance of classification with different values for β. To demonstrate the advantage of introducing the diversity regularizer, we introduce two novel baselines. The first one is that we remove the whole class-aware branch denoted as $\beta(w/o)$, and the other one is that we save the class-aware branch without the diversity regularization denoted as $\beta = 0$. The experimental results are shown in Table I. We observe that the best performance is achieved by setting $\beta = 1e-3$. It points out that our proposed diversity regularization is benefit for learning the discriminative feature representation. Comparing with the performance of $\beta(w/o)$ and $\beta = 0$, we can see that the performance is degenerated when the class-aware branch is directly introduced ($\beta = 0$). This could be explained by that without the diversity regularization, class-aware branch is more focus on learning the seen patterns, which would be sensitive to the sketch variations. Thus, β is fixed to $1e-3$ in all our experiments.

D. Sketch classification on TU-Berlin

In this section, we conduct the experiments to validate our proposed model on sketch classification. First, we randomly select $\{8, 16, 24, 32, 40, 48, 56, 64\}$ from each category to develop 8 types of training sets. The corresponding remaining images are treated as the test set. The validation sets are developed by selecting 20% samples from the training sets. For each type, we use the same training set to learn the model with different number of landmarks and the fixed testing set is used to evaluate the performance of learned classifiers. While ResNet-101 is chosen as the basic deep neural network. Moreover, the number of landmarks is set to $\{5, 10, 15, 20, 25, 30, 40, 50, 60\}$. We further develop an ensemble method, which fuses all the results based on different number of landmarks.

To demonstrate the advantage of our model, we conduct four types of baselines as follows:

- "only landmarks" shows that we remove the offset layer in representative landmarks discovery module and the whole category-aware representation learning module.
- "landmarks+offset" is that we only use the landmarks discovery module and remove the whole category-aware representation learning module.
- "landmarks+diversity" indicates that we remove the offset layer in representative landmarks discovery module and save the whole category-aware representation learning module.
- "landmarks+offset+class-aware" represents that we only remove the diversity regularization in the category-aware representation learning module.

The full model of our proposed method is represented as "landmarks+offset+diversity". The comparison results are shown in Fig.9. It is evident that with our discovered landmarks and diversity regularization term more accurate classification prediction can be achieved, demonstrating that our idea of introducing landmarks from CNN-side output maps is more effective than directly using the whole sketch images. There are also experimental results on proofing the effectiveness of each component. Specifically, we can see that landmarks with offset in general yield better performance than the baselines without offsets ("only landmarks" VS "landmarks+offset" and "landmarks+diversity" VS "landmarks+offset+diversity"). Comparing "only landmarks" with "landmarks+offset", the experimental results show a slight improvement by introducing the offset layer. The experimental results can be explained by that the feature representations of baselines have limited discrimination. With the landmarks, image features are focusing on the image local region, which may be easily dominated by the repeatable patterns. Therefore, even with the offsets, there exists a little improvement on the performance. To further demonstrate the effectiveness of the offset layer, we conduct one more baseline "landmarks+diversity" as shown in Fig. 9 and Table II. Comparing "landmarks+diversity" with "landmarks+offset+diversity", we can observe that there is a significant improvement on the performance, which has demonstrated the effectiveness of the offset layer. This indicates that with the dynamical setting, our model can localize the representative landmarks, which demonstrates the advantages of introducing the landmarks discovery. Comparing the results with and without diversity regularization ("only landmarks" VS "landmarks+diversity" and "landmarks+offset" VS "landmarks+offset+diversity"), the experimental results prove that the diversity regularization is helpful to develop the category-specific feature representations. As expected, we can observe that the "Ensemble" achieves the best performance in all the situations. This proves that different categories need different number of landmarks for description. Moreover, the introduced layers play different effect on distinct

