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

Prajwal Gatti[1], Kshitij Gopal Parikh[1], Dhriti Prasanna Paul[1], Manish Gupta[2], Anand Mishra[1]

[1]Indian Institute of Technology Jodhpur [2]Microsoft
{pgatti, parikh.2, paul.4, mishra}@iitj.ac.in, gmanish@microsoft.com

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 Network), 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 "numbats" leads to images of chipmunk, badger, weasel, mongoose, or skunk.

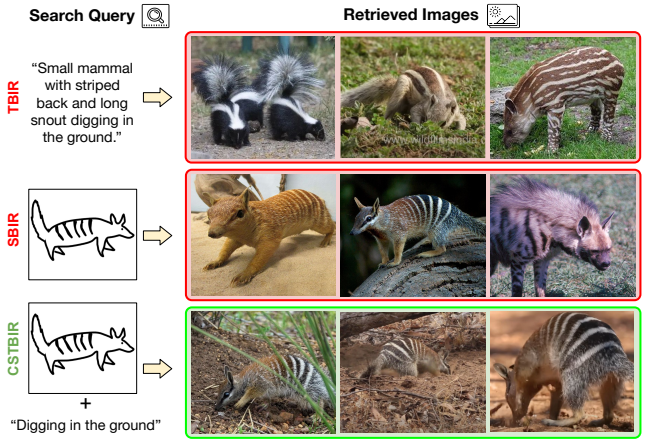

Figure 1: CSTBIR: Composite Sketch+Text Based Image Retrieval: A user wants to search "Numbat digging in the ground" but does not know the word "numbat", and the interaction "digging in the ground" is not easy to sketch.

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 **C**omposite **S**ketch+**T**ext

Based **I**mage **R**etrieval 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, or (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 – STNET 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 STNET, 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; Pang 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).

Query	Dataset	Sketch	Text	Target Image
Sketch	TU-Berlin	Object	None	Focused Object
Sketch	QMUL-Shoe-V2	Object	None	Focused Object
Text	COCO	None	Complete	Complete Scene
Text	Flickr-30K	None	Complete	Complete Scene
Sketch+Text	FS COCO	Scene	Complete	Complete Scene
Sketch+Text	CSTBIR (Ours)	Object	Complementary	Complete Scene

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.

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

Figure 2: Examples from our dataset – CSTBIR. It contains queries composed of a sketch of an object, a natural language text describing its attributes and interactions, and the target natural scene image containing the object. Queried objects from left to right: markhor, bodhran, and penny-farthing (Best viewed in color).

Property	Value
Average sentence length (in words/tokens)	5.4 / 7.7
Number of Unique Images	108K
Number of Unique Sketches	562K
Number of Unique Object Categories	258
Number of Training Instances	1.89M
Number of Validation Instances	97K
Number of Test Instances	5000
Avg % Area Covered by Query	36.7

Table 2: CSTBIR Dataset Statistics

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 $\sim$2.6x times more sketch pixels compared to our dataset[1]. 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 $\mathcal{D} = \{I_i\}_{i=1}^{N}$ of natural scene images with multiple objects, the CSTBIR problem aims to rank the N images according to relevance to the composite $\langle S, T \rangle$ 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. We take the intersection of the object categories between Visual Genome and Quick, Draw! to get 258 intersecting object classes in CSTBIR. This leads to $\sim$108K natural images with $\sim$2M queries in CSTBIR. We pair each query from Visual Genome with the corresponding object's sketch, sampled randomly from 10K sketches taken for each category from Quick, Draw!

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 $\sim$108K images and $\sim$562K sketches. The training dataset consists of $\sim$97K images, $\sim$484K sketches, and $\sim$1.89M queries. On average, the text sentences contain 5.4 words. The dataset also includes a validation set with $\sim$5K images, $\sim$83K sketches, and $\sim$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

[1]For fair comparison in terms of pixels covered by the sketch strokes, we apply thinning to normalize the stroke width for both datasets: (Sangkloy et al. 2022) and ours.

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^T_{CLS} and sketch embedding h^S_{CLS}.

