Abstract
Non-native speakers with limited vocabulary often struggle to name specific objects despite being able to visualize them, e.g., people outside Australia searching for ‘numbats.’ Further, users may want to search for such elusive objects with difficult-to-sketch interactions, e.g., “numbat digging in the ground.” In such common but complex situations, users desire a search interface that accepts composite multimodal queries comprising hand-drawn sketches of “difficult-to-name but easy-to-draw” objects and text describing “difficult-to-sketch but easy-to-verbalize” object’s attributes or interaction with the scene. This novel problem statement distinctly differs from the previously well-researched TBIR (text-based image retrieval) and SBIR (sketch-based image retrieval) problems. To study this under-explored task, we curate a dataset, CSTBIR (Composite Sketch+Text Based Image Retrieval), consisting of ~2M queries and 108K natural scene images. Further, as a solution to this problem, we propose a pretrained multimodal transformer-based baseline, STNET (Sketch+Text Net), that uses a hand-drawn sketch to localize relevant objects in the natural scene image, and encodes the text and image to perform image retrieval. In addition to contrastive learning, we propose multiple training objectives that improve the performance of our model. Extensive experiments show that our proposed method outperforms several state-of-the-art retrieval methods for text-only, sketch-only, and composite query modalities. We make the dataset and code available at: https://vl2g.github.io/projects/cstbir.

Introduction
Traditional text-based image retrieval (TBIR) systems (Li et al. 2020a; Kim, Son, and Kim 2021; Zhang et al. 2020; Lee et al. 2018; Li et al. 2020b) are intuitive for users with strong linguistic abilities. However, non-native speakers or users unfamiliar with particular objects struggle in using such systems to find objects with “elusive” names, e.g., users outside Australia searching for numbats, as shown in Figure 1. Elaborate text descriptions in lieu of the precise object name could provide limited help, even with all the details. For example, “Small mammal with striped back and long snout digging in the ground” as a replacement for “numbat” leads to images of chipmunk, badger, weasel, mongoose, or skunk.

Sketch-based image retrieval (SBIR) systems (Yu et al. 2016; Dey et al. 2019; Song et al. 2017b; Collomosse, Bui, and Jin 2019; Sain et al. 2022) seem to provide an illusory relief in such situations. Although a user can sketch “difficult-to-name but easy-to-draw” objects, (1) users may not have enough time, skills, or tools to draw all the details, leading to ambiguity in sketches; (2) users may be looking for “difficult-to-sketch but easy-to-verbalize” object’s attributes or interaction with the scene. For example, for the query, “numbat digging in the ground”, it is difficult to draw a sketch to represent “digging in the ground”, and even if drawn, the sketch could lead to false positives about “numbat eating an insect” or “numbat searching for termites”.

Such common but complex search situations require novel multimodal search interfaces, allowing seamless text and sketch mix-ups in queries. Such a flexible and natural interface should help the user to draw sketches for “difficult-to-name” objects effortlessly and then complement their creations with text descriptions to define layout, color, pose, and other object characteristics, along with complex interactions with other objects in scenes. We refer to such a novel proposed system as CSTBIR or Composite Sketch+Text.
Based Image Retrieval system.

Although a vast literature exists on TBIR and SBIR, to the best of our knowledge, the CSTBIR problem setting has yet to be studied rigorously. There have been some recent works (Song et al. 2017a; Sangkloy et al. 2022; Chowdhury et al. 2023a) that attempt to solve a simpler version, where: (a) target image collection is focused objects rather than complex natural scenes, (b) sketch is at scene-level rather than object-level, and (c) text description is comprehensive rather than partial (or complementary). This paper proposes a system for the complex CSTBIR setting.

Given input with a rough sketch and a complementary text description (e.g., the sketch of “numbat” and text=“digging in the ground”), an evident approach is to guess the object name from the sketch, complete the text query as “Numbat digging in the ground” and then use TBIR methods. However, such a two-stage method may fail when the sketch represents an object with an ambiguous name (e.g., mouse, bat, crane) and suffers from signal loss when attempting to describe knowledge in the sketch using an object name. Additionally, such two-stage approaches are restricted to closed-world settings where the object names are previously known and may not generalize well to rare or novel objects. Hence, we propose a principled method – STN\_ET that jointly processes text and sketch inputs. More specifically, we propose the following task-specific pretraining objectives for the multimodal transformer: (i) object classification, i.e., predict object name; (ii) sketch-guided object detection, i.e., localize the relevant objects in the image, and (iii) sketch reconstruction, i.e., recreate the query sketch from the multimodal representation of sketch and the image.