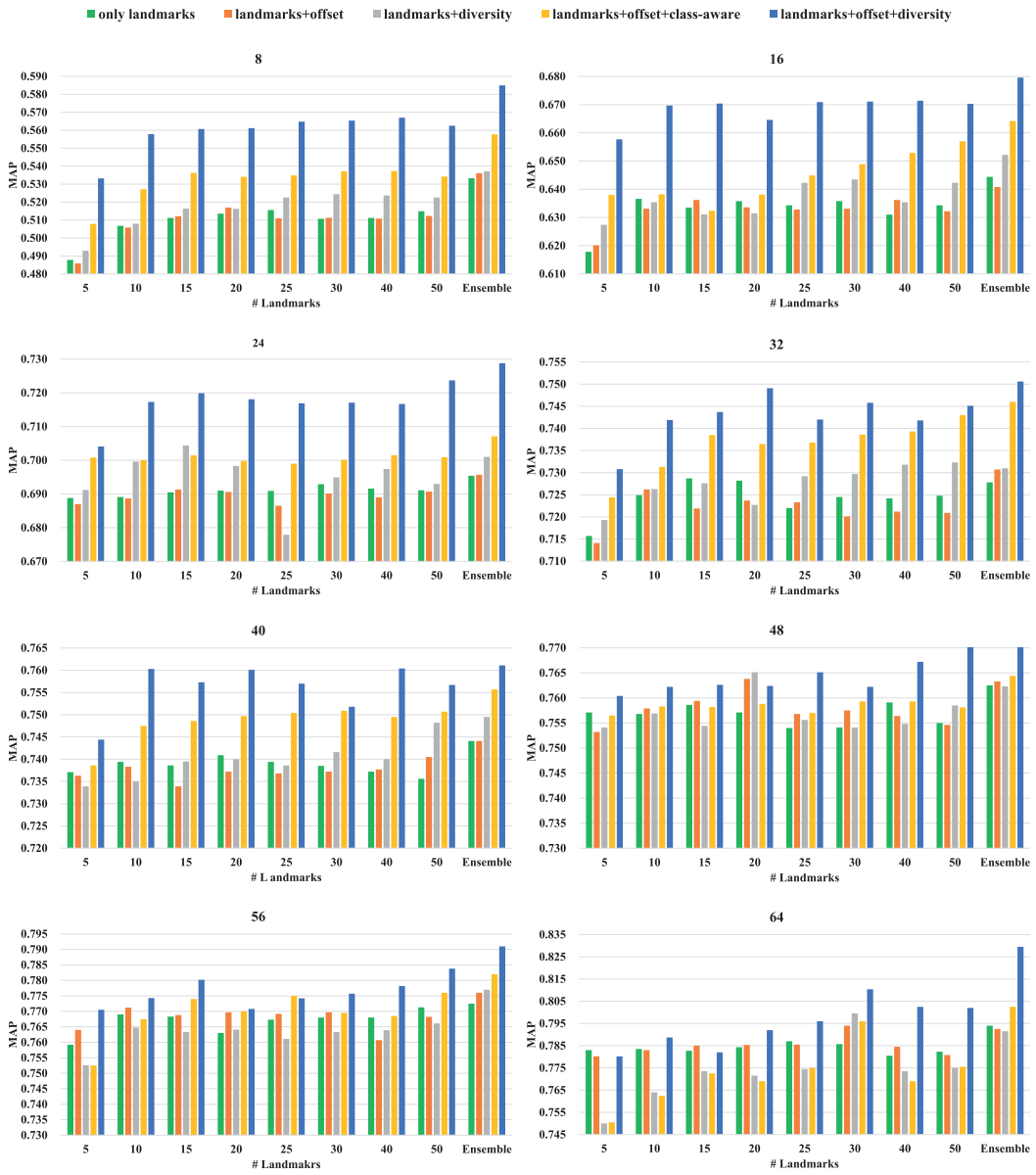


Fig. 9. The comparison of classification results on TU-Berlin sketch benchmark with different number of training samples and different number of landmarks. The green bar indicates that we only use the landmarks without the offset, while the orange bar is adding the offset on the landmarks. The gray bar represents the results by introducing the diversity loss on the landmarks without using the offset layer. The yellow bar indicates that we only remove the diversity regularization in the category-aware representation learning module. And the blue bar shows the results based on our whole deep architecture.

