Tactile Sketch Saliency

Jianbo Jiao[*]
University of Oxford

Ying Cao
City University of Hong Kong

Manfred Lau
City University of Hong Kong

Rynson Lau
City University of Hong Kong

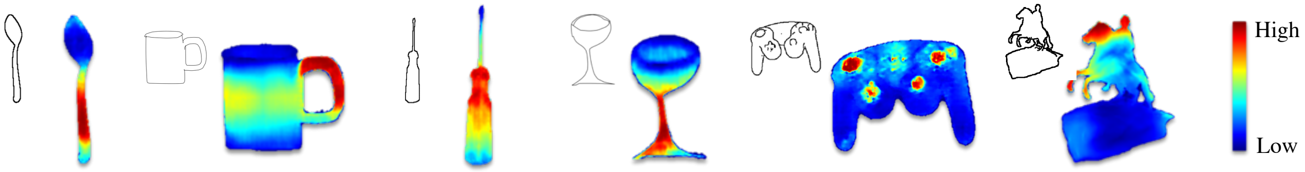


Figure 1: Given an input sketch, our model can predict its tactile saliency map, indicating where people would likely grasp (*e.g.*, for cup), press (*e.g.*, for game controller) or touch (*e.g.*, for statue) the object depicted by the sketch.

ABSTRACT

In this paper, we aim to understand the functionality of 2D sketches by predicting how humans would interact with the objects depicted by sketches in real life. Given a 2D sketch, we learn to predict a tactile saliency map for it, which represents where humans would grasp, press, or touch the object depicted by the sketch. We hypothesize that understanding 3D structure and category of the sketched object would help such tactile saliency reasoning. We thus propose to jointly predict the tactile saliency, depth map and semantic category of a sketch in an end-to-end learning-based framework. To train our model, we propose to synthesize training data by leveraging a collection of 3D shapes with 3D tactile saliency information. Experiments show that our model can predict accurate and plausible tactile saliency maps for both synthetic and real sketches. In addition, we also demonstrate that our predicted tactile saliency is beneficial to sketch recognition and sketch-based 3D shape retrieval, and enables us to establish part-based functional correspondences among sketches.

CCS CONCEPTS

• **Computing methodologies** → **Interest point and salient region detections**; **Image representations**; *Multi-task learning.*

KEYWORDS

Saliency, Sketch, Tactile

ACM Reference Format:
Jianbo Jiao, Ying Cao, Manfred Lau, and Rynson Lau. 2020. Tactile Sketch Saliency. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413588

1 INTRODUCTION

Sketching is one of the intuitive ways for visual communication. The increasing availability and widespread use of sketches have motivated many recent works on sketch analysis (*e.g.*, sketch recognition [31, 42, 52, 54], segmentation [20, 29, 43], consolidating [30] and abstraction [39]), and applications (*e.g.*, sketch-based image retrieval [37, 40, 45, 46, 51] and sketch-based 3D modeling [2, 6, 8, 36, 44]).

When looking at a sketch, humans can recognize not only what object it is depicting, but also where to touch or grasp this object. Therefore, a fundamental question to ask is: *whether a computer is able to reason about how humans interact with an object depicted by a sketch?* Such knowledge can provide useful information about the functionality of a sketch, which in turn helps facilitate visual interpretation of sketches and informs the algorithms for sketch-based 2D/3D content synthesis.

While recent works have focused on analyzing the visual and semantic properties of sketches, in this paper, we make an initial effort to understand the functional aspect of sketches, by addressing the *tactile sketch saliency* prediction problem. Unlike visual saliency that describes where people will *look* at, the tactile saliency of a sketch reflects where people will *grasp, press, or touch* an object as depicted by the sketch. Here the concept of tactile saliency on sketches does not involve any interactions with real objects, but virtually consider where the depicted object could be interacted with. Given a 2D sketch, we aim to predict a tactile saliency map, where the value of each pixel represents its likelihood of being

[*] Work done while Jianbo Jiao was at City Unviersity of Hong Kong.
Ying Cao and Rynson Lau are the corresponding authors.

MM '20, October 12–16, 2020, Seattle, WA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413588

grasped, pressed, or touched. To this end, we propose a learning-based framework to directly map a sketch to its tactile saliency map in an end-to-end manner. Our key observation is that knowing the 3D structure and object category of a sketch can benefit the prediction of which part of the sketch to touch. Hence, our model jointly solves three mutually beneficial tasks: tactile saliency prediction, depth estimation, and sketch recognition in a unified framework. To train our model, a dataset of sketches with labeled tactile saliency is required. However, pixel-wise annotation of saliency values on a large number of sketches is labor-intensive and expensive. To construct our training dataset, we propose a data synthesis strategy, where we render a collection of 3D models with available tactile information to form sketches and transfer the depth and tactile information from the 3D models to the sketches directly.

Our experiments on both synthetic and real sketches show that the proposed method can predict accurate and plausible tactile saliency consistently. Examples are illustrated in Figure 1. Through extensive qualitative and quantitative evaluations, we demonstrate that our model, although trained on synthetic data, can favorably generalize to real sketches. We have also explored a few applications that can benefit from learning to predict tactile sketch saliency. In particular, we show that our predicted tactile saliency can help improve sketch recognition and sketch-based 3D shape retrieval, and makes it possible for us to establish sparse correspondences between functional parts across a collection of sketches.

Our primary contributions are: 1) We introduce the concept of tactile sketch saliency, as an initial effort towards understanding the functional aspects of 2D sketches; 2) We propose a joint reasoning approach for tactile sketch saliency prediction; 3) We show how to leverage synthetic data to learn tactile sketch saliency and propose a data synthesis pipeline, which utilizes 3D shapes to generate 2D tactile sketch saliency data for training; 4) We demonstrate that such tactile saliency can impact many sketch analysis tasks. We believe that the insights brought by our work in learning sketch functionality and how it may benefit other vision tasks could be important contributions to the community and potentially lead to a new research direction.

2 RELATED WORK

2.1 Saliency Prediction

Visual saliency prediction has been an important research problem in computer vision [21]. A recent review of this problem can be found in [4]. Many algorithms have been proposed to model visual saliency using hand-crafted low-level features [7, 12, 34] and high-level features [15, 22, 33]. Recently, there is a trend in using deep convolutional neural networks (CNNs) to automatically learn hierarchical features [3, 17, 19, 28, 32, 47, 48, 55, 56]. In our work, we also make use of CNNs for saliency prediction, but focus on functional saliency rather than visual saliency.

