Finding Badly Drawn Bunnies

Lan Yang\textsuperscript{1,2}  Kaiyue Pang\textsuperscript{2}  Honggang Zhang\textsuperscript{1}\textsuperscript{*}  Yi-Zhe Song\textsuperscript{2}

\textsuperscript{1} PRIS, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China
\textsuperscript{2} SketchX, CVSSP, University of Surrey, United Kingdom

\{ylan, zhhg\}@bupt.edu.cn, \{kaiyue.pang, y.song\}@surrey.ac.uk

Abstract

As lovely as bunnies are, your sketched version would probably not do it justice (Fig. 1). This paper recognises this very problem and studies sketch quality measurement for the first time – letting you find these badly drawn ones. Our key discovery lies in exploiting the magnitude (L\textsuperscript{2} norm) of a sketch feature as a quantitative quality metric. We propose Geometry-Aware Classification Layer (GACL), a generic method that makes feature-magnitude-as-quality-metric possible and importantly does it without the need for specific quality annotations from humans. GACL sees feature magnitude and recognisability learning as a dual task, which can be simultaneously optimised under a neat cross-entropy classification loss. GACL is lightweight with theoretical guarantees and enjoys a nice geometric interpretation to reason its success. We confirm consistent quality agreements between our GACL-induced metric and human perception through a carefully designed human study. Notably, we demonstrate three practical sketch applications enabled for the first time using our quantitative quality metric.

1. Introduction

Everybody \textit{can} sketch, the debate is on \textit{how well}. With the proliferation of touchscreen devices, the urge to sketch is ever more pronounced. This is not to mention the broad range of sketch-enabled applications – from recognition \[10,26,46,64\], parsing \[24,45,63\], recreating \[8,16,28,50\], to leveraging sketch as a query modality for image search \[3,29,38,42,43\] and visual content manipulation \[5,20,57,69\], or even enabling a drawing agent that excels human at a Pictionary-like sketching game \[2\].

This paper recognises this very “how well” problem and proposes a learnable metric that for the first time tells us just \textit{how badly} drawn my bunny (or any other sketch) is – so from the collection of bunny sketches in Fig. 1(a) to an ordered list of bunnies from worse to best in Fig. 1(b). As interesting as the problem sounds in its own right, it also underpins many facets of sketch research at large. These include but not limited to (i) disentangling human factor in model prediction, \textit{i.e.} whether the model is bad or the sketch is; (ii) facilitating better representation learning, \textit{i.e.} making data-driven models less prone to overfit by learning against specific quality level; (iii) sketching assistance applications, \textit{e.g.} helping users to move towards a better bunny.

Quantifying sketch quality is non-trivial. The first obstacle is the fatal lack of existing sketch datasets annotated with human quality ratings. This essentially renders most of the recent works on image quality assessment that predicts human opinion scores inapplicable \[31,47,52,68\]. The sketch-specific trait as a vector representation of sequential coordinates also sets it apart from another line of works on trying to model human-interpretable image quality directly from low-level statistical distortions \[1,21,58,66\] – compared with visual artefacts such as noise, blur and compression, sketch quality is more of a subjective interpretation of holistic visual concepts.

Figure 1. (a) Not every free-hand bunny sketch is of equal quality. (b) We contribute an annotation-free solution (GACL) for discriminating quality between bunny (and many other category) sketches. We show quality discovery under GACL supports reasonable level of quality examination from humans.