Overall, we make the following main contributions in this paper: (i) We study an important and under-explored task, namely CSTBIR. (ii) Toward this novel setting, we contribute a large dataset of \( \sim 2M \) queries and \( \sim 108K \) natural scene images. (iii) For CSTBIR, we pre-train a multimodal Transformer STN\_ET, designed to handle sketch and text as inputs, using multiple loss functions: contrastive loss, object classification loss, sketch-guided object detection loss, and sketch reconstruction loss. (iv) Our proposed model outperforms several competitive text-only, sketch-only, and sketch+text baselines.

### Related Work

Image retrieval systems can answer queries expressed using hand-drawn sketches (SBIR), text (TBIR), a combination of sketch and text (CSTBIR), color layout, concept layout (Zhou, Li, and Tian 2017), visual features (Tian, Newsam, and Boakye 2023; Dodds et al. 2020), or location-sensitive tags (Gomez et al. 2020). We review existing work on TBIR, SBIR, and multimodal query-based IR.

**Sketch-Based Image Retrieval (SBIR):** It allows the flexibility to easily specify the qualitative characteristics using sketches (Yu et al. 2016; Dey et al. 2019; Song et al. 2017b). Following the initial work on sketch recognition (Sun et al. 2012), earlier SBIR studies mainly focused on convolutional neural networks (CNN) (Yu et al. 2016; Liu et al. 2017) which was soon followed by various Transformer (Vaswani et al. 2017)-based architectures (Ribeiro et al. 2020; Chowdhury et al. 2022). Deep Siamese models with triplet loss have also been explored (Yu et al. 2016; Collomosse, Bui, and Jin 2019). Several specialized SBIR settings have also emerged such as Zero Shot-SBIR (Pandey et al. 2020; Dey et al. 2019; Dutta and Akata 2019), fine-grained SBIR (Liu et al. 2020; Bhunia et al. 2022; Pung et al. 2019, 2017; Ling et al. 2022; Bhunia et al. 2020; Song et al. 2017b), and category-level SBIR (Sain et al. 2021; Bhunia et al. 2021; Sain et al. 2022).

**Text-Based Image Retrieval (TBIR):** Popular methods for TBIR include alignment of input text and the corresponding input image using pretrained multimodal Transformer methods like VisualBERT (Li et al. 2020a) and ViLT (Kim, Son, and Kim 2021). Further, cross-attention-based models (Zhang et al. 2020; Lee et al. 2018) and models that use object tags detected in images (Li et al. 2020b) have also been proposed. Recently, contrastive learning methods (Jia et al. 2021), along with zero-shot learning (Radford et al. 2021), have been shown to achieve state-of-the-art results.

**Multimodal Query Based Image Retrieval:** Several systems have been built to consume multimodal input for image retrieval. Earlier works used reference images and text as an attribute on a category-level retrieval (Kovashka, Parikh, and Grauman 2012; Han et al. 2017). Input text data was more elaborated to provide improved results (Guo et al. 2018; Vo et al. 2019). While such earlier systems used CNNs, more recent systems (Song et al. 2023; Baldrati et al. 2022) leverage Transformers. Further, some studies (Changpinyo et al. 2021; Pont-Tuset et al. 2020) explored the setting where the user simultaneously uses both speech and mouse traces as the query. Lastly, (Nakatsuka, Hamasaki, and Goto 2023) search images relevant to input music.

It is not always possible to have an input reference image for image retrieval; instead, a sketch (along with text description) is used, which gives more flexibility. Image retrieval using hand-drawn sketches and textual descriptive data has been under-explored.

Detailed sketch and text input have been used to (a) retrieve e-commerce product images using CNNs and LSTMs (Song et al. 2017a), and (b) retrieve scene images using CLIP (Sangkloy et al. 2022; Chowdhury et al. 2023a). However, in several practical scenarios, (a) the sketch is