In the area of robotics, there are some studies on affordance detection [23, 35, 41, 53]. However, these works focus on inferring contact and grasping regions from the 3D geometry of an object/scene, while our work, for the first time, introduces the tactile saliency concept to sketches and aims at inferring the functional/tactile saliency from 2D sketches. Our work is closely related to a recent work on computing the tactile saliency on 3D meshes [24] from their projected depth maps. Compared with their work that regresses saliency values from continuous depth maps, our problem setting is more challenging, as we aim at predicting a dense saliency map from a sketch with sparse strokes. Therefore, rather than using a single-task network as in [24], we propose a multi-task architecture to facilitate the learning of our model.

2.2 Sketch Analysis

There have been a lot of research efforts on sketch analysis. A detailed literature review is beyond the scope of this paper. Readers may refer to recent works on sketch recognition [10, 31, 42, 52, 54], segmentation [20, 29, 43, 57], and sketch-based retrieval for 2D images [9, 37, 40, 50, 51] and 3D shapes [6, 8, 11, 27, 46, 49].

In recent years, deep CNNs have been widely applied to solve various sketch-related problems. For example, Yu *et al.* [52] proposed a multi-scale, multi-channel network for sketch recognition, which outperforms humans. Sangkloy *et al.* [40] published a large-scale dataset of sketch-photo pairs and used it to train a cross-domain CNN for fine-grained sketch-based image retrieval. Wang *et al.* [46] used Siamese networks to learn feature representations of both sketches and 2D views of 3D models for sketch-based 3D shape retrieval. A recent work [39] proposed a sketch abstraction model through reinforcement learning of a stroke removal policy without degrading its recognizability. In contrast to these prior works that seek to understand the visual content and semantics of sketches, we focus on the unexplored topic of analyzing the functionality of sketches by considering how humans would interact with real-world objects depicted by sketches.

3 OUR APPROACH

We present a learning-based model for tactile sketch saliency prediction. One naive approach is to train a single-task encoder-decoder (SED) network to directly map an input 2D sketch to a saliency map. Such a basic model, however, does not work well since regressing a continuous saliency map from a set of sparsely-distributed strokes is inherently ambiguous. To address this challenge, we propose to jointly reason about three properties of a sketch: tactile saliency, depth and semantic categories via a multi-task network architecture. Our intuition is that knowing what the object is and its 3D structure can help determine where humans can interact with the object.

3.1 Direct Regression of Tactile Saliency

The SED model follows an encoder-decoder architecture [1, 38]. While the encoder extracts a sketch representation from the input sketch, the decoder takes the feature representation to generate a tactile saliency map. SED is fully convolutional without using any pooling layers or fully-connected layers, which can help preserve spatial information throughout the intermediate representations. This is desirable for our problem since we aim to predict a pixel-wise dense saliency map, which relies on some local information from the input sketch. To propagate low-level information towards the output more efficiently, we use skip connections between encoder and decoder. SED is trained by minimizing a saliency loss L_s that penalizes ℓ_2 distance between a predicted saliency map and the ground truth.

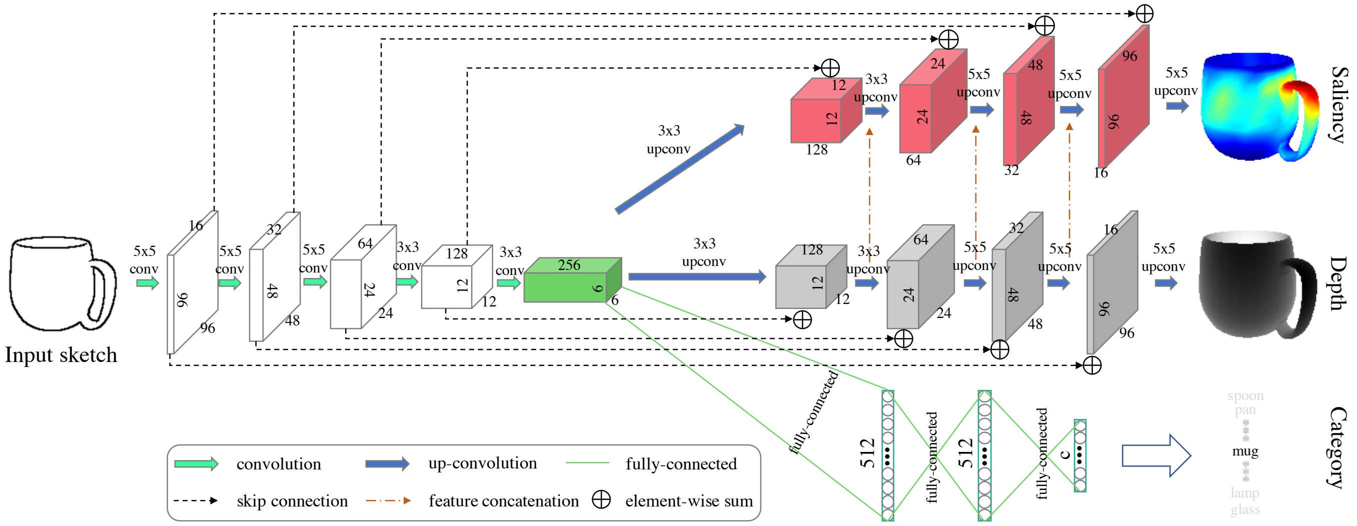


Figure 2: The proposed joint reasoning framework. An input sketch is first fed to the encoder (white cubes with green arrows) to obtain a sketch representation (the green cube), which is then used by three branches for three different tasks: tactile saliency prediction, depth prediction, and sketch classification.

While our network takes as input the entire sketch, it is also possible to predict the saliency based on local patches, as in [24]. However, we have empirically found that patch-based prediction does not give meaningful results for our problem, mainly due to the sparsity nature of sketches where local patches are mostly empty and may not contain enough information for reliable saliency prediction.