*corresponding author
In this paper, we provide the first stab at learning a sketch-specific quality score (metric) without the reliance on human quality annotations. Core to our technical solution is the insight that such a score is inherent to the feature space geometry underlying a recognition task (Fig. 6). We first single out the orderless feature geometry learned off a typical Softmax formulation (Eq. 4) to be the main culprit for the failed quality discovery. This is because Softmax will constantly push sketch features to be close to the class centre and thus undermines any potential geometry formation. The intuition is then that a real-valued feature magnitude can already induce a quality metric, if the underlying feature geometry can satisfy the following property: the better quality a sketch, the closer its feature to the class centre. To encourage the integrity of learned feature magnitude, we enforce its optimisation to be convex and that a global optima is guaranteed. We also show that under mild mathematical approximation, sketch quality score under GACL is a de facto margin value to the class decision boundary, yielding a nice semantics to the quality defined (that the better the quality, the farther away a sketch is from its class decision boundary). We develop four specific instantiations of GACL and conduct human study respectively to provide some assurance on the ordering of the quality scores learned. Asking humans to devise a strictly-pairwise global ordering is however not feasible [12, 25], especially when sketch quality perception can be highly subjective. For that, we resort to the psychology literature [11, 15, 51] and adopt a set-based approach where human gets to rank coarsely at set-level other than individual sketches. Results off 12,800 trials (40 participants each doing 320 trials) have human agreeing with the learned quality ordering 92.61% of the time on average across 8 carefully selected sketch categories.

Importantly, we showcase the practical benefits of modelling sketch quality in three applications: (i) quality-aware sketch recognition that contributes the new state-of-the-art recognition performance; (ii) quality-guided sketch generation that pushes the envelope of sketch manipulation task beyond generating conceptually correct sketches; (iii) quality-enabled sketch attribution that helps sketch practitioners to identify malicious user input.

2. Related Work

Sketch research. Apart from constantly raising the performance bar on various sketch perceptual tasks, recent computer vision works for human sketch data have been additionally focusing on two unique aspects: (i) pixel/vector dichotomy: should sketch be processed as raster pixel image [39, 44, 57, 64] or vector graphic [16, 26, 36, 49] compiled as a sequence of points, or the combination of both [50, 53, 61, 62]? Current explorations suggest that better performance is often obtained when the two modalities are encoded cooperatively as one unified representation for either generative or discriminative task. (ii) “can’t sketch” reality: unlike clicking a tag or typing a search keyword, sketching is a slow and skillful process. Users can be worrying about inaccurate results because of their poor renderings and consequently not motivated enough to sketch at the first place. Existing solutions include allowing users to stop early in a sketching episode so that their goals can be achieved with earliest/easiest strokes [2, 3, 23] or a real-time drawing assistant that lowers rendition barriers [34, 48, 60]. We study a new sketch problem of computational quality modelling, which can potentially benefit many ongoing sketch research – from improving discriminative performance (Sec. 4.2) to introducing a beautification objective into existing generative models (Sec. 4.3).

Image quality assessment. Existing literature on image quality assessment (IQA) draws a distinction between approaches that require an input reference and those do not. Referenced-based algorithms [18, 21, 66] assume the availability of pristine and distorted image pairs so that the quality gap can be measured, where the no-reference or blind IQA [59, 65, 67] loosens the pairing constraint by instead exploiting from a carefully curated image set processed with several known fixed distortion types (e.g. noise, blurring, corruptions and compression artefacts). One particular line of blind IQA works [13, 33, 52, 68] is how to accurately predict subjective human quality ratings provided by datasets like AVA [37] and LIVE [14], which is not applicable here due to the lack of a similarly annotated dataset. We too approach sketch quality assessment as a blind IQA problem and propose a novel solution by leveraging sketch feature magnitude as a promising quality metric to bypass the laborious and expensive human annotation step.