<table>
<thead>
<tr>
<th>Query</th>
<th>Dataset</th>
<th>Sketch</th>
<th>Text</th>
<th>Target Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sketch</td>
<td>TU-Berlin</td>
<td>Object</td>
<td>None</td>
<td>Focused Object</td>
</tr>
<tr>
<td>Sketch</td>
<td>QMUL-Shoe-V2</td>
<td>Object</td>
<td>None</td>
<td>Focused Object</td>
</tr>
<tr>
<td>Text</td>
<td>COCO</td>
<td>None</td>
<td>Complete</td>
<td>Complete Scene</td>
</tr>
<tr>
<td>Text</td>
<td>Flickr-30K</td>
<td>None</td>
<td>Complete</td>
<td>Complete Scene</td>
</tr>
<tr>
<td>Sketch+Text</td>
<td>SF COCO</td>
<td>Scene</td>
<td>Complete</td>
<td>Complete Scene</td>
</tr>
<tr>
<td>Sketch+Text</td>
<td>CSTBIR (Ours)</td>
<td>Object</td>
<td>Complementary</td>
<td>Complete Scene</td>
</tr>
</tbody>
</table>

Table 1: Comparison of datasets with CSTBIR. The CSTBIR is the only dataset that demands searching over a database of natural scene images using queries of object sketch and partial complementary natural language sentences.
object-level, very rough, and not elaborate, and (b) the text is partial (complementary to sketch) and not self-contained. Unfortunately, no previous work exists for such a (complex) practical setting. Our contributed dataset, CSTBIR, and the proposed method addresses this setting in this paper. Compared to (Sangkloy et al. 2022) where sketch covers 100% area of the image to be retrieved, in our dataset, sketches cover only 36.7% area of the matching scene image on average. In our dataset, sketches are less complex than in (Sangkloy et al. 2022), which contain ∼2.6x times more sketch pixels compared to our dataset1. Table 1 shows these comparisons of CSTBIR with other existing image retrieval datasets (Eitz, Hays, and Alexa 2012; Yu et al. 2016; Lin et al. 2014; Young et al. 2014; Chowdhury et al. 2022).

The CSTBIR Problem and Dataset

Given a hand-drawn sketch $S$, a complementary text $T$ and a database $D = \{I_i\}_{i=1}^N$ of natural scene images with multiple objects, the CSTBIR problem aims to rank the $N$ images according to relevance to the composite $(S, T)$ query.

Due to the lack of a suitable dataset, we curate the CSTBIR dataset, where each sample consists of a hand-drawn sketch, a partial complementary text description, and a relevant natural scene image. The natural scene images in the database have multiple object categories, attributes, relationships, and activities. Although this dataset does not have “difficult-to-name” objects, it is a reasonable proxy. We also evaluate using a manually curated separate test set of “difficult-to-name” objects.

The natural images and text descriptions are taken from Visual Genome (Krishna et al. 2017). The dense annotations in this dataset allow us to frame multiple queries related to an image, each of which pertains to a particular object in the image. The hand-drawn sketches are taken from the Quick, Draw! dataset (Ha and Eck 2018). Annotators have drawn these sketches in <20 seconds; hence, they are rough and lack the exact details as that of an image, which aligns with the challenging real-world setting of this task. Quick, Draw! has over 50M sketches across 345 categories.

Table 2 shows basic statistics of the dataset. The dataset has been split into train, validation, and test based on the corresponding splits from Visual Genome for the scene images. The dataset has a total of ∼108K images and ∼562K sketches. The training dataset consists of ∼97K images, ∼484K sketches, and ∼1.89M queries. On average, the text sentences contain 5.4 words. The dataset also includes a validation set with ∼5K images, ∼83K sketches, and ∼97K queries. Further, it contains three test sets: Test-1K, Test-5K, and Open-Category set. Test-1K includes 1K queries and corresponding 1K natural scene images in the gallery. Test-5K is a more challenging set that contains 4K queries and 5K gallery images. All the sketch object categories in Test-1K and Test-5K sets are present during training. However, the scene and sketch images in the test set were not part of the training or validation set. We created the Open-Category test set to evaluate the model on novel object categories unseen at train time, which contains 70 novel object categories (of which 50 are “difficult-to-name”) and corresponding sketches.

Figure 2 shows a few examples from the dataset. For further data analysis, we performed part of speech tagging on text descriptions using NLTK. We visualize these statistics in the Appendix as word clouds for the top few adjectives (object attribute indicating words), verbs (action indicating words), and prepositions (position indicating words) for the text descriptions in the CSTBIR dataset.

The Proposed STNET Model for CSTBIR

For the task of sketch and text-based image retrieval, we introduce STNET (Sketch+Text Network), a novel multimodal architecture. It comprises three independent Transformer-based encoder networks based on the pre-trained CLIP model (Radford et al. 2021). The overall architecture of STNET is illustrated in Figure 3. Next, we describe the working and architectural details of STNET in the
following subsections.