3.2 Joint Reasoning of Multiple Properties

Based on the SED model (saliency branch in Figure 2), we develop our joint reasoning framework to incorporate more semantic and structural information to guide the training process, as shown in Figure 2. Specifically, we add a depth map prediction branch to predict the depth information of the input sketch, and a classification branch to predict the category of the sketch. The encoder is shared among the three branches so that it can learn a common sketch representation (green cube) for the three tasks. In order for our network to complete the three tasks well, the encoder should learn a sketch representation that captures the semantics, structure as well as functionality of the input sketch. The depth branch maps the shared sketch representation to a depth map. We concatenate the feature maps of the up-convolutional layers in the depth branch with those of the corresponding up-convolutional layers in the saliency branch, in order to provide multi-level structural features for the saliency prediction. In addition, skip connections are linked from the encoder to both the saliency and depth branches, in order to propagate high-frequency information to the reconstruction process. The classification branch receives the shared sketch representation and outputs the probabilities of object categories. It has two fully-connected layers followed by a softmax layer. ReLU [14] activations are used in each layer of the network.

To train our multi-task network, in addition to saliency loss L_s, we also define a ℓ_2 loss L_d for depth prediction and a cross-entropy loss L_c for sketch classification. Our final loss is a weighted sum of the three task-specific losses: $L = \xi_1 \cdot L_s + \xi_2 \cdot L_d + \xi_3 \cdot L_c$, where $\{\xi_i\}$ are weight parameters to control the contribution of each loss term.

4 TACTILE SALIENCY DATA SYNTHESIS

To train our network, we need a dataset of 2D sketches with tactile saliency and depth annotations. Unfortunately, such a dataset is not available and can be expensive to acquire since it requires labeling the saliency and depth of each sketch in a pixel-wise manner. Hence, we address this problem by proposing a new synthetic training dataset.

To construct our training dataset, we first start with the same 3D shape dataset obtained from [24]. In this dataset, each 3D model already has its dense 3D tactile saliency data generated by their model, where each vertex has a tactile saliency value indicating how likely a person would grasp, press, or touch the location when interacting with the 3D object. There are a total of 19 shape categories for the three tactile modalities, *i.e.*, grasp, press and touch. Although there might be other tactile modalities, as an initial step towards this direction, we choose these commonly observed interaction modalities in this work and leave a more thorough study for future research. To further increase the diversity of our training data, besides the 19 shape categories from [24], we introduce 6 additional shape categories: including 3 categories for the grasp modality (backpack, spray bottle, and handbag) and 3 categories for the touch modality (horse statue, sit-man statue, and human with bird statue). Statues are chosen here since they have been used for studying touch saliency in [24] and shows that people have high consistency in touching for statues. For the new categories and shapes, we generate the 3D tactile saliency data using the same model [24] as for the other 19 categories, which predicts vertex-wise 3D tactile saliency given a 3D shape. In total, we have

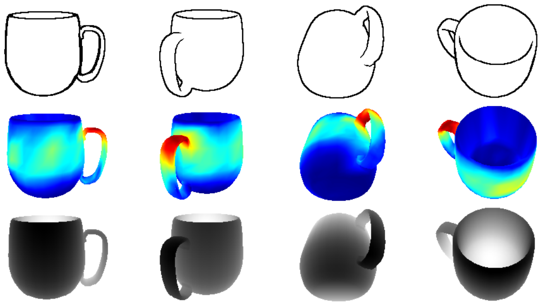


Figure 3: Synthesis of 2D training data from 3D data. Synthesized 2D sketches (top row), 2D tactile saliency maps (middle row), and depth maps (bottom row) from a 3D mug model at different viewpoints (*i.e.*, different columns).

collected 3D shapes in 25 shape categories, which cover many of the shape categories in ShapeNet [5] that humans can interact with in a tactile sense.

Given all the collected 3D shapes, we obtain the 2D data (*i.e.*, sketches, 2D tactile saliency, and depth maps) via projection as follows. For each 3D model M_i from a particular viewpoint v, we generate the 2D synthetic data in the form of (I_{iv}, S_{iv}, D_{iv}), where I_{iv} is a sketch of M_i for viewpoint v, S_{iv} is the corresponding 2D tactile saliency map, and D_{iv} is the corresponding depth map. To generate 2D tactile saliency map S_{iv} and depth map D_{iv}, we use ray casting. To generate 2D sketch I_{iv}, we take the depth map D_{iv} and perform Sobel edge detection to generate an outline of the projected 3D shape. All generated 2D sketches, tactile saliency maps, and depth maps have a resolution of 200 × 200. We end up with around 1,800 synthetic sketches in our dataset. Figure 3 shows some examples of the synthetic data.

5 EXPERIMENTS

In this section, we evaluate our model quantitatively and qualitatively on both synthetic and real sketches, test the importance of network design choices, and demonstrate the applications of our model in various sketch analysis tasks.

5.1 Implementation Details

Our network is trained end-to-end using mini-batch stochastic gradient descent, with a momentum of 0.9 and a batch size of 1. We use an initial learning rate of 10^{-5} and divide it by 10 for every 20 epochs until 10^{-7}. We initialize our network with the modified Xavier initializer [16], and train it until convergence. The weights for the saliency, depth, and classification loss terms are set to 0.4, 0.3, and 0.3, respectively. We split our dataset so that 85% of it is for training and the remaining 15% for testing. We perform data augmentation on the training set by random cropping, flipping, and rotation (with a degree between [-10, 10]). The augmentation results in a total of 2 million training examples. Dataset and implementations are available online[1].

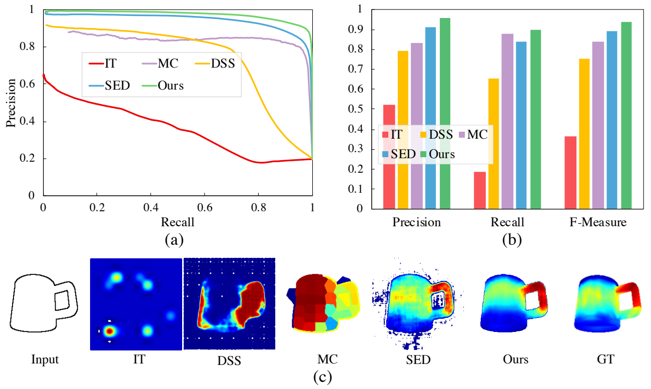


Figure 4: Comparison of our model against some baselines, including three methods on visual saliency prediction (IT [21], DSS [18] and MC [55]) and our SED: (a) precision-recall curves; (b) precision, recall, and F-measure values; (c) a visual comparison of the results from all these methods.

5.2 Evaluation on Synthetic Data

We first evaluate the effectiveness of our model on our synthetic test dataset.

