Supplementary Material for "Sketch-BERT: Learning Sketch Bidirectional Encoder Representation from Transformers by Self-supervised Learning of Sketch Gestalt"

Anonymous CVPR submission

Paper ID 4710

1. Frameworks for Different Downstream Tasks

We first illustrate the frameworks of how to leverage the Sketch-BERT on different sketch tasks in Fig. 1. Formally, a sketch is denoted as $s = (s_1, s_2, \ldots, s_n) \in R^{n\times 5}$, $s^{ps} \in R^{n\times 2}$; $s^{st} \in R^{n\times 3}$ represent position and state parts for a sketch; the mask is defined as $m \in \{0,1\}^{n\times 5}$. For different tasks, we use $f_*(\cdot)$ to represent the Sketch-BERT model with different outputs. **Sketch Gestalt**. As the Fig. 1(a) shows, for sketch gestalt, Sketch-BERT takes a masked sketch in sequential type and then predicts the masked parts. We denote the input as $s_{mask} = s_{gt} \cdot m, m$; for convenience, we also define the selection operation: $s[1-m], s[m]$ as selected masked points and selected unmasked points; Sketch-BERT model outputs the completed sketch $s_{comp} = f_{ges}(s_{mask}, m)$, To train the Sketch-BERT model on sketch gestalt, we utilize l_1 loss for position values and cross entropy loss for state values, the loss functions are shown in the following

$$
\begin{aligned}
L_{ps} &= \|(s^{ps}_{comp} - s^{ps}_{gt})[1-m^{ps}]\|_1 && (1)\\
L_{st} &= CE(s^{st}_{comp}[1-m^{st}], s^{st}_{gt}[1-m^{st}]) && (2)\\
L_{ges} &= L_{ps} + L_{st}. && (3)
\end{aligned}
$$

In test procedure, the predicted points will be combined with the incomplete sketch from the input to get the final completed output sketch. **Sketch Recognition/Classification**. Then, for sketch recognition task, we add a [CLS] label to get the classification label at the same position. The Sketch-BERT model will output the predicted class label $T_{pred} = f_{cls}(s)$ and we define the true class label as T. We employ the standard cross entropy loss to fine-tune the Sketch-BERT as shown in Fig. 1(b).

$$L_{cls} = CE(T_{pred}, T) \quad (4)$$

Sketch Retrieval. The framework for Sketch Retrieval is more complicated, where three weight-sharing Sketch-BERT models are used for extracting features for retrieval. We define the query sketch, positive sketch and negtive sketch as s_q, s_p, s_n; the class label for s_q, s_p is defined as T_p; Sketch-BERT will output a retrieval feature for each input $r_s = f_{ret}(s)$; a further fully-connected layer for classification is defined as $g_{cls}(f_{ret}(s))$. To train the model, we employ both triplet loss and cross entropy loss as shown in Fig. 1(c).

$$
\begin{aligned}
L_{tri} &= \max(0, \|r_{s_q} - r_{s_p}\|_2 - \|r_{s_q} - r_{s_n}\|_2 && (5)\\
&\quad + margin) && (6)\\
L_{cls} &= CE(g_{cls}(f_{ret}(s_q)), T_p) && (7)\\
&\quad + CE(g_{cls}(f_{ret}(s_p)), T_p) && (8)\\
L_{ret} &= L_{tri} + L_{cls} && (9)
\end{aligned}
$$

where the $margin$ is set to 1.

2. More Qualitative Examples for Sketch Gestalt

In this section, we show more examples from 20 classes, *bread,stethoscope,floor lamp,cloud,shovel,guitar,tiger,crayon,violin,onion,banana, flashlight,cake,camel,streetlight,ambulance,wine bottle,diamond,helmet,hammer*, for sketch gestalt task in Fig. 2, 3, 4. There are 6 examples in each class and these instances are a part of the sketches of the user study we conduct further.