Margin-based learning. Margin is an important concept for representation learning before the deep learning wave (e.g. SVM [7] is also known as soft margin classifier), and even more so when deep learning sweeps across computer vision fields today (e.g. contrastive [17] or triplet ranking loss [56]). Most relevant to ours is the idea of encapsulating margin into a Softmax-based classification model. By modifying the vanilla Softmax via the insertion of an either fixed or adaptive margin, many representative Softmax variants have been proposed [9, 27, 30, 35, 44, 55] to boost feature discriminativeness with the same goal of ensuring within-class variation is smaller than between-class difference. We have shown analytically that quality score learned
3. Methodology

The goal of this paper is to obtain a score-based metric $q(\cdot)$ that quantifies sketch quality. Given sketch sample $x_i$ with category label $y_i \in \{1, 2, \ldots, C\}$, our key finding is that, during the training of a sketch recognition network $f(\cdot)$ and under certain mild conditions, sketch feature magnitude ($L_2$ norm) can automatically encode the computational metric needed for quality discrimination, i.e., $q_i \equiv q(x_i) = ||f(x_i)||_2$. We will first introduce necessary preliminaries before describing our proposed method for $q(\cdot)$ to be a good proxy for quality discovery.

3.1. Preliminaries and Discussions

In a conventional Softmax-based classification layer, the training objective for a sample $x_i$ being classified as its ground truth category $y_i$ is formulated as:

$$L_{sm}(x_i) = -\log \frac{e^{W_{y_i}^T f(x_i) + b_{y_i}}}{\sum_{j=1 \atop j \neq y_i}^C e^{W_{y_j}^T f(x_i) + b_{y_j}}}$$

(1)

where $f(x_i) \in \mathbb{R}^d$ is the extracted deep feature of the i-th sketch sample belonging to the class $y_i$. $W \in \mathbb{R}^{d \times C}$ denotes the weights of all $C$ class centres with $B \in \mathbb{R}^C$ as bias terms. We transform $W_{y_j}^T f(x_i)$ to $||W_{y_j}|| ||f(x_i)|| \cos \theta_{i,y_j}$ where $\theta_{i,y_j}$ is the angle (i.e. cosine distance) between $f(x_i)$ and $W_{y_j}$. For ease of analyse, we further eliminate the bias term and set $||W_{y_j}||$ to 1. This gives us a modified Softmax formulation as follows:

$$\bar{L}_{sm}(x_i) = -\log \frac{e^{||f(x_i)|| \cos \theta_{i,y_i}}}{\sum_{j=1 \atop j \neq y_i}^C e^{||f(x_i)|| \cos \theta_{i,y_j}}}$$

(2)

Assume that each class has the same number of samples and that all samples are well-separated, we could obtain the lower bound of $\bar{L}_{sm}$ as (details in the supplementary):

$$\bar{L}_{sm} \geq \log(1 + (C - 1)e^{-\frac{C}{1 ||f(x_i)||}})$$

(3)

Astute readers may already notice the catastrophic implication under the loss function $\bar{L}_{sm}$: the optimisation process can be dominated towards maximising $||f(x_i)||$ and completely independent of $\theta$ at its worst, derailing from the very goal of categorisation. Indeed, Eq. 3 tells us that the minimisation process of $\bar{L}_{sm}(x_i)$ can take place on $||f(x_i)||$ only. Solving this problem, however, requires more than an naive unit normalisation of $||f(x_i)||$. To see this clearer, imagine the extreme ideal case where a sample is infinitely close to its centre. As such, the gradient of $\bar{L}_{sm}(x_i)$ w.r.t the ground-truth label $y_i$ is $1 - \frac{e^1}{e^1 + (C-1)e^0}$ (0.931 when $C=100$, 0.993 when $C=1000$), which means that the model will undesirably back-propagate large gradients even when samples are well separated. To get around this seemingly opposing role of feature magnitude, similar compromise is often undertook, where $||f(x_i)||$ is first cancelled out from the formulation (i.e. $||f(x_i)|| = 1$) and replaced with a global scalar $s$ to simulate its critical effect for numerical stability under cross-entropy loss optimisation. We are now ready to write down a normalised version of Softmax as enjoyed by many existing works [9, 30, 54, 55]:

$$L_{norm-sm} = -\log \frac{e^s \cos \theta_{i,y_i}}{\sum_{j=1 \atop j \neq y_i}^C e^s \cos \theta_{i,y_j}}$$

(4)

where the exact value of $s$ is empirically set.