Evaluation Metrics. Since the ground truth saliency maps for the synthetic sketches are known, we measure the performance of our model using two common metrics for saliency detection: PR-curve and F-Measure [21]. Most commonly used evaluation metrics for saliency detection are mainly based on pixel-wise mean square error (MSE). However, for our problem, we are more interested in the relative saliency between two points (*i.e.*, whether one point is more salient than the other) than the absolute saliency value at a single point. Further, these metrics require pixel-wise ground truth saliency maps, which are quite expensive to obtain for real sketches. Thus, inspired by [24], we instead propose a point-pair based evaluation metric especially designed for our problem. Given a predicted tactile saliency map, we first generate several pairs of points, where each pair consists of two randomly-sampled points (within the depicted object). We then compare the saliency values at the two points to determine their *relative relation* (*i.e.*, greater or smaller). The relative relation between a pair of points is considered as correctly predicted if it is the same as the ground truth.

Baselines. Since there are no prior works on tactile sketch saliency prediction, we use visual saliency prediction methods as baselines for comparison: a classic method with hand-crafted features (IT) [21] and two recent deep-learning-based methods using multi-scale contexts and features (DSS [18] and MC [55]). We have chosen DSS and MC because their multi-scale approach can possibly deal with the sparsity of sketches better, by combining features from regions of different scales for more reliable prediction. We fine-tune these methods on our training dataset before comparison. Note that we also experiment with training DSS and MC on our training data from scratch, but found the performance is worse than fining-tuning them. The SED is also included for comparison.

Quantitative Results. Figure 4(a) and 4(b) show the quantitative performance on precision-recall and F-measure. We can see that our model outperforms the baselines. A visual comparison of the

[1] https://bitbucket.org/JianboJiao/tactilesketchsaliency

Table 1: Accuracy (%) comparison of our model (Ours) with the baselines using the point-pair based evaluation metric on both synthetic (Synthetic) and freehand (Real) sketch datasets.

	IT [21]	DSS [18]	MC [55]	SED	Ours
Synthetic	57.14	49.25	52.58	75.65	**87.39**
Real	47.95	43.80	44.63	76.45	**82.95**

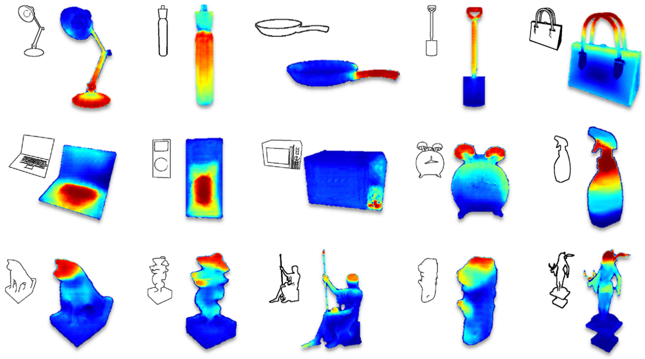


Figure 5: Tactile saliency prediction on synthetic sketches of different object categories. We show example results for grasp saliency (first row), press saliency (second row), and touch saliency (third row).

results with the above settings is shown in Figure 4(c). In addition, we have also evaluated the three methods using our point-pair based evaluation metric. For each test sketch, we set the number of sampled pairs to 100 and run the sampling process 5 times, resulting in 500 pairwise comparisons. We compute a prediction accuracy for each sketch and report an average accuracy over all the sketches in Table 1 (Synthetic). We can see that our method surpasses the baselines by a large margin. The existing visual saliency models (IT, DSS and MC) are especially designed for natural images, whose visual characteristics are quite different from sketches (without colors and rich textures), and hence have poor performances.

Qualitative Results. We show some prediction results from our model on the synthetic sketches of the three tactile modalities in Figure 1 and 5. For the grasp saliency results (first row of Figure 5), our models predicts object handles to be most salient. For the press saliency results (second row of Figure 5), the buttons on the electronic devices are predicted to be most salient. Note that our model can also predict multiple saliency parts on a single object, such as different buttons on a game controller. The touch saliency results (third row of Figure 5) show that the head or top parts of the statues are more salient than the other parts of the statues. These results show that our model can predict plausible tactile saliency for sketches of different object categories.

5.3 Evaluation on Real Sketches

We then explore how well our model can generalize to real freehand sketches.

Table 2: Prediction accuracy on real freehand sketches from 27 categories. The accuracy values for the seen categories are underlined and the overall accuracy is in bold.

Sketch Category	Acc. (%)	Sketch Category	Acc. (%)
mug	100.00	shovel	82.86
tennis racket	97.30	axe	82.22
pen	93.55	cup	81.58
wineglass	93.10	cellphone	80.00
hammer	93.02	bottle opener	79.49
toothbrush	91.43	door handle	79.07
teapot	91.30	flashlight	76.74
knife	90.91	computer-mouse	76.67
microphone	90.91	fork	76.09
spoon	89.47	floor lamp	70.59
baseball bat	84.78	screwdriver	65.71
comb	84.09	candle	61.36
cooking pan	83.33	watch	61.11
ice-cream-cone	82.98	Average	**82.95**

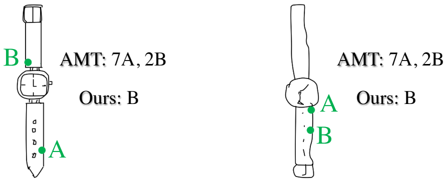


Figure 6: Our model fails to predict correctly on watch sketches. The human responses (from 9 participants) collected using AMT and our predictions are shown. Note that where to touch seems to be ambiguous even for humans in these examples. Participants tend to choose the strap to be more salient in the left example, but the dial part to be more salient in the right example.

Dataset and Evaluation Method. To evaluate our model, we have selected 27 categories of freehand sketches (Table 2), which are more relevant to tactile interaction by hand, from the TU Berlin sketch dataset [10]. The other categories are relatively ambiguous for tactile interaction, and hence difficult to evaluate quantitatively (see examples in Figure 8). For each category, we randomly select 10 sketches, resulting in a test dataset of 270 freehand sketches. Among these 27 categories in the test set, 11 appear in our training set, while the others do not.