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 (224×224) and then spatially sliced into a 16×16 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 $\tilde{H}^I$. The obtained values represent attention scores α_{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^I_{AVG}. Mathematically, $\alpha_{IS} = \text{Softmax}(\tilde{H}^I \times h^S_{CLS})$, $H^I = \alpha_{IS} \odot \tilde{H}^I$ and $h^I_{AVG} = \frac{1}{m}\sum_{i=1}^{m} H^I_i$, where h^I_{AVG} represents the global average pooled embedding of the image encoder.

STNET Training

STNET follows multiple task-specific training objectives.

Contrastive Training ($\mathcal{L}_{CT}$) 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} and h^I_{AVG} to obtain the contrastive loss ($\mathcal{L}_{CT}$) 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.

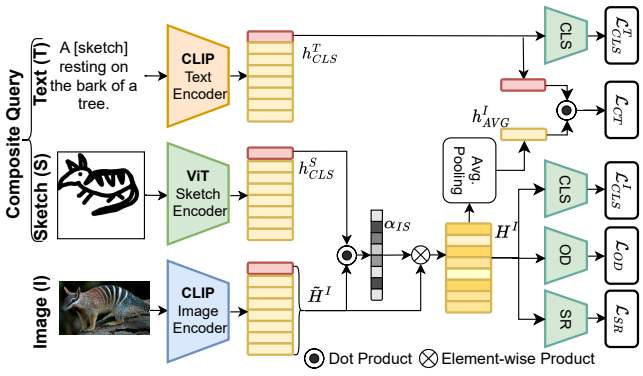

Figure 3: Overview of the proposed method, STNET for the CSTBIR problem.

Object Classification ($\mathcal{L}^T_{CLS}$ and $\mathcal{L}^I_{CLS}$) 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 $\mathcal{L}^T_{CLS}$ and $\mathcal{L}^I_{CLS}$, respectively.

Sketch-Guided Object Detection ($\mathcal{L}_{OD}$) 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 16×16 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 ($\mathcal{L}_{OD}$) loss as done in (Redmon et al. 2016).

Sketch Reconstruction ($\mathcal{L}_{SR}$) 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-

ure 3. 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 $\mathcal{L}_{SR} = \alpha\mathcal{L}_{BCE} + \beta\mathcal{L}_{DICE}$.

Our overall loss is the sum of all five losses $\mathcal{L}_{CT}+\mathcal{L}^{T}_{CLS}+\mathcal{L}^{I}_{CLS} + \mathcal{L}_{OD} + \mathcal{L}_{SR}$.

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 page[2].

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-3.5[3] 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.

	Method	R@10 ↑		R@20 ↑		R@50 ↑		R@100 ↑		MdR ↓	
		T1K	T5K	T1K	T5K	T1K	T5K	T1K	T5K	T1K	T5K
Sketch	Doodle2Search	14.3	3.6	24.5	6.7	36.2	14.5	45.7	24.4	129.0	573.5
	DeepSBIR	5.2	1.6	8.8	3.0	18.9	5.7	27.4	9.5	258.5	1288.0
	ViT-Siamese	20.4	5.2	34.2	9.9	51.0	22.2	62.6	34.9	48.0	233.0
Text	VisualBERT	23.3	7.6	35.9	15.4	40.8	27.8	54.0	40.2	46.0	246.0
	ViLT	28.1	10.5	42.7	16.5	60.2	30.1	74.3	43.8	30.0	163.0
	CLIP	50.6	24.2	63.1	33.7	78.8	49.1	86.7	62.5	10.0	52.0
Sketch+Text	TIRG	31.9	10.4	44.2	17.3	62.8	31.6	73.2	45.4	27.5	128.0
	Taskformer	22.4	9.3	35.6	14.8	42.3	27.6	53.8	38.3	48.0	204.0
	Two-stage	67.0	34.8	77.4	46.9	88.6	64.7	**93.7**	**76.2**	5.0	24.0
	Two-stage (desc)	60.1	30.5	73.7	41.7	85.5	59.6	91.6	72.0	7.0	32.0
	STNET (Ours)	**73.7**	**38.7**	**80.6**	**50.0**	**89.4**	**64.6**	93.5	74.5	**3.0**	**20.5**

Table 3: Image retrieval results on CSTBIR Test-1K (T1K) and Test-5K (T5K). Higher values are preferred for R@K (Recall@K) and lower for MdR (Median Rank).