One problem with Eq. 4 is its inclination to treat every sketch sample equally recognisable – all $\cos \theta_{i,y_i}$ is optimised towards the same optimum of being as close to the class centre as possible. This loss of instance discrimination comes against to how we perceive human sketch data in practice, where people can draw dramatically different bunnies while retaining recognisability (Fig. 1). These bunnies are certainly not of equal quality, and nor should their feature distances be the same to the class centre. A natural question to ask is then that instead of over-simplifying the role of feature magnitude $||f(\cdot)||$ to a constant scalar, can we exploit it to encourage the establishment of quality semantics within the same class so that $\cos \theta_{i,y_i} > \cos \theta_{j,y_i}$ when $x_i$ is of significantly better quality than $x_j$? We give an affirmative answer to this question. We show by carefully tuning the interplay between $||f(\cdot)||$ and $\cos \theta$ into a unified framework (Sec. 3.2), $||f(\cdot)||$ promotes a quality-aware feature geometry space and that turns itself into a promising quality indicator as verified in our empirical evaluation.

3.2. Geometry-Aware Classification Layer

Eq. 4 presents a magnitude-agnostic classification loss. We aim to inject feature magnitude $||f(x_i)||$ as a learnable variable into Eq. 4 so that it adaptively works with the classification objective $\cos \theta_{i,y_i}$, and consequently induces an instance-discriminative feature space geometry that permits quality discovery. For that, we introduce a new formulation upon Eq. 4 by replacing $s \cos \theta_{i,y_i}$ with a compound function $A(q_i, \theta_{i,y_i})$.

We name it Geometry-Aware Classification Layer (GACL). Denoting $\sum_{j=1 \atop j \neq y_i}^C e^{s \cos \theta_{i,y_j}}$ as $R$, 

\[1\] For notation simplicity, we use $q_i$ and $\theta_{i,y_i}$ to represent $||f(x_i)||$ and $\theta_{i,y_i}$, respectively.
GACL transforms Eq. 4 to:

\[ L_{\text{GACL}}(q_i, \theta_y) = -\log \frac{e^{A(q_i, \theta_y)}}{e^{A(q_i, \theta_y)} + R} \]  

The success of GACL thus relies critically on the design choice of \( A(q_i, \theta_y) \), for which we define three necessary constraints for its success.

**Geometry constraint.** If \( q_i \) is a good proxy for quality measurement, it should be larger than \( q_j \) when \( \theta_y \), is geometrically lying closer to the class centre than that of \( \theta_{y_j} \), i.e. \((q_i - q_j)(\theta_{y_i} - \theta_{y_j}) \leq 0\).

**Condition on \( A(q_i, \theta_y) \).** Given two sketches with different value pairs of \((q_i, \theta_{y_i})\) and \((q_j, \theta_{y_j})\), we assume that both has reached optimal recognisability/optimisation equilibrium \( L_{\text{GACL}}(q_i, \theta_y) = L_{\text{GACL}}(q_j, \theta_y) \). We perform a Taylor expansion of the left-hand side:

\[ L_{\text{GACL}}(q_i, \theta_y) \approx L_{\text{GACL}}(q_j, \theta_y) + (q_i - q_j) \nabla_q L_{\text{GACL}} + (\theta_{y_i} - \theta_{y_j}) \nabla_\theta L_{\text{GACL}} \]  

where we have dropped higher order terms. To ensure \((q_i - q_j)(\theta_{y_i} - \theta_{y_j}) \leq 0\), it is then easy to obtain the condition to which \( A(q_i, \theta_y) \) must satisfy:

\[ \frac{\nabla_q A(q_i, \theta_y)}{\nabla_\theta A(q_i, \theta_y)} > 0 \]  

**Optimality constraint.** Assuming the value range of \( q_i \) is bounded in \([l_q, u_q]\), we require that \( L_{\text{GACL}} \) always has an optimal solution \( q^*_i \) between \([l_q, u_q]\) in order to prescribe a valid quality metric.