Unlike our synthetic training data where ground truth saliency maps are known a prior, our real test dataset has no ground truth saliency maps for quantitative evaluation. Hence, we use the point-pair based metric introduced in Section 5.2 for evaluation. We have conducted a user study using Amazon Mechanical Turk (AMT). Given a sketch and a pair of labeled points (A and B) on it, the participants were asked to choose the point that they would more likely grasp, press, or touch via two alternative forced choice. Our evaluation consists of 1,350 pairs for comparison (27 categories × 10 sketches per category × 5 pairs of (A,B) points per sketch). For each

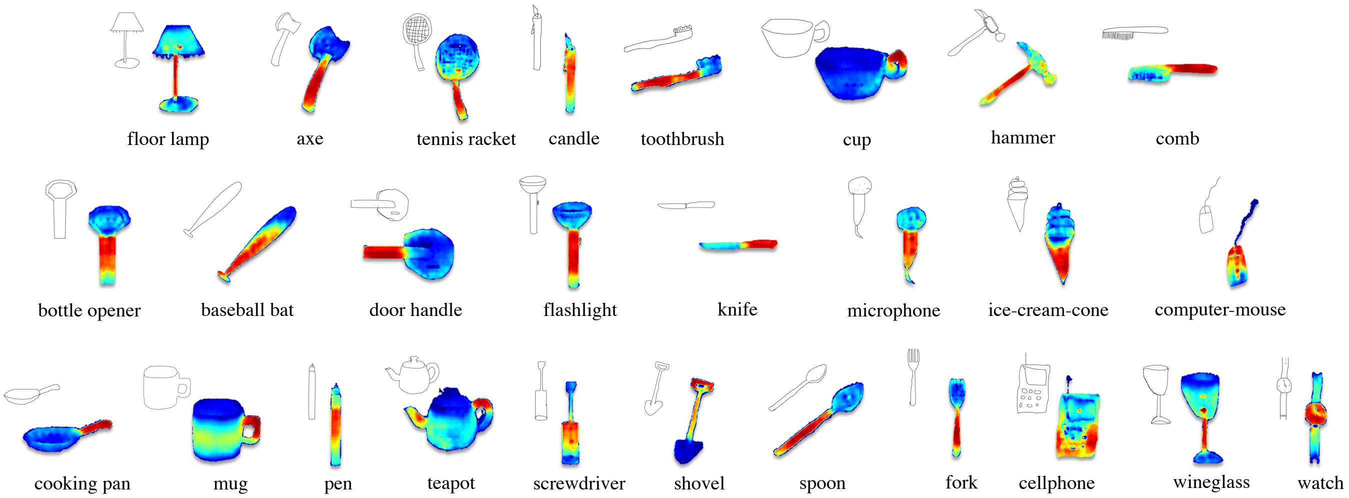


Figure 7: Tactile saliency prediction on 27 categories of freehand sketches. The 11 categories at the bottom row appear in our training dataset (*i.e.*, seen categories), while the rest do not (*i.e.*, unseen categories).

pair, we predict saliency values at A and B using our model, and collect the choices from 9 users to decide the "human ground truth" choice. We apply a *consistency check* by ignoring a pair if it does not have a high consistency among the participant choices [13] (pairs with 5-4 or 6-3 responses are ignored). We end up with 1,000 "human ground truth" cases for evaluation.

Quantitative Results. Table 2 shows the performance of our model. We can see that although our model is only trained on synthetic data, it generalizes well to real sketches, with an overall accuracy of 82.95%, and even to the unseen categories (*e.g.*, tennis racket, hammer and toothbrush). Specifically, average accuracy for the 11 seen categories and 16 unseen categories are 83.32% and 82.70%, respectively. Note that the accuracy for "watch" is relatively lower. This is mainly because there are many possible ways of grasping a watch. It is often ambiguous even for humans to decide which point is more salient. Figure 6 shows two example cases that our model predicts incorrectly for "watch". In general, our model performs well for the object categories whose tactile parts are defined clearly and consistently across instances (*e.g.*, the left part of Table 2), but less well for those with ambiguous tactile regions. We also present the performance of the baseline models in Table 1 (Real). Since our model is trained on synthetic data, it performs slightly worse on real freehand sketches whose visual appearance differs from that in our synthetic data.

Qualitative Results. Figure 7 shows example results of tactile saliency predicted by our model on real sketches. Although some sketches are rough and have distorted shapes, our model can still output good saliency maps. In particular, although 16 of the categories are unseen by our model during training, our model can still generalize well and predict plausible tactile saliency results.

5.4 Ablation Study

We perform an ablation study to investigate the importance of each design choice of our model. In particular, we compare our full model (*Ours*) with its four variants: 1) removing the skip connections (*w/o Skip*); 2) removing the depth branch (*w/o Depth*); 3) removing the cross links (*w/o CrsLk*) from the depth branch to the saliency branch; 4) removing the classification branch (*w/o Cls*).

Table 3: Ablation study on synthetic (Synthetic) and freehand (Real) sketch datasets. Performances are measured using F-Measure (FM) and the point-pair based evaluation metric (PP).

		w/o Skip	w/o Depth	w/o CrsLk	w/o Cls	Ours
Synthetic	FM	0.901	0.900	0.908	0.918	0.940
	PP	83.96%	85.30%	85.26%	87.17%	87.39%
Real	PP	77.26%	77.12%	76.96%	77.17%	82.95%

Table 3 shows the results. Since the ground truth dense saliency maps are unavailable for real sketches, FM is only reported for the synthetic data. We can see that our full model achieves the best performance. Specifically, when we remove the depth branch or the cross links, the performance drops significantly. This means that the 3D structural information of sketches plays an important role in tactile saliency prediction. In addition, if we remove the skip connections between the encoder and decoder, the performance also drops. We have empirically found that adding skip connections can result in smoother predictions qualitatively. Finally, without the classification branch, the performance declines slightly. This experiment confirms the advantage of our multi-task joint reasoning framework for tactile saliency prediction.

5.5 Generalization to Objects with Ambiguous Tactile Saliency

We are interested in knowing how well our model may perform when applied to some unseen object categories with ambiguous tactile properties such as fruits, plants and animals. Figure 8 shows

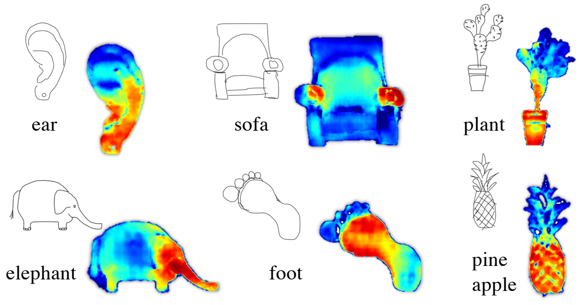


Figure 8: Objects with ambiguous tactile saliency.

the predicted results. Although humans will unlikely have a consistent view on the tactile saliency of these objects, we can see that our model can still predict some plausible tactile saliency for them.