TABLE II

COMPARISON RESULTS ON TU-BERLIN WITH STATE-OF-THE-ART METHODS BASED ON DIFFERENT NUMBER OF TRAINING IMAGES

Methods	Low (8)	Medium (40)	High (64)
ResNet-101 [33]	51.36%	73.03%	78.25%
SketchNet [23]	58.04%	73.54%	77.33%
Sketch-a-Net [41]	57.58%	72.96%	77.95%
Alexnet-FC-GRU [54]	57.95%	71.39%	79.95%
Our method (only landmarks)	51.07%	73.85%	78.57%
Our method (landmarks + offset)	51.63%	73.12%	79.40%
Our method (landmarks + diversity)	52.44%	74.16%	79.65%
Our method (landmarks + offset + class-aware)	53.71%	77.60%	79.60%
Our method (landmarks + offset+diversity)	56.54%	75.18%	81.40%
Our method(Ensemble)	**58.50**%	**76.11**%	**82.95**%

categories. When we fuse these components, there exists a significant improvement on the performance. In some case, the performance is degenerated with the number of landmarks growing, which can be explained by that too many landmarks may make the representations focusing on the repeatable patterns instead of discriminative regions.

TABLE III

COMPARISON RESULTS (TOP-1 ACCURACY) OF CATEGORY-LEVEL SKETCH BASED IMAGE RETRIEVAL ON SKETCHY DATASET USING DIFFERENT CROSS-MODALITY METHODS

Methods	Top-1 Accuracy								
ResNet-101 [33]	0.7434								
Deep shape matching [43]	0.7782								
DSH [16]	0.7331								
GDH [42]	0.7529								
Our model	Number of landmarks								
	5	10	15	20	25	30	40	50	Ensemble
only landmarks	0.6203	0.6522	0.6091	0.6388	0.5695	0.6523	0.6272	0.6979	0.7762
landmarks+offsets	0.7308	0.7331	0.7191	0.7013	0.7062	0.6752	0.6656	0.7128	0.8053
landmarks+diversity	0.6560	0.7076	0.6369	0.7003	0.7217	0.6814	0.6947	0.7265	0.7821
landmarks+offsets+class-aware	0.7450	0.7771	0.7179	0.7640	0.7573	0.7329	0.7682	0.7794	0.8218
Full model (landmarks+offsets+diversity)	**0.7964**	**0.8012**	**0.7947**	**0.7815**	**0.7726**	**0.7443**	**0.7679**	**0.7997**	**0.8569**

Comparisons with state-of-the-art: To conduct the experiments for comparing with the state-of-the-art methods, three types of training sets are selected with different number of training samples. The low (8) denotes that we use 8 sketch images for each category to train our model, the medium (40) indicates the number of training samples is 40, and the high (64) implies that the number of training samples is 64. For the test set, we use the remaining sketch images for each type. Expected for the Ensemble method, we use the number of landmarks for our model fixed as 30. Furthermore, we also report the classification results by combining all the prediction results based from different number of landmarks, termed as "our method (Ensemble)". The state-of-the-art comparison on TU-Berlin dataset is listed in Table II. To make a fair comparison, we reimplemented the state-of-the-art methods [22], [40], [53] based on their descriptions in the papers, which were then trained based on the same training set.

From the table, we can observe that our model achieves a new state-of-the-art classification reaching to **82.95**% compared to [22], [53] and [40] on TU-Berlin dataset. It is clear that our model is significantly better than previous methods. In particular, comparing with [53], the best performing method in the literature, it is evident that our approach outperforms [53] by a significant margin. It is worth noting that in [22] the authors introduce a large collection of real images as context, while our model is using a much smaller training set to achieve a better performance. Moreover, comparing with the baseline of ResNet-101, our proposed model also achieves improvements, which demonstrates the advantages of our proposed method. However, the method may degenerate without the diversity regularization. This demonstrate our intuition that the landmarks would confuse the similar structures belonging to different categories.