**Condition on \( A(q_i, \theta_y) \).** We assume \( L_{\text{GACL}} \) is a convex function of \( q_i \) (i.e. \( \nabla^2_q L_{\text{GACL}}(q_i, \theta_y) \geq 0 \Rightarrow \nabla^2_q L_{\text{GACL}}(q_i, \theta_y) \leq 0 \)), which naturally yields to a global optima. The existence of an optimal solution \( q^*_i \) in \([l_q, u_q]\) then translates to the following condition to be held:

\[ \nabla_q L_{\text{GACL}}(l_q, \theta_y) < 0 \text{ and } \nabla_q L_{\text{GACL}}(u_q, \theta_y) > 0 \]  

Since the value of \( -\frac{R}{e^{A(q_i, \theta_y)} + R} \) is bounded under two specific values \([l_q, u_q]\) and calls for more efforts to satisfy. Rather than handcrafting the possible values of \([l_q, u_q]\) in the lens of microscope, we propose a more principled strategy by introducing a score regulariser \( G(q_i) \):

\[ \nabla_q L_{\text{GACL}} = -\frac{R}{e^{A(q_i, \theta_y)} + R} \nabla_q A(q_i, \theta_y) + \lambda_g \nabla_q G(q_i) \]  

Since the value of \( -\frac{R}{e^{A(q_i, \theta_y)} + R} \) always remains positive in all instantiations, we simply need to set \( \nabla_q G(u_q) = 0 \) to meet \( \nabla_q L_{\text{GACL}}(u_q, \theta_y) > 0 \). We implement \( G(q_i) \) as \( \frac{1}{q_i} + \frac{l_q}{u_q} q_i \) and then focus on achieving \( \nabla_q L_{\text{GACL}}(l_q, \theta_y) < 0 \) for each instantiation scenario in the following discussion.

\[ A(q_i, \theta_y) = (1 - q_i) s \cos \theta_y \]  

Rewriting Eq. 10 gives us:

\[ \nabla_q L_{\text{GACL}} = \frac{R s \cos \theta_y}{e^{A(q_i, \theta_y)} + R} + \lambda_g \nabla_q G(q_i) \]  

We know \( 0 < \frac{R s \cos \theta_y}{e^{A(q_i, \theta_y)} + R} < 1 \). It is then sufficient to ensure \( \nabla_q L_{\text{GACL}}(l_q, \theta_y) < 0 \) if \( \lambda_g \nabla_q G(l_q) < -s \). And since \( \nabla_q G(l_q) = -\frac{1}{q_i^2} - \frac{1}{u_q^2} \), we conclude by requiring \( \lambda_g > -\frac{s^2 u_q^2}{l_q^2 - u_q^2} \). We set \( l_q = 0.1, u_q = 0.3, s = 64 \) in our implementation.

---

2We apply a linear scaling on \( q_i \) in practice to make it work in the proper value range \([l_q, u_q]\), which is omitted here for simplicity.
A(qi, θyi) = s cos (qiyi). Rewriting Eq. 10 gives us:
\[ \nabla_q L_{GACL} = \frac{R s}{e^{A(qi, \theta_{yi})} + R} \sin(qiyi) + \lambda_g \nabla_q G(qi) \]

(12)

We consider similar analysis as above, where given 0 < \( \frac{R \sin(qiyi)}{e^{A(qi, \theta_{yi})} + R} < 1 \), we enforce \( \lambda_g \nabla_q G(l_q) < -\frac{s \pi}{2} \Rightarrow \lambda_g > -\frac{s \pi u_{qy}^2}{2(1 - u_{qy}^2)} \) to meet \( \nabla_q L_{GACL}(l_q, \theta_{yi}) < 0 \). We set \( l_q = 1.1, u_{qy} = 1.25, s = 64 \) in our implementation.