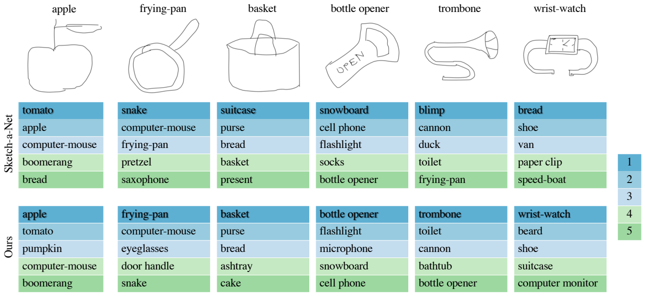
6 APPLICATIONS

6.1 Sketch Recognition

We first explore using our learned representation as features for sketch recognition. We replace the encoder in a state-of-the-art sketch recognition model, Sketch-a-Net [52], with our trained encoder and fine-tune the whole model on the training dataset of [52]. We then compare the performance of Sketch-a-Net with and without using our learned representation on the full test dataset of [52] containing 20,000 sketches from 250 categories. As we are unable to train their multi-scale architecture due to its large memory requirement, we use their single-scale architecture instead. This means that the recognition performance reported here may not be state-of-the-art. However, it can still reflect if our learned representation is useful.

The accuracy of the original Sketch-a-Net is 69.28%, which is increased to 70.65% after using our learned representation. This suggests that our representation can help improve sketch recognition. Note that the test set contains 250 categories, in which 236 categories are unseen by our model during training. We believe a higher performance gain can be expected when training our model on more categories. We have also experimented with using the learned representation by removing our saliency branch, but find that the accuracy drops (69.45%). This again confirms the importance of tactile saliency prediction in helping sketch recognition. Figure 9 shows some visual results with and without using our learned representation. We can see that access to tactile information can help resolve some ambiguous and challenging cases in sketch recognition, especially when the sketches of different categories have similar shape and structure. For example, while the original Sketch-a-Net classifies a frying pan as snake or computer mouse, our revised Sketch-a-Net with the learned representation can correctly recognize it via the additional tactile cues.

We have also tried directly feeding the concatenation of an input sketch and our predicted tactile saliency into Sketch-a-Net for sketch recognition. As we are unable to train our model on the categories outside our training dataset (due to the lack of 3D shapes and 2D tactile saliency ground truth), we directly use our model trained on our training set. Thus, there are 236 out of 250 testing categories unseen by our model. In this case, we still observe a marginal performance gain (69.75%). In addition, when we evaluate the models only on the categories that are seen by our model during training, our model achieves a much better performance (85.71%) than Sketch-a-Net (78.57%).



	apple	frying-pan	basket	bottle opener	trombone	wrist-watch
Sketch-a-Net	tomato	snake	suitcase	snowboard	blimp	bread
	apple	computer-mouse	purse	cell phone	cannon	shoe
	computer-mouse	frying-pan	bread	flashlight	duck	van
	boomerang	pretzel	basket	socks	toilet	paper clip
	bread	saxophone	present	bottle opener	frying-pan	speed-boat
Ours	apple	frying-pan	basket	bottle opener	trombone	wrist-watch
	tomato	computer-mouse	purse	flashlight	toilet	beard
	pumpkin	eyeglasses	bread	microphone	cannon	shoe
	computer-mouse	door handle	ashtray	snowboard	bathtub	suitcase
	boomerang	snake	cake	cell phone	bottle opener	computer monitor

Figure 9: Sketch recognition results. Top: input sketches. Middle: results of the Sketch-a-Net. Bottom: results of Sketch-a-Net augmented by our learned representation. For each method, we show the top 5 predictions (from top to bottom).

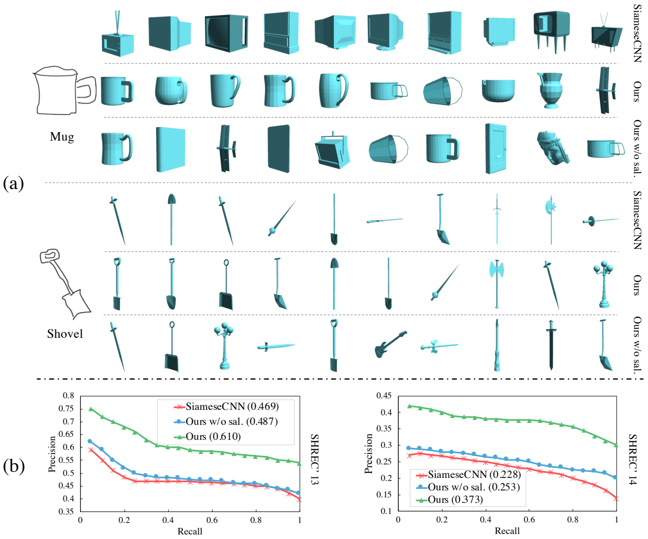


Figure 10: Sketch-based 3D shape retrieval using the SiameseCNN representation and our learned representation. (a) Qualitative results where query sketches are shown on the left and the top 10 retrieved shapes on the right by SiameseCNN (top), Ours (middle), and Ours without the saliency branch (bottom). The retrieved shapes are ranked from left (more similar) to right (less similar). (b) Quantitative comparison where precision-recall curves on SHREC'13 and '14 are plotted, with mean average precision (mAP) scores shown in the brackets.

6.2 Sketch-based 3D Shape Retrieval

We have also investigated if our learned sketch representation is useful for sketch-based 3D shape retrieval. Given a query 2D sketch,

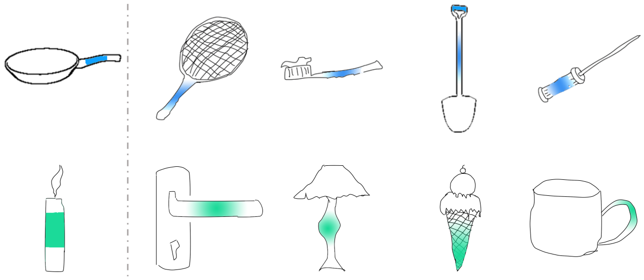


Figure 11: Part-based functional correspondence. The input sketches with colored query regions are shown on the left, while the functionally similar regions found in other sketches are shown on the right.

we need to find the corresponding 3D shape from a database of 3D shapes. Following previous methods [11, 46, 49], we project each 3D shape to 2D sketches from multiple viewpoints, and use our representation to compute the similarity between a query sketch and the 2D projections.