This can be explained as follows: First, different from existing methods using all the sketch to extract feature representation, we propose to extract the discriminative parts for features. This can be robust to the large variations of sketch images, especially the styles of painting. Second, the offsets for each landmark would further help for localizing the latent semantic parts of the sketch. Last but not least, the diversity regularization term provides the category-specific structures for each category, which is useful for developing feature representations.

E. Sketch Retrieval on Sketchy

In this subsection, we verify the discriminative ability of our learned feature representations on sketch based image retrieval. In this experiment, 200 sketches are randomly selected from each category to form the query set, and 50 sketches are randomly chosen to construct the validation set, and the remaining images are used as the training set. Moreover, all the real images are treated as the gallery images. We set the number of landmarks as {5, 10, 15, 20, 25, 30, 40, 50}. To reduce the domain gap, all the real images are preprocessed with the HED edge detector [57]. Then, the edge map of each real image is repeated 3 times to develop a 3-channel input.

We implemented and trained the ResNet-101 [32] as the baseline that takes the 3-channel image as input. To train this model, the real images are first preprocessed by the edge detector to achieve their edge maps, which is repeated 3 times to generate a 3-channel input. Then, we resize the longest axis of our images to 256 pixels and pad the shorter axis with white pixels such that the input image is 256×256. Next, the ResNet-101 network pre-trained on the ILSVRC12 is employed to initialize the parameters of ResNet-101. Finally, we use the last fully connected layer of ResNet-101 as feature representations to compute the similarities between query sketch and real images. We also introduce four types baselines "only landmarks", "landmarks+offset", "landmarks+diversity", and "landmarks+offsets+class-aware", which are defined in Sec. IV-D. Differently, we extract the last fully connected layer of representative landmarks discovery module is used as the feature representations. Finally, we use "Full model" to indicate our whole framework, which uses the feature representation from two modules. For details, we extract the last fully connected layer from the landmarks discovery module, and the representations from the category-aware representation module with max pooling operations. Then, we concatenate these two representations to develop the final feature representation. To further demonstrate the advantage of our model, we also introduce some recent methods on