\[ A(qi, \theta_{yi}) = s \cos (\theta_{yi} + q_i) \]
Rewriting Eq. 10 gives us:
\[ \nabla_q L_{GACL} = \frac{R s}{e^{A(qi, \theta_{yi})} + R} \sin(\theta_{yi} + q_i) + \lambda_g \nabla_q G(qi) \]

(13)

Similarly, we require \( \lambda_g \nabla_q G(l_q) < -s \Rightarrow \lambda_g > -\frac{s \pi^2 u_{qy}^2}{2(1 - u_{qy}^2)} \) to meet \( \nabla_q L_{GACL}(l_q, \theta_{yi}) < 0 \). We set \( l_q = 0.45, u_{qy} = 0.65, s = 64 \) in our implementation.

\[ A(qi, \theta_{yi}) = s \cos \theta_{yi} - q_i \]
Rewriting Eq. 10 gives us:
\[ \nabla_q L_{GACL} = \frac{R s}{e^{A(qi, \theta_{yi})} + R} + \lambda_g \nabla_q G(qi) \]

(14)

Similarly, we require \( \lambda_g \nabla_q G(l_q) < -1 \Rightarrow \lambda_g > -\frac{s \pi^2 u_{qy}^2}{2(1 - u_{qy}^2)} \) to meet \( \nabla_q L_{GACL}(l_q, \theta_{yi}) < 0 \). We set \( l_q = 0.35, u_{qy} = 0.8, s = 64 \) in our implementation.

3.4. Demystifying qi, as Quality Metric

In this section, we provide a different perspective to the role of qi, which comes with a new mathematical interpretation: Under mild approximation, qi is the feature distance to class decision boundary, echoing well as a dual task with \( \theta_{yi} \) (i.e. co-optimisation), which is the feature distance to class centre. Quality discrimination is then encoded in qi from such geometrical semantic establishment, as illustrated in Fig. 2. To see clearly how qi represents the distance to the decision boundary, we first review how Softmax is derived as a classification objective. A general formulation to classify an instance \( x_i \) among C classes is:
\[ \max_{j \neq i} \left( \cos \theta_{y_j} - \cos \theta_{yi} \right) \]

(15)

which is the raw “hardmax” implying that the target logit score should be greater than the rest. By smoothing the two max functions with mathematical approximations, we arrive at the normalised softmax in Eq. 4. The problem with Eq. 15 is that it completely disregards the within-class feature distribution, where samples are treated equally so long as they belong to the same class label, and thus undermines any potential quality discovery. Our assumption is that samples with better quality should be pulled farther away from

\( \frac{\log(1 + e^{\log(1 + e^{\cos \theta_{y_j} - \cos \theta_{yi} - m_j})})}{e^{\cos \theta_{yi} - m_i} + \sum_{j=1, j \neq i}^C e^{\cos \theta_{y_j}}} \)

(17)

Now we can see that apart from global normalisation term \( s \), Eq. 17 is exactly our GACL framework with \( A(qi, \theta_{yi}) \) instantiated as \( s \cos \theta_{yi} - q_i \), where feature magnitude qi becomes \( m_i \). We omit the elaborations for proving other three \( A(qi, \theta_{yi}) \) instantiations and believe the discussions above can serve our purpose in providing intuitions on why GACL permits quality discovery: the synergistic interplay between qi and \( \theta_{yi} \) yields the feature space geometry that importantly gives rise to the notion of quality.

4. Experiments

Settings. We evaluate our approach on the largest human free-hand sketch dataset to date, QuickDraw [16], which is collected via an online game and where the players are asked to sketch a given category name in less than 20 seconds. QuickDraw contains 345 object categories with each containing 70k, 2.5k, 2.5k samples for training, validation and testing respectively. We follow the tradition [41, 62] of using 7k samples per category for training and all testing data for evaluation (862k sketches in total). We im-
implement $f(\cdot)$ as a two-layer BiLSTM [19] with 1024 hidden units, and classification head $W$ with MLPs of dimension 2048-1024-345. Adam [22] optimiser is adopted with initial learning rate 1e-3 and a per-epoch cosine annealing schedule for gradient warm restarts [32]. We train each individual trial for 10 epochs with a batch size of 256, and pre-process vector sketch data to absolute coordinates normalised within range [0, 1]. Lastly, we denote our four instantiations of GACL (Sec. 3.3) as Ours-Sca, Ours-Mul, Ours-Add and Ours-Cos respectively.