For comparison, we choose the representation from a CNN model (SiameseCNN), which is used in a recent sketch-based 3D shape retrieval method [46]. For a fair comparison, both representations are integrated into the multi-view projection framework proposed by [46] and tested on SHREC'13 [25] and SHREC'14 [26]. Figure 10(a) shows the qualitative results. Compared with SiameseCNN, our representation can take advantage of the tactile information in both the query sketch and 2D projections of the 3D shapes (*e.g.*, where to grasp a shovel) for similarity computation, and thus retrieve more relevant 3D shapes even though some of the 3D shapes may look somewhat different from the query sketch. We have also tried to retrieve using our learned representation without the saliency branch, which gives worse results. This shows that learning tactile sketch saliency is crucial to extracting effective features for this task. Figure 10(b) shows the corresponding quantitative comparison. We can see that our full model outperforms the other two by a large margin.

6.3 Part-based Functional Correspondence

Finally, we have explored using the predicted tactile sketch saliency to find functionally similar regions across different sketches. This problem is particularly challenging since it involves matching the regions that may have different geometric shapes across different sketches with varying global structures. Given a query region on a sketch, the goal is to find functionally similar regions on other sketches. We do this by first computing the saliency values of the query region as well as the saliency values on other sketches, and normalize them to be in [0, 1]. We can then identify the functionally similar region in a sketch by sliding a window over the sketch to find the best matching region to the query region, measured by pixel-wise saliency difference.

Figure 11 shows some results. The top row shows the correspondence among sketches across different categories, while the bottom row shows that even for sketches whose functionally corresponding parts have quite different shapes, our method can still give promising results. This experiment demonstrates that with the help of our tactile sketch saliency, part-based correspondence across different sketches can be established in a functionally meaningful way. It is worth noting that since our method is purely based on tactile sketch saliency, it can only find correspondences among regions with high tactile saliency (*i.e.*, regions likely to be grasped/pressed/touched). We believe that more sophisticated methods are needed for this problem, which is left as future research.

7 CONCLUSION

In this paper, we have investigated the tactile sketch saliency problem, and learned a deep model for predicting tactile saliency on input sketches using just synthetically generated training data. We have demonstrated that our model can achieve promising results on both synthetic and real sketches, and have a real impact on many sketch analysis problems. In addition to the depth and classification tasks used in our model, other types of tasks may also be helpful to our problem. For example, surface normal estimation can provide structure information as well. However, it is difficult to enumerate all possible tasks in a single paper. Thus, in this paper, we explore the two auxiliary tasks that can provide the information we need, and leave investigating the effect of other possible tasks as an interesting direction to explore in the future. As a future work, we would also like to explore a wider range of applications of our model in sketch analysis. For example, our model can be integrated into a sketch abstraction model [39] to help ensure that abstracted sketches are both recognizable and functionally valid. We envision our work as just a first step towards the challenging goal of understanding the functional properties of sketches, which could enable a human-centric and physical interpretation of 2D sketches. We hope that our work will inspire the community to further explore this direction.

ACKNOWLEDGMENTS

Manfred Lau acknowledges the Hong Kong Research Grants Council (General Research Fund numbers 11206319 and 11205420).

REFERENCES

[1] Yoshua Bengio et al. 2009. Learning deep architectures for AI. *Foundations and Trends in Machine Learning* 2, 1 (2009), 1–127.

[2] Mikhail Bessmeltsev, Nicholas Vining, and Alla Sheffer. 2016. Gesture3D: posing 3D characters via gesture drawings. *ACM TOG* 35, 6 (2016).

[3] Ali Borji. 2019. Saliency Prediction in the Deep Learning Era: Successes and Limitations. *IEEE TPAMI* (2019).

[4] Ali Borji, MingMing Cheng, Huaizu Jiang, and Jia Li. 2015. Salient Object Detection: A Benchmark. *IEEE TIP* 24, 12 (2015), 5706–5722.

[5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

[6] Jiaxin Chen, Jie Qin, Li Liu, Fan Zhu, Fumin Shen, Jin Xie, and Ling Shao. 2019. Deep Sketch-Shape Hashing with Segmented 3D Stochastic Viewing. In *Proc. CVPR*.

[7] MingMing Cheng, Niloy Mitra, Xiaolei Huang, Philip Torr, and ShiMin Hu. 2015. Global contrast based salient region detection. *IEEE TPAMI* 37, 3 (2015), 569–582.

[8] Guoxian Dai, Jin Xie, and Yi Fang. 2018. Deep correlated holistic metric learning for sketch-based 3d shape retrieval. *IEEE TIP* 27, 7 (2018), 3374–3386.

[9] Anjan Dutta and Zeynep Akata. 2019. Semantically Tied Paired Cycle Consistency for Zero-Shot Sketch-based Image Retrieval. In *Proc. CVPR*.

[10] Mathias Eitz, James Hays, and Marc Alexa. 2012. How do humans sketch objects? *ACM TOG* 31, 4 (2012).

[11] Mathias Eitz, Ronald Richter, Tamy Boubekeur, Kristian Hildebrand, and Marc Alexa. 2012. Sketch-based shape retrieval. *ACM TOG* 31, 4 (2012).

[12] Dashan Gao and Nuno Vasconcelos. 2007. Bottom-up saliency is a discriminant process. In *Proc. ICCV*.

[13] Elena Garces, Aseem Agarwala, Diego Gutierrez, and Aaron Hertzmann. 2014. A similarity measure for illustration style. *ACM TOG* 33, 4 (2014).

[14] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Networks. *Aistats* 15, 106 (2011).

[15] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. 2012. Context-aware saliency detection. *IEEE TPAMI* 34, 10 (2012).

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. ICCV*.

[17] Shengfeng He and Rynson Lau. 2016. Exemplar-Driven Top-Down Saliency Detection via Deep Association. In *Proc. CVPR*.

[18] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. 2019. Deeply Supervised Salient Object Detection with Short Connections. *IEEE TPAMI* 41, 4 (2019), 815–828.

[19] Qibin Hou, Ming-Ming Cheng, Xiao-Wei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. 2017. Deeply supervised salient object detection with short connections. In *Proc. CVPR*.

[20] Zhe Huang, Hongbo Fu, and Rynson Lau. 2014. Data-driven segmentation and labeling of freehand sketches. *ACM TOG* 33, 6 (2014).