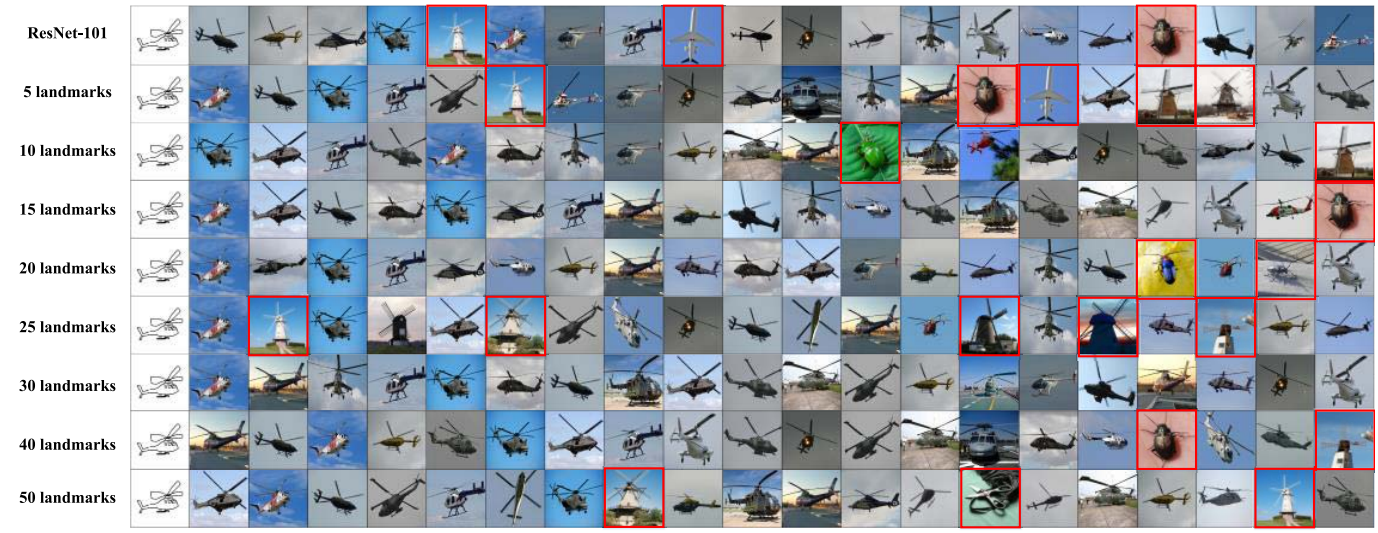


Fig. 10. Examples of sketch based image retrieval results on the Sketchy dataset. Qualitative comparison with different number of landmarks on the same query sketch is presented. The red bounding boxes indicates the false positive results.

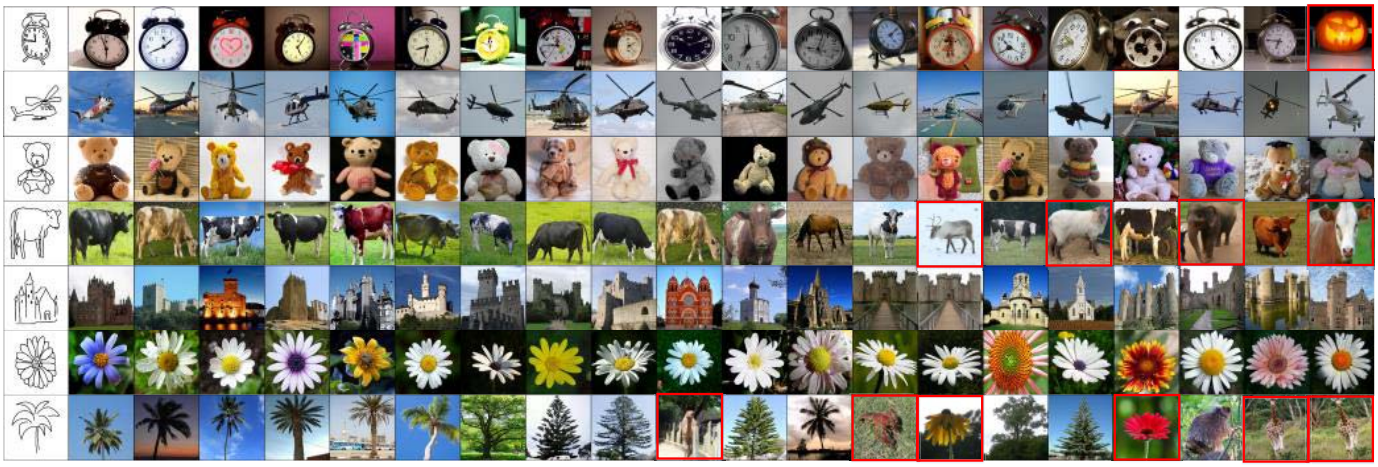


Fig. 11. The visualization of our sketch based image retrieval on Sketchy dataset: five example query sketches with their top-20 retrieval results on Sketchy dataset. Red bounding box indicates the false positives.

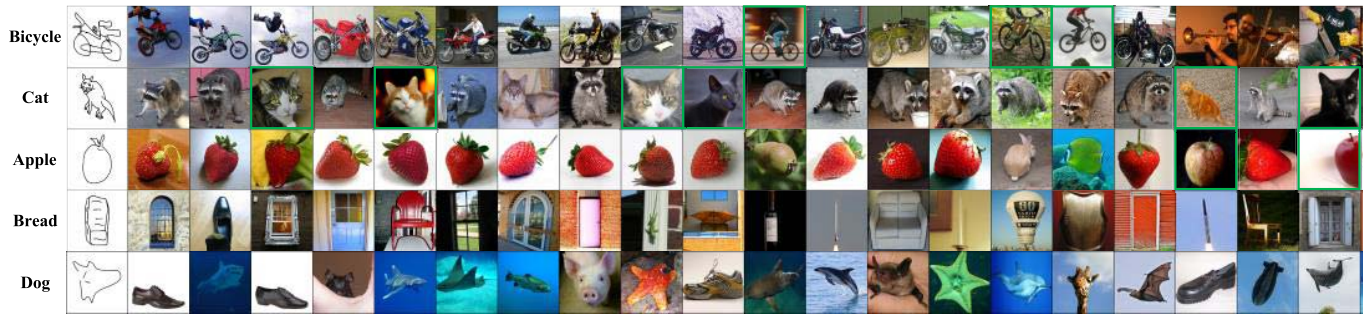


Fig. 12. Illustration of bad cases of our proposed method with their top-20 retrieval results on Sketchy dataset. The text indicates the category label. And the green bounding box indicates the correct prediction.