3. Quantitative Results for Sketch Gestalt Task

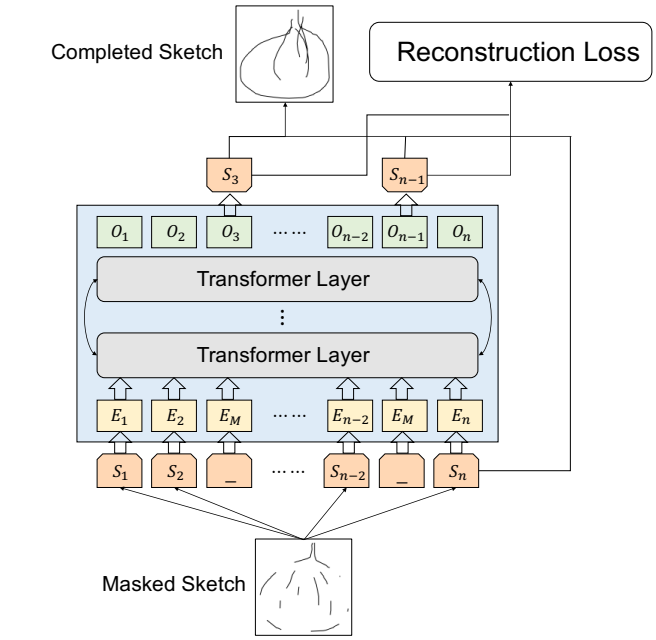
We conduct a user study to evaluate different models for sketch gestalt task. 200 test sketches are sampled from 20 classes with 10 samples in each class. The results from different methods will be shown to 10 participants where each participant is assigned test sketches from 6 class *i.e.*, each sketch result will be evaluated by three persons. Particularly, we ask each participant to rank the results in three different levels (1) Result from Sketch-BERT is much better,

(2)Results from Sketch-BERT and SketchRNN are almost the same, (3) Result from SketchRNN is much better.

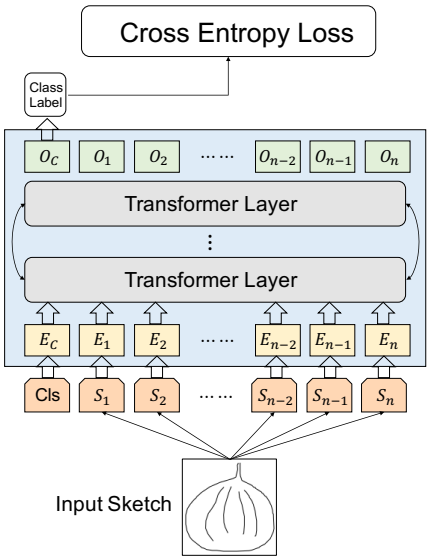
Models	Preference	Equal	Preference(%)
Sketch-BERT	352	197	58.67 %
SketchRNN	51	197	8.50 %

Table 1. The results of user Study on sketch gestalt task.

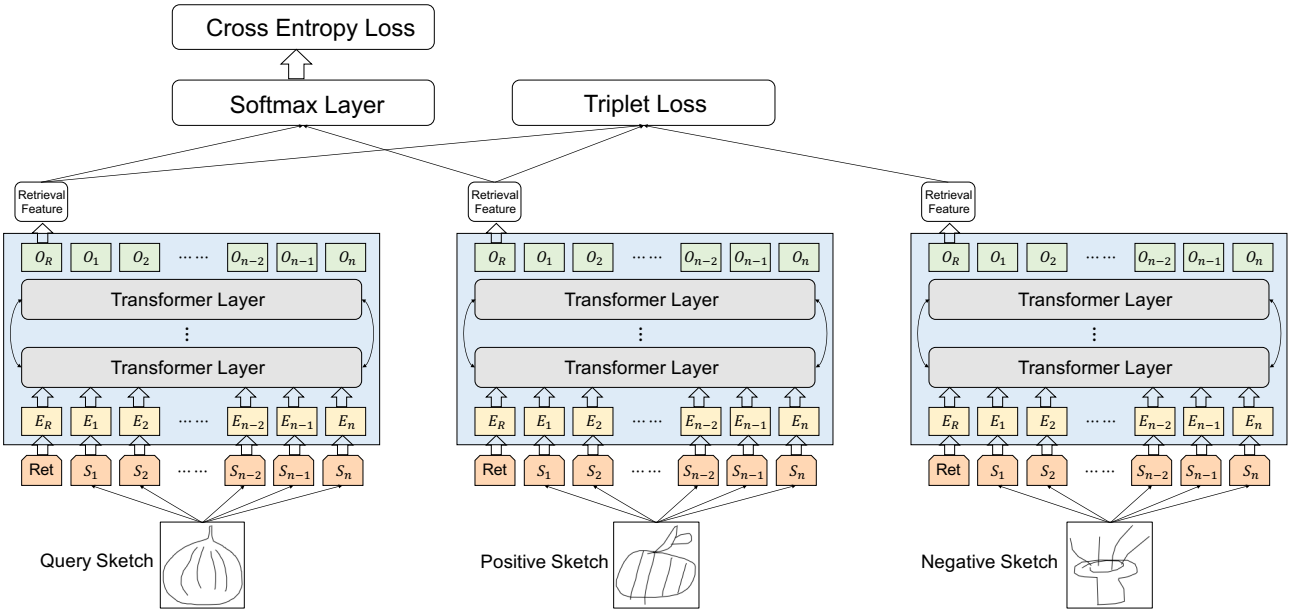
We show the collected data in Tab. 1. Preference and Preference (%) represent the number and percentage of the participants choose the corresponding method. The Equal number is the same for two methods since it means the number of people who think two methods perform similarly in their examples. We can find our Sketch-BERT has a 58.67 % preference rate which is much higher than the SketchRNN with 8.50%. It demonstrates the capacity of our Sketch-BERT on such sketch gestalt task.