**Query (Sketch+Text) Encoding**

In CSTBIR, the query consists of a text sentence and a hand-drawn sketch. We independently encode these two inputs using a pretrained CLIP text encoder and a pretrained Vision Transformer (ViT) (Dosovitskiy et al. 2021) encoder.

Given a query text sentence $T$, we tokenize it using a Byte-Pair-Encoding scheme according to the learned vocabulary of the text encoder as $F_T = [CLS, t_1, t_2, \ldots, t_n]$, where each $t_i$ represents a sub-word token, and $CLS$ is the global pool token. Given the query sketch image $S$, we use a pretrained ViT encoder which is fed the input $F_S = [CLS, s_1, s_2, \ldots, s_m]$, where each $s_i$ is the embedding of an image patch. As the ViT encoder is pretrained on the ImageNet-21K dataset (Ridnik et al. 2021), we first train it on the sketch data for the classification task to adapt it for the sketch domain. This trained encoder is then used to embed the sketch input. Overall, this results into text embedding $h^{CLS}_T$ and sketch embedding $h^{CLS}_S$.

**Image Encoding**

To utilize the benefits of large-scale pretraining, we use the pretrained CLIP-ViT image encoder. Similar to the ViT encoder in the sketch, a candidate scene image $I$ is reshaped to a fixed size ($224 \times 224$) and then spatially sliced into a $16 \times 16$ grid of non-overlapping image patches. Further, these image slices are then reshaped into a sequence of embeddings before passing it to the ViT for further processing.

As our problem focuses on queries related to objects in natural scenes, it would be beneficial for our model to focus on the object being queried in the scene image. To enable this, we would like to use the sketch input $S$ to localize or attend to the corresponding object of interest in the image. Specifically, as shown in Figure 3, we use the pooled output of the sketch encoder $h^{S}_{CLS}$ to calculate dot-product attention over the output embeddings of the image encoder $H^I$. The obtained values represent attention scores $a_{IS}$ over the spatial regions of the image as well as the $CLS$ token. We obtain weighted values of image embeddings $H^I$, which are then average pooled to get the final image embedding $h^{AVG}_I$. Mathematically, $a_{IS} = \text{Softmax}(H^I \times h^{CLS}_S)$, $H^I = a_{IS} \odot H^I$ and $h^{AVG}_I = \frac{1}{m} \sum_{i=1}^{m} H^I_i$, where $h^{AVG}_I$ represents the global average pooled embedding of the image encoder.

**STNET Training**

STNET follows multiple task-specific training objectives.

**Contrastive Learning ($L_{CT}$)** We adopt a batch-wise contrastive learning strategy akin to CLIP (Radford et al. 2021) to facilitate image retrieval. Given a batch of $N$ paired (query, image) samples from the train set, we aim to maximize the cosine similarity of the image and query embeddings of the $N$ real pairs in the batch while minimizing the cosine similarity of the embeddings of the $N(N - 1)$ incorrect pairings. We use conditional sampling to ensure uniqueness, i.e., a query does not match multiple images and vice versa. Particularly, we use the InfoNCE loss between $h^{T, CLS}_I$ and $h^{S, AVG}_I$ to obtain the contrastive loss ($L_{CT}$) as done in CLIP. Further, as our model utilizes the pretrained CLIP, which lacks joint modeling of text, sketch, and image modalities, we propose three additional training losses to be optimized concomitantly with the contrastive objective.

**Object Classification ($L_{CLS}^C$ and $L_{CLS}^T$)** Given that the CSTBIR problem focuses on object-specific queries, we propose separately predicting the object name from the text sentence and image inputs. To this end, we train the text and image encoders for the multi-class classification objective to predict the object’s class from the $C$ object categories available in the train set. Since the object’s label is not mentioned in the text sentence or is always prominent in the image, this objective requires the model to use the contextual information from both modalities to predict the object class. We refer to the classification losses computed using text encodings and the image encodings as $L_{CLS}^T$ and $L_{CLS}^C$, respectively.

**Sketch-Guided Object Detection ($L_{OD}$)** To aid the localization of the query-relevant object while encoding the image information, inspired by the recent literature in the sketch-guided object detection problem (Tripathi et al. 2020; Chowdhury et al. 2023b; Tripathi et al. 2023), we propose the training objective of sketch-guided localization of the object. Specifically, given the sketch-attended embeddings $H^I$ from the image encoder, we utilize the embeddings corresponding to the $16 \times 16$ spatial grids. Following the implementation from YOLO (Redmon et al. 2016), we transform the output embeddings of the ViT network to predict an output of shape $S \times S \times (5B + C)$, where $S \times S$ represents the image grid size, each predicting $B$ bounding boxes, and $C$ class probabilities. We use $S = 7$, $B = 2$, and we have $C = 258$ classes in our train set, so we predict a $7 \times 7 \times 268$ output tensor. Finally, we use intersection over union (IoU) to calculate the multipart object detection ($L_{OD}$) loss as done in (Redmon et al. 2016).