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.).

[2]https://vl2g.github.io/projects/cstbir

[3]https://chat.openai.com/

M	Query	Objective	R@10	R@20	R@50	R@100	MdR
1	S	$\mathcal{L}_{CT}$	20.2	33.7	50.9	62.9	50.5
2	T	$\mathcal{L}_{CT}$	50.6	63.1	78.8	86.7	10.0
3	T+S	$\mathcal{L}_{CT}$	68.4	77.2	85.6	89.8	5.0
4	T+S	$\mathcal{L}_{CT} + \mathcal{L}_{OD} + \mathcal{L}_{SR}$	69.4	80.4	85.6	90.4	5.0
5	T+S	$\mathcal{L}_{CT} + \mathcal{L}_{CLS} + \mathcal{L}_{SR}$	70.4	79.6	86.2	91.1	5.0
6	T+S	$\mathcal{L}_{CT} + \mathcal{L}_{CLS} + \mathcal{L}_{OD}$	71.2	79.0	87.0	93.0	4.0
7	T+S	$\mathcal{L}_{CT} + \mathcal{L}_{CLS} + \mathcal{L}_{OD} + \mathcal{L}_{SR}$	**73.7**	**80.6**	**89.4**	**93.5**	**3.0**

Table 4: Ablation study for STNET 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. ($\mathcal{L}_{CT}$: contrastive loss, $\mathcal{L}_{CLS}$: classification loss, $\mathcal{L}_{OD}$: object-detection loss, and $\mathcal{L}_{SR}$: sketch-reconstruction loss). Higher values are preferred for recall and lower ones for MdR. $\mathcal{L}_{CLS}$=$\mathcal{L}^{T}_{CLS}$+$\mathcal{L}^{I}_{CLS}$.

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 STNET 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 STNET 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 ouris 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 STNET 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 ($\mathcal{L}_{CT}$) computed using sketch modality alone (Model 1). Model 2, which is trained with just $\mathcal{L}_{CT}$ computed using only text modality, performs better. This broadly indicates that the information in text is higher than in sketch, which makes sense since our sketches are quite rough. Using text and sketch for contrastive loss computation (Model 3) leads to further improvements. Note that we do not perform dot-product attention between sketch and image in Model 1; rather, we employ contrastive learning between their encoders. Our full proposed model, STNET (Model 7), consists of all the loss functions: contrastive loss ($\mathcal{L}_{CT}$), object classification loss using text encodings ($\mathcal{L}^{T}_{CLS}$), object classification loss using image encodings ($\mathcal{L}^{I}_{CLS}$), sketch-guided object detection loss ($\mathcal{L}_{OD}$) and sketch reconstruction loss ($\mathcal{L}_{SR}$). Models 4, 5 and 6 are trained by removing classification ($\mathcal{L}^{T}_{CLS}$+$\mathcal{L}^{I}_{CLS}$) loss, object detection loss ($\mathcal{L}_{OD}$) and sketch reconstruction loss ($\mathcal{L}_{SR}$) respectively from the overall STNET model. Broadly, removing any of the three losses leads to degradation in performance across all metrics compared to the full STNET model (Model 7). The degradation worsens when the $\mathcal{L}_{CLS}$ is removed (Model 4).

Method	R@10↑	R@25↑	R@50↑	R@100↑	MdR↓
ViT-Siamese	6.3	8.6	14.5	23	241.0
CLIP	21.6	30.6	39.4	47.6	71.0
Two-Stage	29.0	38.2	48.8	54.8	63.0
STNET (Ours)	**37.2**	**45.3**	**62.3**	**71.7**	**27.5**

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.

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 STNET to the top sketch-only (ViT-Siamese), text-only (CLIP), and sketch+text (Two-Stage) baselines. Although STNET 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 STNET encodes generic visual semantics from sketches, it is more robust to this complex setting than all the baselines.