sketch based image retrieval as the baselines, which are Deep shape matching [42], DSH [16] and GDH [41]. To achieve a fair comparison, we have re-implemented those methods by using the same training and testing sample in our experiments.

We report the top-1 accuracy as shown in Table III. Some qualitative results on the Sketchy dataset are shown in Figs. 10 and 11. It is clear that the proposed approach is able to generate better retrieval results, which demonstrates the effectiveness our proposed model. We believe that this

is probably because the effective landmarks discovery and category-specific representation can yield more discriminative structural representations to bridge the domain gap.

Our proposed method leads to superior results with **85.69**% accuracy, and achieves significant improvements over the best-performing competing methods. The reason is that our framework can discover the representative landmarks both for the sketches and real images, which enables to bridge the domain gap between the sketches and real images. Furthermore, the diversity representation is able to capture the category-specific structure features for each category.

There are also some bad cases for our proposed method as shown in Fig. 12. We can observe that because of the drawing perspective, the sketch "bicycle" is more similar to a "motorbike" than a bicycle. While the "bread" and "dog" are very ambiguous, i.e. hard to determine their category just based on the sketches. In this case, we need the drawers to provide more text descriptions or other contextual information for determining the category.

V. Conclusion

In conclusion, we have proposed a novel architecture, named, landmark-aware network, for weakly supervised sketch recognition and retrieval. Our model is composed of two novel modules the representative landmarks discovery module and the category-aware representation learning module. We show that enabling representation model with representative landmarks and category-aware features, through different modules, leads to a large performance gain on the task of sketch classification and sketch based image retrieval. Experiments on TU-Berlin and Sketchy yield promising results and demonstrate the validity of the proposed landmarks-based approach. In the future, we will focus on the computational efficiency of extracting the discriminative feature representations, which is an essential component for online sketch retrieval. More interesting directions would involve developing more complex architectures using our proposed approach to solve even more challenging vision tasks.

References

[1] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: Internet image montage," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 124:1–124:10, Dec. 2009.

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

[3] J. Collomosse, T. Bui, M. Wilber, C. Fang, and H. Jin, "Sketching with style: Visual search with sketches and aesthetic context," in *Proc. ICCV*, Oct. 2017, pp. 2660–2668.

[4] T. Bui and J. Collomosse, "Scalable sketch-based image retrieval using color gradient features," in *Proc. ICCV Workshops*, Dec. 2015, pp. 1–8.

[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*, Oct. 2017, pp. 5551–5560.

[6] X. Qian, X. Tan, Y. Zhang, R. Hong, and M. Wang, "Enhancing sketch-based image retrieval by re-ranking and relevance feedback," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 195–208, Jan. 2016.

[7] 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.

[8] D. Dixon, M. Prasad, and T. Hammond, "iCanDraw: Using sketch recognition and corrective feedback to assist a user in drawing human faces," in *Proc. SIGCHI*, 2010, pp. 897–906.

[9] X. Han, C. Gao, and Y. Yu. (2017). "DeepSketch2Face: A deep learning based sketching system for 3D face and caricature modeling." [Online]. Available: https://arxiv.org/abs/1706.02042

[10] S. Nagpal, M. Singh, R. Singh, M. Vatsa, A. Noore, and A. Majumdar, "Face sketch matching via coupled deep transform learning," in *Proc. ICCV*, Oct. 2017, pp. 5419–5428.

[11] Y. Song, L. Bao, Q. Yang, and M.-H. Yang, "Real-time exemplar-based face sketch synthesis," in *Proc. ECCV*, 2014, pp. 800–813.