**Sketch Reconstruction ($L_{SR}$)** Similar to the object detection training objective, which facilitates the localization of the relevant objects, we introduce the task of sketch reconstruction using the image features $H^I$ as illustrated in Fig-
We employ eight blocks of Convolution-BatchNorm-ReLu as done in (Isola et al. 2017) to upsample the information to a reconstructed sketch tensor of size $1 \times 224 \times 224$. Further to train the sketch-reconstruction module, we utilize a combination of Binary Cross Entropy loss and the DICE loss (Sudre et al. 2017) as $L_{SR} = \alpha L_{BCE} + \beta L_{DICE}$.

Our overall loss is the sum of all five losses $L_{CT} + L_{CLS} + L_{OD} + L_{SR}$.

We measure the distance between the query and the image embedding during retrieval using cosine similarity. We provide the implementation details for STNet in the Appendix and make code and dataset available at our project page³.

### Experiments and Results

#### Baseline Models

We compare STNet extensively with competitive image retrieval baselines.

**Sketch-based Image Retrieval (SBIR):** SBIR is a prominently studied domain in the literature. In our setup, from our composite queries, we only take sketches and drop text to experiment with these baselines. We choose two representative and competitive SBIR methods as our baselines: Doodle2Search (Dey et al. 2019) and DeepSBIR (Yu et al. 2016). We also create a Vision Transformer-based SBIR baseline, viz. ViT-based Siamese Network. This network comprises two ImageNet-pre-trained ViT-based encoders for sketch and image modalities, trained using the InfoNCE loss (Oord, Li, and Vinyals 2018).

**Text-based Image Retrieval (TBIR):** These baselines perform retrieval using only the text part of the query while ignoring the sketch component. We choose the following three modern approaches in this category: VisualBERT (Li et al. 2020a), ViLT (Kim, Son, and Kim 2021), and CLIP (Radford et al. 2021).

**Composite Query-based Image Retrieval:** These baselines perform retrieval using the sketch and text inputs. We compare our proposed method, STNet, with the following baseline methods: TIRG (Vo et al. 2019) and TaskFormer (Sangkloy et al. 2022), and a two-stage model. We trained the TIRG model from scratch using our dataset. For TaskFormer, we finetuned the publicly available checkpoint using our dataset and our reproduced code for training. We adhered to the hyperparameter configurations outlined in their respective papers for these models. For the two-stage method, in the first stage, we use a ViT trained for sketch classification to get an object name from the sketch. Next, in the second stage, we obtain the full-text query by inserting the predicted object name into the incomplete text and then using pretrained CLIP for image retrieval.

Finally, we experimented with another baseline, “two-stage (desc)”. This method’s first stage is the same as the “two-stage”. In the second stage, rather than using the class name, we obtain the full-text query by inserting the description of the predicted object into the incomplete text and then using a pretrained CLIP model for image retrieval. The description for each of the 258 object names is chosen randomly from seven different sets of descriptions annotated per object name. Five of these object description sets are obtained automatically, while the other two are manually annotated.

Automated descriptions were generated by using ChatGPT-3.5³ on Mar 14, 2023. We used the following five prompts to obtain five different description sets: (i) “Describe the following words with visual descriptions in 4 to 10 words.” (ii) “Describe the following words with visual descriptions as a 15-year-old kid in 4 to 10 words.” (iii) “Describe the following words with visual descriptions as a 35-year-old in 4 to 10 words.” (iv) “Describe the following words with visual descriptions as a 55-year-old in 4 to 10 words.” (v) “Describe the following words with visual descriptions as a non-native English speaker in 4 to 10 words.” The human annotators were asked to write descriptions with 4 to 10 words that included visual attributes without mentioning the object’s name.

We use two metrics: Recall@K and Median Rank (MdR). Recall@K is the percentage of times the ground truth image is retrieved within the top K results across all queries in the test set; the higher, the better. Median Rank is the median of the rank of ground truth image in the retrieved set across all queries in the test set; the lower, the better.