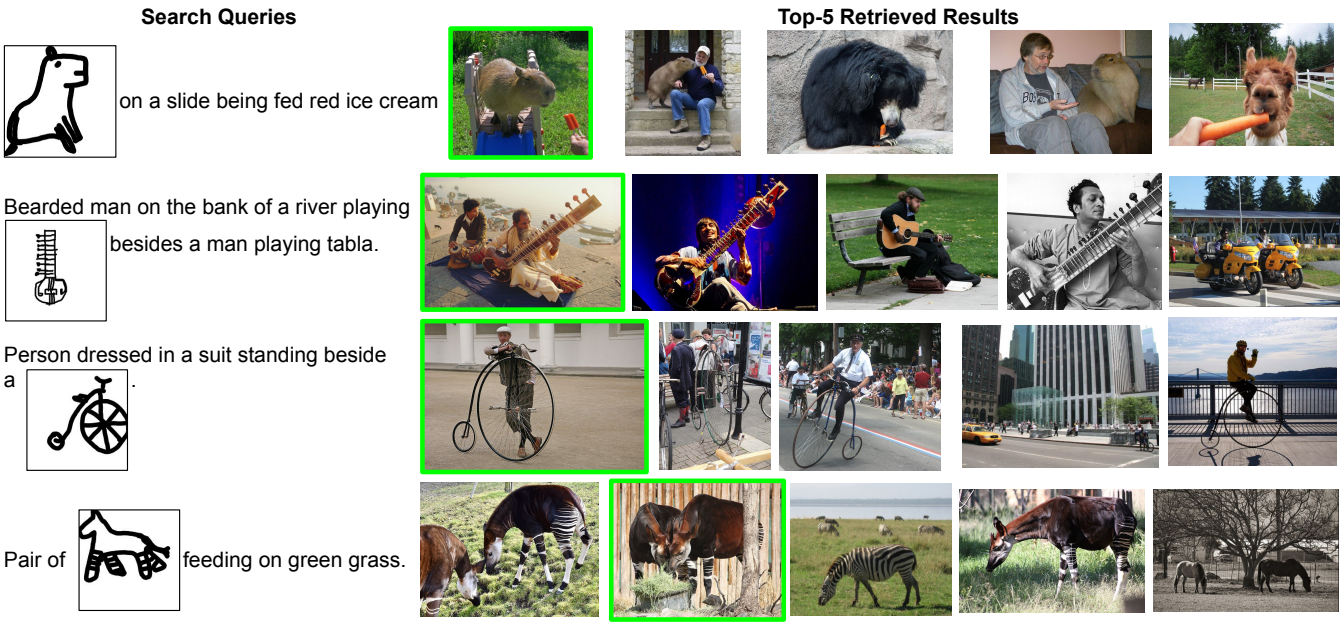

Figure 4: Qualitative results of our STNET. We show top-5 retrieved results for the multimodal (sketch+text) queries shown in left most column. From top to bottom, the sketch are for capybara, sitar, penny-farthing, and okapi. The ground truth image is shown with a green frame. (Best viewed in color).

Method	R@10↑	R@25↑	R@50↑	R@100↑	MdR↓
ViT-Siamese	41.5	50.3	58.6	63.1	17.0
CLIP	50.6	63.1	78.8	86.7	10.0
Two-Stage	61.4	72.5	82.8	89.3	7.0
STNET (Ours)	**70.3**	**81.8**	**90.7**	**95.6**	**3.0**

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 STNET 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 STNET 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 STNET. More details are in the Appendix.

Qualitative Analysis

We show a few retrieval results from the CSTBIR dataset for our model STNET 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, STNET, 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. STNET 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

Baldrati, 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.; Chowdhury, P. N.; Yang, Y.; Hospedales, T. M.; Xiang, T.; and Song, Y.-Z. 2021. Vectorization and rasterization: Self-supervised learning for sketch and handwriting. 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*.