[12] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in *Proc. CVPR*, Jul. 2017, pp. 5400–5409.

[13] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. ICCV*, Dec. 2015, pp. 945–953.

[14] G. Dai, J. Xie, and Y. Fang, "Deep correlated holistic metric learning for sketch-based 3D shape retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3374–3386, Jul. 2018.

[15] A. Borji and L. Itti, "Human vs. computer in scene and object recognition," in *Proc. CVPR*, Jun. 2014, pp. 113–120.

[16] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, "Deep sketch hashing: Fast free-hand sketch-based image retrieval," in *Proc. CVPR*, Jul. 2017, pp. 2862–2871.

[17] G. Tolias and O. Chum. (2017). "Asymmetric feature maps with application to sketch based retrieval." [Online]. Available: https://arxiv.org/abs/arXiv:1704.03946

[18] F. Huang, Y. Cheng, C. Jin, Y. Zhang, and T. Zhang, "Deep multimodal embedding model for fine-grained sketch-based image retrieval," in *Proc. ACM SIGIR*, 2017, pp. 929–932.

[19] J. Xie, G. Dai, F. Zhu, and Y. Fang, "Learning barycentric representations of 3D shapes for sketch-based 3D shape retrieval," in *Proc. CVPR*, Jul. 2017, pp. 5068–5076.

[20] C. Zhang *et al.*, "Generalized latent multi-view subspace clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[21] A. Creswell and A. A. Bharath, "Adversarial training for sketch retrieval," in *Proc. ECCV*, 2016, pp. 798–809.

[22] H. Zhang, S. Liu, C. Zhang, W. Ren, R. Wang, and X. Cao, "Sketchnet: Sketch classification with web images," in *Proc. CVPR*, Jun. 2016, pp. 1105–1113.

[23] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[24] K. Lenc and A. Vedaldi, "Understanding image representations by measuring their equivariance and equivalence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 991–999.

[25] T. S. Cohen and M. Welling. (2014). "Transformation properties of learned visual representations." [Online]. Available: https://arxiv.org/abs/arXiv:1412.7659

[26] I. Rocco, R. Arandjelović, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proc. CVPR*, vol. 2, Jul. 2017, pp. 6148–6157.

[27] I. Rocco, R. Arandjelović, and J. Sivic, "End-to-end weakly-supervised semantic alignment," in *Proc. CVPR*, Jun. 2018, pp. 6917–6925.

[28] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. (2017). "Rethinking atrous convolution for semantic image segmentation." [Online]. Available: https://arxiv.org/abs/arXiv:1706.05587

[29] K.-H. Kim, S. Hong, B. Roh, Y. Cheon, and M. Park. (2016). "PVANET: Deep but lightweight neural networks for real-time object detection." [Online]. Available: https://arxiv.org/abs/arXiv:1608.08021

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.

[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, vol. 4, 2017, p. 12.

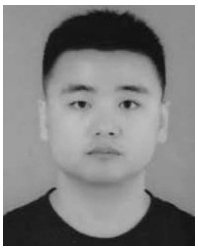
[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.

[33] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Trans. Graph.*, vol. 31, no. 4, pp. 44–52, 2012.

[34] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Trans. Graph.*, vol. 35, no. 4, p. 119, Jul. 2016.

