Composite Sketch+Text Queries for Retrieving Objects with Elusive Names and Complex Interactions

Prajwal Gatti\textsuperscript{1}, Kshitij Parikh\textsuperscript{1}, Dhrity Prasanna Paul\textsuperscript{1}, Manish Gupta\textsuperscript{2}, Anand Mishra\textsuperscript{1}

\textsuperscript{1}Indian Institute of Technology Jodhpur, \textsuperscript{2}Microsoft

### The CSTBIR Task

**Search Query**

```
"Small mammal with striped back and long snout digging in the ground."
```

**Retrieved Images**

![Images of small mammals and snouts digging in the ground]

---

**Given:** a hand-drawn sketch $S$, a complementary text $T$ and a database $D$ of $N$ natural scene images with multiple objects

**Rank:** $N$ images according to relevance to composite $(S, T)$ query.

---

### CSTBIR Dataset

#### Query

- Pair of \textsuperscript{markhor}, \textsuperscript{bodhran}, \textsuperscript{penny-farthing} climbing cliffs on a sunny day.
- People admiring a \textsuperscript{markhor} displayed on a table.
- Person dressed in a suit standing beside a \textsuperscript{bodhran}.

#### Target Image

- (objects: markhor, bodhran, penny-farthing)

---

- Natural images and text descriptions from Visual Genome and sketches from Quick, Draw!
- Train (~1.89M queries, ~97K images, 258 object classes)
- Validation (~5K images, ~97K queries)
- Test-1K: 1K queries, 1K images
- Test-5K: 4K queries, 5K images
- Open-Category set: 750 queries, 70 objects, 1K images.

---

### STNet: Sketch+Text Network

**Training objectives:** (i) Contrastive Training, (ii) Object Classification, (iii) Sketch-Guided Object Localization, and (iv) Sketch Reconstruction

---

### Results

<table>
<thead>
<tr>
<th>Method</th>
<th>R@10</th>
<th>R@20</th>
<th>R@50</th>
<th>R@100</th>
<th>MDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>DoodieSearch</td>
<td>14.3</td>
<td>3.6</td>
<td>2.6</td>
<td>1.0</td>
<td>129.0</td>
</tr>
<tr>
<td>DeepSBIR</td>
<td>20.4</td>
<td>8.9</td>
<td>9.9</td>
<td>11.0</td>
<td>233.0</td>
</tr>
<tr>
<td>VIT-Siamese</td>
<td>7.5</td>
<td>1.6</td>
<td>8.8</td>
<td>1.0</td>
<td>75.5</td>
</tr>
<tr>
<td>VisualBERT</td>
<td>23.3</td>
<td>7.5</td>
<td>35.9</td>
<td>15.4</td>
<td>40.8</td>
</tr>
<tr>
<td>VIT</td>
<td>28.1</td>
<td>10.5</td>
<td>42.7</td>
<td>16.3</td>
<td>60.2</td>
</tr>
<tr>
<td>CLIP</td>
<td>50.6</td>
<td>2.4</td>
<td>63.1</td>
<td>33.7</td>
<td>78.8</td>
</tr>
<tr>
<td>TIRG</td>
<td>31.9</td>
<td>10.4</td>
<td>44.2</td>
<td>17.3</td>
<td>62.8</td>
</tr>
<tr>
<td>Taskformer</td>
<td>22.4</td>
<td>9.5</td>
<td>35.6</td>
<td>14.8</td>
<td>23.6</td>
</tr>
<tr>
<td>Two-stage</td>
<td>67.0</td>
<td>34.8</td>
<td>74.9</td>
<td>46.9</td>
<td>86.4</td>
</tr>
<tr>
<td>Two-stage (desc)</td>
<td>60.1</td>
<td>30.5</td>
<td>73.7</td>
<td>41.7</td>
<td>83.5</td>
</tr>
</tbody>
</table>

Table 1: Performance comparison on the CSTBIR test-1K/5K.

### Ablation study for STNet model

### Summary

CSTBIR: Composite Sketch+Text Based Image Retrieval Task

- New Dataset: containing ~2M queries and ~108K natural images
- STNet: Pre-trained multilmodal transformer based method with task-specific training objectives.

---

Paper and resources available here: [https://vl2g.github.io/projects/cstbir/](https://vl2g.github.io/projects/cstbir/)