(a) Sketch-BERT for Sketch Gestalt



(b) Sketch-BERT for Sketch Recognition



(c) Sketch-BERT for Sketch Retrieval

Figure 1. Overview structures of Sketch-BERT for Sketch Gestalt, Sketch Recognition and Sketch Retrieval tasks.

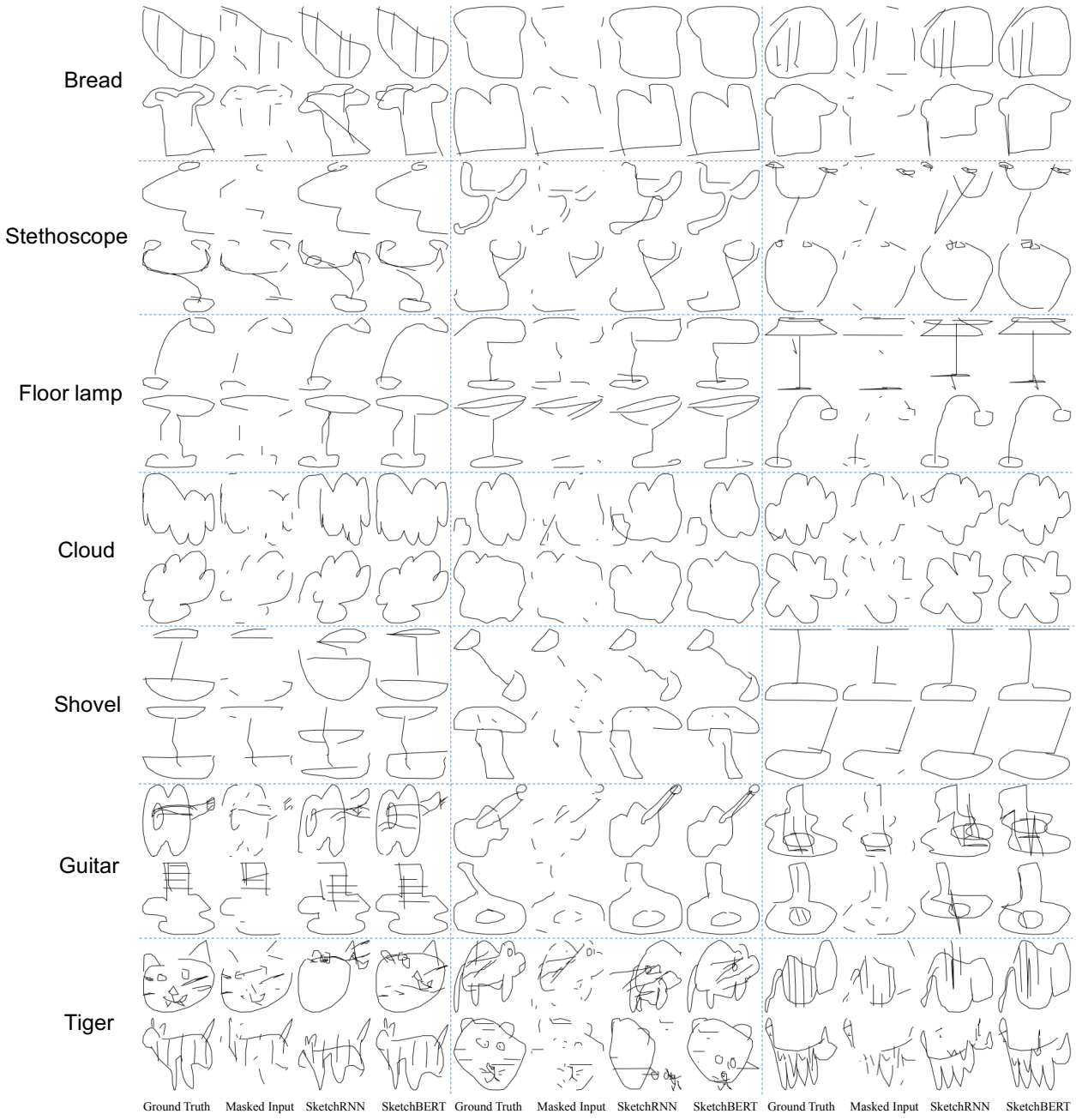


Figure 2. Part (a):Qualitative Results from 20 classes for Sketch Gestalt.

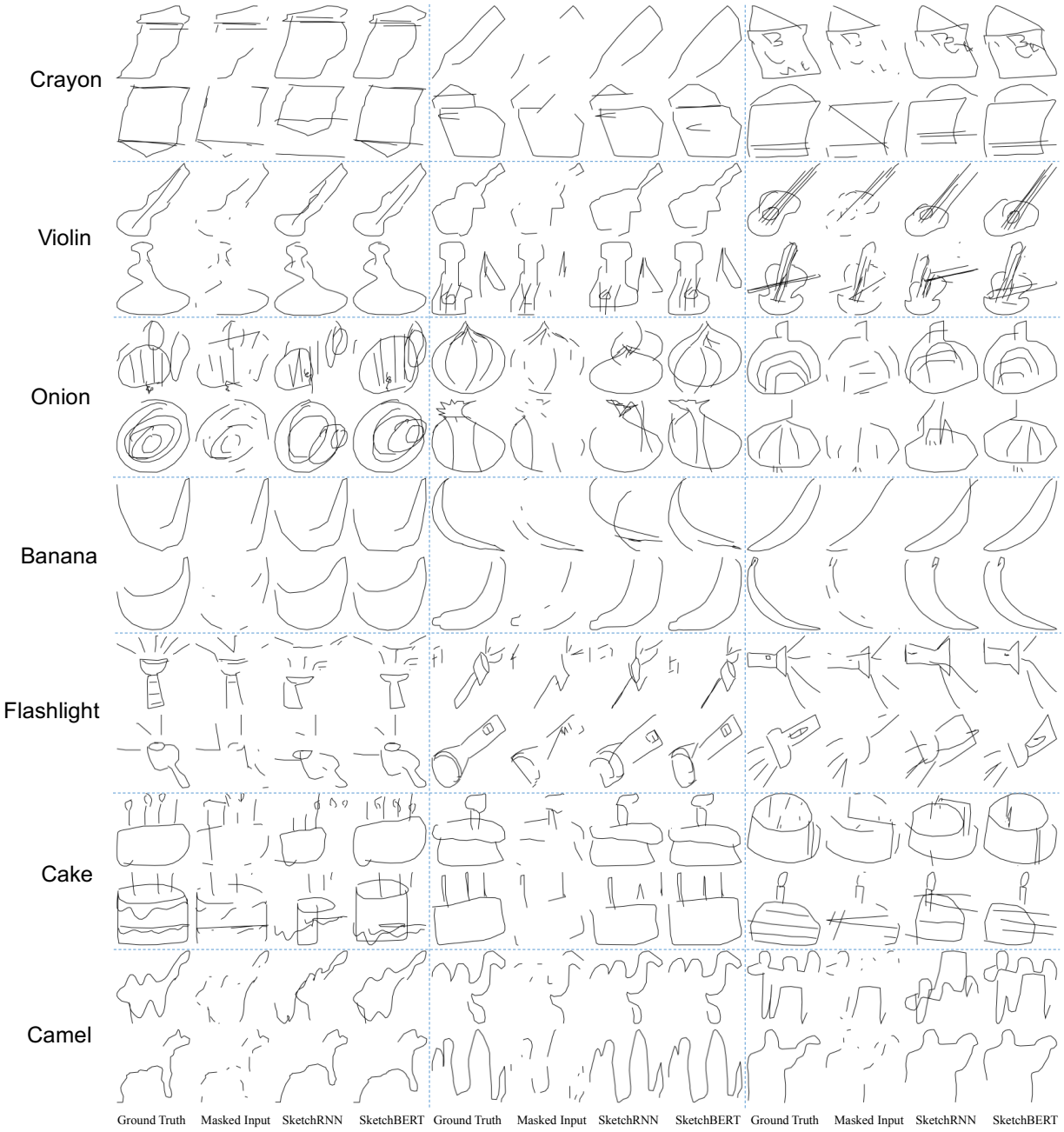


Figure 3. Part (b):Qualitative Results from 20 classes for Sketch Gestalt.