4.1. GACL Supports Sketch Quality Discovery

For empirical evaluation of $q(\cdot)$, we select 8 out of 345 categories in QuickDraw based on the complexity, variety and semantic richness rules outlined in [24]. In Fig. 3, we first qualitatively visualise their distribution of $q(\cdot)$ under different GACL instantiations and demonstrate some exemplary sketch samples separated apart by dramatically different $q$ values. It can be seen that $q(\cdot)$ encodes sketch quality discriminatively in a reasonable way to viewers. Samples corresponding to smaller $q$ values are often aesthetically less pleasing, hard to recognise or simply incomplete and unreliable sketch data. On the other hand, $q(\cdot)$ works from a wide range of perspectives to interpreting good sketch quality, including smooth and coherent levels of visual structure rendering (e.g. umbrella), local conceptual semantics highlights (e.g. passport) and holistic visual aesthetics and richness (e.g. angel). It is also understandable that $q(\cdot)$ learned by Ours-Sca/Ours-Mul/Ours-Add/Ours-Cos are noticeably different (shading areas) given the different value domains they are designed to work on. All four GACL instantiations however show similar trend of score distribution change, indicating the possibility of a unified metric depending on the granularity of quality support evaluated on, as confirmed in our quantitative evaluation.

It needs more careful inspection when coming to quantitatively evaluating sketch quality. The common way of achieving this by measuring the difference between model predictions and human ground-truth ratings does not apply here as we lack of relevant annotations. We further argue that such approach would be flawed even we recruit human participants and collect their quality opinions on individual sketches – it’s hard to obtain objective and accurate scores...
with consensus give the subjective and abstract nature of free-hand sketch data. Inspired by the psychological findings on using set-based approach for more stable human behaviour in complex visual tasks [11,51], we evaluate the local quality rankings of GACL as a way [12,25] to coarsely examine its efficacy as a continuous global scale. Specifically, we recruit 40 participants with each undertaking 320 independent trials. We divide the scores underlying a quality metric \( q \cdot \) into 3-quantile and form three sketch sets with each containing random samples from its corresponding score range. Each participant is then required with a binary action on the question: “do you agree with the quality order between the presented sketch sets?” In Fig. 4(a), we plot the percentage of “yes” answers for each category. The mean percentage of 97.18% with standard deviation 0.38% among four GACL instantiations confirms the consistent efficacy of our quality discovery method. We further conduct a similar four-quantile task (Fig. 4(b)), where participants agree with the quality rankings 88.04% of all time.

### 4.2. Quality-Aware Sketch Recognition

One potential benefit of the proposed GACL framework is to contribute a competitive sketch recognition model as a byproduct – representation learning is discriminating between the quality of sketch instances and thus generalises better by less overfitting on lower quality data as per similar findings in the recent literature [4,6]. To verify, we first visualise the relationship between \( q \cdot \) and \( \theta \) in Fig. 6 and confirm that sketch instances of better quality (larger \( q \) values) tend to be more easily recognised (smaller \( \theta \) values) under Ours-Cos, while such phenomenon is failed to be observed in models trained by conventional Softmax loss. We further compare the performance between Ours-Cos and contemporary sketch recognition baselines in Tab. 1. It can be seen that our approaches achieve consistently and significantly improvements over their no-quality-attended counterpart (vs. BiLSTM), and even beat the state-of-the-art sketch recognition work of a noticeable margin without any bells and whistles (vs. SketchAA).