[35] S. A. Eslami, N. Heess, C. K. Williams, and J. Winn, "The shape boltzmann machine: A strong model of object shape," *Int. J. Comput. Vis.*, vol. 107, no. 2, pp. 155–176, 2014.
[36] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[37] F. P. Tasse and N. Dodgson, "Shape2vec: Semantic-based descriptors for 3D shapes, sketches and images," *ACM Trans. Graph.*, vol. 35, no. 6, p. 208, 2016.
[38] Z. Wu *et al.* (2014). "3D ShapeNets: A deep representation for volumetric shapes." [Online]. Available: https://arxiv.org/abs/arXiv:1406.5670
[39] F. Wang, L. Kang, and Y. Li, "Sketch-based 3d shape retrieval using convolutional neural networks," in *Proc. CVPR*, Jun. 2015, pp. 1875–1883.
[40] 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.
[41] J. Zhang *et al.*, "Generative domain-migration hashing for sketch-to-image retrieval," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 297–314.
[42] F. Radenovic, G. Tolias, and O. Chum, "Deep shape matching," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 751–767.
[43] W. Kim, B. Goyal, K. Chawla, J. Lee, and K. Kwon. (2018). "Attention-based ensemble for deep metric learning." [Online]. Available: https://arxiv.org/abs/arXiv:1804.00382
[44] M. Opitz, G. Waltner, H. Possegger, and H. Bischof. (2018). "Deep metric learning with BIER: Boosting independent embeddings robustly." [Online]. Available: https://arxiv.org/abs/arXiv:1801.04815
[45] D. Novotny, D. Larlus, and A. Vedaldi. (2017). "Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching." [Online]. Available: https://arxiv.org/abs/arXiv:1704.04749
[46] C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker, "Universal correspondence network," in *Proc. NIPS*, 2016, pp. 2414–2422.
[47] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned invariant feature transform," in *Proc. ECCV*, 2016, pp. 467–483.
[48] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. ICCV*, Oct. 2017, pp. 764–773.
[49] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. NIPS*, 2016, pp. 379–387.
[50] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, Dec. 2015, pp. 1440–1448.
[51] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. CVPR*, Jun. 2014, pp. 580–587.
[52] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.
[53] R. K. Sarvadevabhatla, J. Kundu, and V. Babu R, "Enabling my robot to play pictionary: Recurrent neural networks for sketch recognition," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 247–251.
[54] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1–9.
[55] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. (2014). "Return of the devil in the details: Delving deep into convolutional nets." [Online]. Available: https://arxiv.org/abs/arXiv:1405.3531
[56] C. Szegedy *et al.* (2014). "Going deeper with convolutions." [Online]. Available: https://arxiv.org/abs/arXiv:1409.4842
[57] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. ICCV*, Dec. 2015, pp. 1395–1403.

Hua Zhang received the Ph.D. degree in computer science from the School of Computer Science and Technology, Tianjin University, Tianjin, China, in 2015. He is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include computer vision, sketch-based image retrieval, multimedia, and machine learning.



Peng She received the B.E. degree in computer science from Chengdu. He is currently a graduate student with the School of The Ministry of Education Key Laboratory, Yunnan Normal University, China. His current research interests include computer vision and sketch retrieval.

Yong Liu was born in 1986. He received the Ph.D. degree in computer science from Tianjin University, Tianjin, China, in 2016. He is currently an Associate Researcher at the Institute of Information Engineering, Chinese Academy of Sciences. His research interests include large-scale kernel methods, large-scale model selection, and machine learning.

Jianhou Gan received the B.S. degree in computer science education and the M.S. degree in mathematics from Yunnan Normal University, Kunming, China, in 1998 and 2004, respectively, and the Ph.D. degree in computational metallurgy from the Kunming University of Science and Technology, China, in 2016. He is currently the Vice Director of the Key Laboratory of Educational Informatization for Nationalities, Yunnan Normal University. His current research interests include knowledge engineering and educational informatization for nationalities.

Xiaochun Cao received the B.E. and M.E. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA. He has been a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, since 2012. After graduation, he spent about 3 years at ObjectVideo Inc. as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He has authored and coauthored more than 200 journal and conference papers. He is a fellow of the IET. He is on the Editorial Board of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. His dissertation was nominated for the University of Central Florida's University-Level Outstanding Dissertation Award. In 2004 and 2010, he was a recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition.

Hassan Foroosh (M'02–SM'03) is currently a CAE Link Professor of computer science at the University of Central Florida (UCF). He has authored and coauthored over 150 peer-reviewed journal and conference papers. He has been serving on the editorial boards and the organizing committees of various IEEE TRANSACTIONS, conferences, and working groups. His research has been supported by the NSF, NASA, ONR, DIA, and various industries.