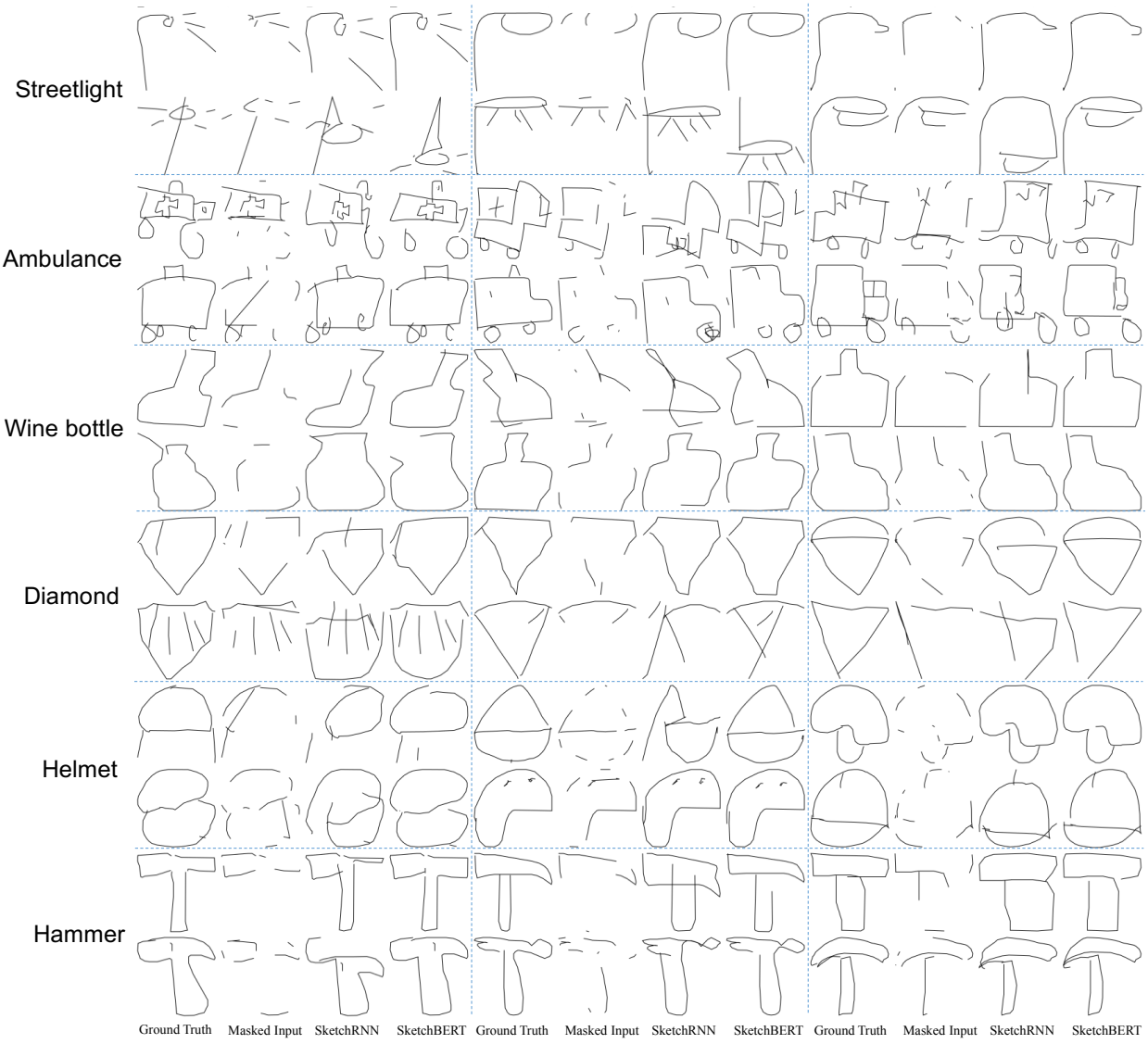


Figure 4. Part (c):Qualitative Results from 20 classes for Sketch Gestalt.