### 4.3. Quality-Guided Sketch Generation

In this section, we show \( q \cdot \) learned under GACL can be used to guide sketch generative models towards higher quality exploration in a post-hoc iterative manner (see supplementary for concrete implementation). A decisive factor affecting the synthesis outcome is the hyperparameter term \( \alpha \) that balances the weighting between self-reconstruction and quality improvement – in our setting, a larger \( \alpha \) value prefers the former. We showcase some examples in Fig. 5 between generation process under two distinct \( \alpha \) values and can observe that (i) our learned quality metric \( q \cdot \) is indeed a useful drop-in module to enriching existing generative sketch models with a quality dimension. By sliding
4.4. Quality-Enabled Sketch Attribution

One practical bottleneck for sketch model deployment today is the general lack of method for user sketch attribution – when the poor model performance is detected, developers can not know whether it comes down to model capacity itself or the malicious\(^4\) sketch input. In this section, we aim to examine to what extent can our learned \(q(\cdot)\) benefits such purpose. Intuitively, if the \(q\) value of a sketch is greater than a threshold value \(q_\tau\), we deem it as a benign user input (better quality) or which otherwise as a malicious one. We collect human opinions of their binary decision on whether a given sketch is maliciously intended or not and form an annotated test set of 2000 sketches with 1,000 for each sketch type as human ground truth. We adopt \(F_1\) score [40] as our evaluation metric for its ability to balance the performance between precision and recall. Result in Fig. 7(a) shows that with a raw \(q(\cdot)\) scorer can achieve the best \(F_1\) score of 61.25\% for sketch attribution.

With some additional efforts into analysing the attribution disagreements between our method and human annotators, we arrive at one interesting observation: some irregular sketches with excessively long and fragmented strokes or irrelevant personalised decorations deviating from the main rendering objective (Fig. 7(b)) are often treated as non-malicious inputs by human judges, contrary to our model predictions. We term these sketches as out-of-distribution (OOD) data and devise a way to prevent our model from attributing them to malicious inputs. The key insight is that despite \(q\) values for OOD and malicious sketch are all low, the stroke subset of the former can justify a much larger \(q\) value because it does encapsulate a well recognisable visual object – just with noisy visual outliers perturbing model predictions. This means given a sketch input and if its \(q\) is lower than the threshold value \(q_\tau\), we can chip in one extra conditioning step before we decide to categorise it into malicious input or not. Specifically, we simply test one stroke at a time and remove it from the input if that can lead to a noticeable increase on \(q\) value. We treat a sketch as OOD, i.e. non-malicious input, if there exists a partial composition of its strokes that reach to a \(q\) value more than a pre-set threshold \(q_{\text{max}}\) (Fig. 7(b)). We denote such method as \(q_{\text{max}}(\cdot)\) and compare with \(q(\cdot)\) using two different \(q_{\text{max}}\) values in Fig. 7(a). Significant improvements can be observed by taking into account our modelling on OOD sketch input.

5. Conclusion

We have presented a method for quantifying human freehand sketch quality. Without relying on supervision from human quality opinion annotations for learning, our proposed solution GACL is able to stand up to the test of a human study by showing human-agreeable results on sketch quality discrimination. We also demonstrate three practical use cases benefited from successful sketch quality modelling. We hope our work can be of help to sketch practitioners who seek for further application advancements and are held back by the lack of a proper quality metric. Moreover, we expect GACL is not constrained to work with vector sketch of point-based representation only and leave the exploration of its efficacy for more data modalities (e.g. raster, 3D) as future work.

Acknowledgements We thank the anonymous reviewers for their valuable comments. This work was supported in part by the National Natural Science Foundation of China (NSFC) under grant # 62076034. We especially thank the China Scholarship Council (CSC) for funding the first author to conduct the entirety of this project at SketchX under # 202006470075.

---

\(^4\)Random scribbles that do not conform to any semantic concept.
References


[33] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. TIP, 2017. 2