### Results on Test-1K and Test-5K

Table 3 shows our main results on both test sets. Our proposed method, STNet, outperforms all baseline methods. Multiple sketch+text-based image retrieval models are better than text-based models, which are better than sketch-based image retrieval models. This is mainly because neither the sketches nor the incomplete text can answer the queries accurately. Amongst sketch-based image retrieval models, ViT-based Siamese networks perform the best. Among text-based image retrieval models, CLIP performs the best. STNet is better than the two-stage model (except for R@100) because the object name may not completely cover the semantics in the sketch and, even worse, may suffer from ambiguous object names (e.g., mouse, bat, star, etc.).

---

1. https://vl2g.github.io/projects/cstbir
2. https://chat.openai.com/

<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></td>
<td>T1K</td>
<td>T5K</td>
<td>T1K</td>
<td>T5K</td>
<td>T1K</td>
</tr>
<tr>
<td>Doodle2Search</td>
<td>14.3</td>
<td>3.6</td>
<td>24.5</td>
<td>6.7</td>
<td>36.2</td>
</tr>
<tr>
<td>DeepSBIR</td>
<td>5.2</td>
<td>1.6</td>
<td>8.8</td>
<td>3.0</td>
<td>18.9</td>
</tr>
<tr>
<td>ViT-Siamese</td>
<td>20.4</td>
<td>5.2</td>
<td>34.2</td>
<td>9.9</td>
<td>51.0</td>
</tr>
<tr>
<td>CLIP</td>
<td>50.6</td>
<td>24.2</td>
<td>63.1</td>
<td>33.7</td>
<td>78.8</td>
</tr>
<tr>
<td>VisualBERT</td>
<td>23.3</td>
<td>7.6</td>
<td>35.9</td>
<td>15.4</td>
<td>40.8</td>
</tr>
<tr>
<td>ViLT</td>
<td>28.1</td>
<td>10.5</td>
<td>42.7</td>
<td>16.5</td>
<td>60.2</td>
</tr>
<tr>
<td>CLIP</td>
<td>50.6</td>
<td>24.2</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>9.3</td>
<td>35.6</td>
<td>14.8</td>
<td>42.3</td>
</tr>
<tr>
<td>Two-stage</td>
<td>67.0</td>
<td>34.8</td>
<td>77.4</td>
<td>46.9</td>
<td>88.6</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>85.5</td>
</tr>
<tr>
<td>STNet (Ours)</td>
<td>73.7</td>
<td>38.7</td>
<td>80.6</td>
<td>50.0</td>
<td>89.4</td>
</tr>
</tbody>
</table>

³https://chat.openai.com/
The two-stage model (desc) is expected to avoid some of the drawbacks of the two-stage model. However, descriptions of object names are often not natural (e.g., a description for “grass” is “green plant used for landscaping and grazing animals”) and are still quite generic. Similarly, consider objects like boat, yacht, ship, and ferry. It is difficult to describe these in a differentiating manner but easy to sketch. Hence, both the two-stage model and STN are better than the two-stage model (desc).

Considering the other sketch+text-based image retrieval models, TIRG (Vo et al. 2019) and Taskformer (Sangkloy et al. 2022), our proposed model STN performs massively better. The poor performance of TIRG is because it does not use any pretraining for text. Also, the image pretraining in TIRG uses ResNet-17 (He et al. 2015) (trained on ImageNet dataset), which has been shown to lead to poorer image embeddings compared to CLIP (Radford et al. 2021). For Taskformer, we finetuned the publicly available checkpoint using ours because the initial checkpoint has been trained on a dataset where the (a) images in the collection are focused object images, unlike scene images in our dataset, (b) sketches are elaborate and not crudely drawn, and (c) text is self-contained and not incomplete. In other words, the samples on our dataset, CSTBIR, are more challenging (closer to practical settings) compared to data used to train Taskformer. We also experimented with training the Taskformer model from scratch but did not see any improvements. Finally, Taskformer does not use sketch reconstruction and object detection losses, which cater to the object-centric nature of our dataset, as shown in Table 4.

Ablation Study

Our overall STN model consists of several components. To understand the importance of each component, we perform several ablations as shown in Table 4.