[21] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI* 20, 11 (1998), 1254–1259.

[22] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *Proc. ICCV*.

[23] Peter Kaiser and Tamim Asfour. 2018. Autonomous Detection and Experimental Validation of Affordances. *IEEE Robotics and Automation Letters* 3, 3 (2018), 1949–1956.

[24] Manfred Lau, Kapil Dev, Weiqi Shi, Julie Dorsey, and Holly Rushmeier. 2016. Tactile mesh saliency. *ACM TOG* 35, 4 (2016).

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

[26] B. Li, Yijuan Lu, Chunyuan Li, Afzal Godil, Tobias Schreck, Masaki Aono, Qiang Chen, Nihad Karim Chowdhury, Bin Fang, Takahiko Furuya, Henry Johan, Ryuichi Kosaka, Hitoshi Koyanagi, Ryutarou Ohbuchi, and Atsushi Tatsuma. 2014. SHREC'14 Track: Large Scale Comprehensive 3D Shape Retrieval. In *Eurographics Workshop on 3D Object Retrieval*. .

[27] Changjian Li, Hao Pan, Yang Liu, Xin Tong, Alla Sheffer, and Wenping Wang. 2017. Bendsketch: Modeling freeform surfaces through 2d sketching. *ACM TOG* 36, 4 (2017), 125.

[28] Guanbin Li and Yizhou Yu. 2015. Visual saliency based on multiscale deep features. In *Proc. CVPR*.

[29] Ke Li, Kaiyue Pang, Jifei Song, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Honggang Zhang. 2018. Universal sketch perceptual grouping. In *Proc. ECCV*.

[30] Chenxi Liu, Enrique Rosales, and Alla Sheffer. 2018. Strokeaggregator: Consolidating raw sketches into artist-intended curve drawings. *ACM TOG* 37, 4 (2018), 1–15.

[31] Fang Liu, Xiaoming Deng, Yu-Kun Lai, Yong-Jin Liu, Cuixia Ma, and Hongan Wang. 2019. SketchGAN: Joint Sketch Completion and Recognition with Generative Adversarial Network. In *Proc. CVPR*.

[32] Nian Liu, Junwei Han, and Ming-Hsuan Yang. 2018. PiCANet: Learning pixel-wise contextual attention for saliency detection. In *Proc. CVPR*.

[33] Risheng Liu, Junjie Cao, Zhouchen Lin, and Shiguang Shan. 2014. Adaptive partial differential equation learning for visual saliency detection. In *Proc. CVPR*.

[34] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and HeungYeung Shum. 2011. Learning to detect a salient object. *IEEE TPAMI* 33, 2 (2011), 353–367.

[35] Austin Myers, Ching Lik Teo, Cornelia Fermüller, and Yiannis Aloimonos. 2015. Affordance detection of tool parts from geometric features.. In *Proc. ICRA*.

[36] Luke Olsen, Faramarz Samavati, Mario Costa Sousa, and Joaquim Jorge. 2009. Sketch-based modeling: A survey. *Computers & Graphics* 33, 1 (2009), 85–103.

[37] Sarthak Parui and Anurag Mittal. 2014. Similarity-invariant sketch-based image retrieval in large databases. In *Proc. ECCV*.

[38] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei Efros. 2016. Context encoders: Feature learning by inpainting. In *Proc. CVPR*.

[39] Umar Riaz Muhammad, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. 2018. Learning deep sketch abstraction. In *Proc. CVPR*.

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

[41] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. 2014. SceneGrok: Inferring action maps in 3D environments. *ACM TOG* 33, 6 (2014), 212.

[42] Rosália Schneider and Tinne Tuytelaars. 2014. Sketch classification and classification-driven analysis using fisher vectors. *ACM TOG* 33, 6 (2014).

[43] Rosália Schneider and Tinne Tuytelaars. 2016. Example-Based Sketch Segmentation and Labeling Using CRFs. *ACM TOG* 35, 5 (2016).

[44] Tianjia Shao, Dongping Li, Yuliang Rong, Changxi Zheng, and Kun Zhou. 2016. Dynamic furniture modeling through assembly instructions. *ACM TOG* 35, 6 (2016).

[45] Giorgos Tolias and Ondrej Chum. 2017. Asymmetric feature maps with application to sketch based retrieval. In *Proc. CVPR*.

[46] Fang Wang, Le Kang, and Yi Li. 2015. Sketch-based 3d shape retrieval using convolutional neural networks. In *Proc. CPVR*.

[47] Lijun Wang, Huchuan Lu, Xiang Ruan, and MingHsuan Yang. 2015. Deep networks for saliency detection via local estimation and global search. In *Proc. CVPR*.

[48] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. 2018. Detect globally, refine locally: A novel approach to saliency detection. In *Proc. CVPR*.

[49] Kun Xu, Kang Chen, Hongbo Fu, WeiLun Sun, and ShiMin Hu. 2013. Sketch2Scene: Sketch-based co-retrieval and co-placement of 3D models. *ACM TOG* 32, 4 (2013).

[50] Peng Xu, Yongye Huang, Tongtong Yuan, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, Zhanyu Ma, and Jun Guo. 2018. Sketchmate: Deep hashing for million-scale human sketch retrieval. In *Proc. CVPR*.

[51] Qian Yu, Feng Liu, YiZhe Song, Tao Xiang, Timothy Hospedales, and ChenChange Loy. 2016. Sketch me that shoe. In *Proc. CVPR*.

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

[53] Philipp Zech, Simon Haller, Safoura Rezapour Lakani, Barry Ridge, Emre Ugur, and Justus Piater. 2017. Computational models of affordance in robotics: a taxonomy and systematic classification. *Adaptive Behavior* 25, 5 (2017), 235–271.

[54] Hua Zhang, Si Liu, Changqing Zhang, Wenqi Ren, Rui Wang, and Xiaochun Cao. 2016. Sketchnet: Sketch classification with web images. In *Proc. CVPR*.

[55] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. 2015. Saliency detection by multi-context deep learning. In *Proc. CVPR*.

[56] Ting Zhao and Xiangqian Wu. 2019. Pyramid Feature Attention Network for Saliency Detection. In *Proc. CVPR*.

[57] Changqing Zou, Qian Yu, Ruofei Du, Haoran Mo, Yi-Zhe Song, Tao Xiang, Chengying Gao, Baoquan Chen, and Hao Zhang. 2018. Sketchyscene: Richly-annotated scene sketches. In *Proc. ECCV*.