We start with just the contrastive loss ($L_{CT}$) computed using sketch modality alone (Model 1). Model 2, which is trained with just $L_{CT}$ computed using only text modality, performs better. This broadly indicates that the infor-

<table>
<thead>
<tr>
<th>M</th>
<th>Query</th>
<th>Objective</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>1</td>
<td>S</td>
<td>$L_{CT}$</td>
<td>73.7</td>
<td>50.9</td>
<td>50.7</td>
<td>62.9</td>
<td>50.5</td>
</tr>
<tr>
<td>2</td>
<td>T</td>
<td>$L_{CT}$</td>
<td>70.3</td>
<td>57.6</td>
<td>58.7</td>
<td>68.7</td>
<td>51.0</td>
</tr>
<tr>
<td>3</td>
<td>T+S</td>
<td>$L_{CT}$</td>
<td>74.1</td>
<td>59.2</td>
<td>59.3</td>
<td>71.2</td>
<td>51.0</td>
</tr>
<tr>
<td>4</td>
<td>T+S</td>
<td>$L_{CT} + L_{OD}$</td>
<td>74.5</td>
<td>59.4</td>
<td>59.5</td>
<td>71.5</td>
<td>51.0</td>
</tr>
<tr>
<td>5</td>
<td>T+S</td>
<td>$L_{CT} + L_{OD} + L_{SR}$</td>
<td>74.1</td>
<td>59.2</td>
<td>59.3</td>
<td>71.2</td>
<td>51.0</td>
</tr>
<tr>
<td>6</td>
<td>T+S</td>
<td>$L_{CT} + L_{OD} + L_{SR}$</td>
<td>74.5</td>
<td>59.4</td>
<td>59.5</td>
<td>71.5</td>
<td>51.0</td>
</tr>
<tr>
<td>7</td>
<td>T+S</td>
<td>$L_{CT} + L_{OD} + L_{SR}$</td>
<td>74.5</td>
<td>59.4</td>
<td>59.5</td>
<td>71.5</td>
<td>51.0</td>
</tr>
</tbody>
</table>

Table 4: Ablation study for STN on Test-1K set based on query modalities and training objectives. Models (M) 1 and 2 are text-only (T) and sketch-only (S) query-based methods, resp. Models 3-6 denote objective-based ablations. Model 7 is our final model. ($L_{CT}$: contrastive loss, $L_{CLS}$: classification loss, $L_{OD}$: object-detection loss, and $L_{SR}$: sketch-reconstruction loss). Higher values are preferred for recall and lower ones for MdR. $L_{CLS} = L_{T} + L_{CLS}$. 

Table 5: Performance of image retrieval for object classes that are unseen during training. This measures the ability of the baselines to generalize concepts outside of the training domain. We evaluate this on an Open-Category test set of 750 samples containing 70 unseen object classes.

<table>
<thead>
<tr>
<th>Method</th>
<th>R@10↑</th>
<th>R@25↑</th>
<th>R@50↑</th>
<th>R@100↑</th>
<th>MdR↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-Siamese</td>
<td>61.3</td>
<td>8.6</td>
<td>14.5</td>
<td>23</td>
<td>241.0</td>
</tr>
<tr>
<td>CLIP</td>
<td>21.6</td>
<td>50.6</td>
<td>39.4</td>
<td>47.6</td>
<td>71.0</td>
</tr>
<tr>
<td>Two-Stage</td>
<td>29.0</td>
<td>38.2</td>
<td>48.8</td>
<td>54.8</td>
<td>63.0</td>
</tr>
<tr>
<td>STN (Ours)</td>
<td>37.2</td>
<td>45.3</td>
<td>62.3</td>
<td>71.7</td>
<td>27.5</td>
</tr>
</tbody>
</table>

Results on Open-Category Test Set

In a real-world scenario, the objects in queries may be uncommon or entirely unfamiliar. Considering that the Visual Genome focuses solely on common objects, we curate an Open-Category test set featuring 70 novel object categories under nine overarching classes. Among these, 50 are rare objects that are challenging to name but simple to illustrate, examples being Numbat, Mangosteen, Feijoa, Draw Knife, and Gibraltar Campion. These objects and their corresponding sketches are entirely unseen in the training set. The classes are mentioned in the Appendix. This set includes 750 composite queries and 1K gallery images.

Table 5 showcases results for this experiment, comparing STN to the top sketch-only (ViT-Siamese), text-only (CLIP), and sketch+text (Two-Stage) baselines. Although STN is naturally extensible to novel object categories, the two-stage model requires a pre-defined universe of possible objects to select from. Hence, we first create a set of possible object categories for the two-stage model by augmenting the train set with the 70 additional test categories. ViT can’t extend to new classes during testing, so we use zero-shot CLIP for sketch classification in the two-stage baseline. From Table 5, we observe that (i) the Open-Category setting is difficult as expected. (ii) Since STN encodes generic visual semantics from sketches, it is more robust to this complex setting than all the baselines.
Search Queries

- on a slide being fed red ice cream
- Bearded man on the bank of a river playing besides a man playing tabla.
- Person dressed in a suit standing beside a
- Pair of feeding on green grass.

Top-5 Retrieved Results

Figure 4: Qualitative results of our STN-ET. We show top-5 retrieved results for the multimodal (sketch+text) queries shown in left most column. From top to bottom, the sketch are for capybara, sitar, penny-farthing, and okapi. The ground truth image is shown with a green frame. (Best viewed in color).

<table>
<thead>
<tr>
<th>Method</th>
<th>R@10↑</th>
<th>R@25↑</th>
<th>R@50↑</th>
<th>R@100↑</th>
<th>MdR↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-Siamese</td>
<td>41.5</td>
<td>50.3</td>
<td>58.6</td>
<td>63.1</td>
<td>17.0</td>
</tr>
<tr>
<td>CLIP</td>
<td>50.6</td>
<td>63.1</td>
<td>78.8</td>
<td>86.7</td>
<td>10.0</td>
</tr>
<tr>
<td>Two-Stage</td>
<td>61.4</td>
<td>72.5</td>
<td>82.8</td>
<td>89.3</td>
<td>7.0</td>
</tr>
<tr>
<td>STN-ET (Ours)</td>
<td>70.3</td>
<td>81.8</td>
<td>90.7</td>
<td>95.6</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Table 6: Performance of image retrieval for examples with instance-level sketches. This measures the ability of the baselines to generalize to rich sketches with pose, size, and shape information. We evaluate this on Test-1K (where rich ones have replaced crude sketches).

Performance with Instance-Level Sketches

We have primarily focused on crude sketches. How does STN-ET fare with detailed instance-level sketches—those with pose, size, and shape details? Such sketches require the retrieved image to have a matching object instance. For this experiment, we generate rich synthetic sketches automatically for each image in the Train and Test-1K datasets using the method proposed in (Li et al. 2019). We obtain sketches only for the part of the image covered by the relevant object box. Table 6 shows that STN-ET outperforms all baselines on this complex setting as well. As the sketch becomes more expressive, the two-stage model, converting the sketch to a category name, loses nuanced details, widening its gap with STN-ET. More details are in the Appendix.

Qualitative Analysis

We show a few retrieval results from the CSTBIR dataset for our model STN-ET in Figure 4. Our model correctly retrieves the ground truth image associated with each composite query and ranks several relevant images in the top results. We observe that it can even reason about certain complex visual attributes associated with the queried object (e.g., “okapi” with striped legs). We provide more analyses in the appendix.

Conclusion

We proposed the novel problem of multimodal query-based retrieval on a collection of natural scene images where the query consists of an incomplete text and an accompanying rough sketch. Towards this task, we contributed a novel dataset, CSTBIR, containing ~2M queries and ~103K natural scene images. Further, we also proposed a novel model, STN-ET, which is trained on losses specially designed for the CSTBIR problem: contrastive loss, object classification loss, sketch-guided object detection loss, and sketch reconstruction loss. STN-ET outperforms existing strong baselines by significant margins. CSTBIR could be essential in multiple real-world settings. For example, searching for a product in digital catalogs given its rough sketch and a short description. It can also aid in the search for missing people, given their prominent features with accompanying descriptions from a repository of crowd photos taken by surveillance cameras.

Acknowledgements

This work is supported by the Science and Engineering Research Board (SERB) Grant. DST No: SRG/2021/001948.
References

Baldnati, A.; Bertini, M.; Uricchio, T.; and Del Bimbo, A. 2022. Effective conditioned and composed image retrieval combining CLIP-based features. In CVPR.


Bhunia, A. K.; Sain, A.; Shah, P. H.; Gupta, A.; Chowdhury, P. N.; Xiang, T.; and Song, Y.-Z. 2022. Adaptive fine-grained sketch-based image retrieval. In ECCV.


Changpinyo, S.; Pont-Tuset, J.; Ferrari, V.; and Soricut, R. 2021. Telling the what while pointing to the where: Multimodal queries for image retrieval. In ICCV.


Chowdhury, P. N.; Sain, A.; Bhunia, A. K.; Xiang, T.; Gryaditskaya, Y.; and Song, Y.-Z. 2022. FS-COCO: Towards understanding of freehand sketches of common objects in context. In ECCV.


Ha, D.; and Eck, D. 2018. A Neural Representation of Sketch Drawings. In ICLR.


Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duering, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In ICML.