Bhunia, A. K.; Yang, Y.; Hospedales, T. M.; Xiang, T.; and Song, Y.-Z. 2020. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *CVPR*.

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.; Bhunia, A. K.; Sain, A.; Koley, S.; Xiang, T.; and Song, Y.-Z. 2023a. SceneTrilogy: On Human Scene-Sketch and its Complementarity with Photo and Text. In *CVPR*.

Chowdhury, P. N.; Bhunia, A. K.; Sain, A.; Koley, S.; Xiang, T.; and Song, Y.-Z. 2023b. What Can Human Sketches Do for Object Detection? In *CVPR*.

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*.

Collomosse, J.; Bui, T.; and Jin, H. 2019. Livesketch: Query perturbations for guided sketch-based visual search. In *CVPR*.

Dey, S.; Riba, P.; Dutta, A.; Llados, J.; and Song, Y.-Z. 2019. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*.

Dodds, E.; Culpepper, J.; Herdade, S.; Zhang, Y.; and Boakye, K. 2020. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Dutta, A.; and Akata, Z. 2019. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*.

Eitz, M.; Hays, J.; and Alexa, M. 2012. How do humans sketch objects? *ACM Transactions on graphics (TOG)*.

Gomez, R.; Gibert, J.; Gomez, L.; and Karatzas, D. 2020. Location sensitive image retrieval and tagging. In *ECCV*.

Guo, X.; Wu, H.; Cheng, Y.; Rennie, S.; Tesauro, G.; and Feris, R. 2018. Dialog-based interactive image retrieval. *NeurIPS*.

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

Han, X.; Wu, Z.; Huang, P. X.; Zhang, X.; Zhu, M.; Li, Y.; Zhao, Y.; and Davis, L. S. 2017. Automatic spatially-aware fashion concept discovery. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *CVPR*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*.

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

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*.

Kovashka, A.; Parikh, D.; and Grauman, K. 2012. Whittlesearch: Image search with relative attribute feedback. In *CVPR*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *ECCV*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2020a. What Does BERT with Vision Look At? In *ACL*.

Li, M.; Lin, Z.; Mech, R.; Yumer, E.; and Ramanan, D. 2019. Photo-sketching: Inferring contour drawings from images. In *WACV*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Ling, Z.; Xing, Z.; Zhou, J.; and Zhou, X. 2022. Conditional Stroke Recovery for Fine-Grained Sketch-Based Image Retrieval. In *ECCV*.

Liu, F.; Zou, C.; Deng, X.; Zuo, R.; Lai, Y.-K.; Ma, C.; Liu, Y.-J.; and Wang, H. 2020. Scenesketcher: Fine-grained image retrieval with scene sketches. In *ECCV*.

Liu, L.; Shen, F.; Shen, Y.; Liu, X.; and Shao, L. 2017. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*.

Nakatsuka, T.; Hamasaki, M.; and Goto, M. 2023. Content-Based Music-Image Retrieval Using Self-and Cross-Modal Feature Embedding Memory. In *WACV*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pandey, A.; Mishra, A.; Verma, V. K.; Mittal, A.; and Murthy, H. 2020. Stacked adversarial network for zero-shot sketch based image retrieval. In *WACV*.

Pang, K.; Li, K.; Yang, Y.; Zhang, H.; Hospedales, T. M.; Xiang, T.; and Song, Y.-Z. 2019. Generalising fine-grained sketch-based image retrieval. In *CVPR*.

Pang, K.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2017. Cross-domain Generative Learning for Fine-Grained Sketch-Based Image Retrieval. In *BMVC*.

Pont-Tuset, J.; Uijlings, J.; Changpinyo, S.; Soricut, R.; and Ferrari, V. 2020. Connecting vision and language with localized narratives. In *ECCV*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *CVPR*.

Ribeiro, L. S. F.; Bui, T.; Collomosse, J. P.; and Ponti, M. A. 2020. Sketchformer: Transformer-Based Representation for Sketched Structure. *CVPR*.

Ridnik, T.; Baruch, E. B.; Noy, A.; and Zelnik, L. 2021. ImageNet-21K Pretraining for the Masses. In *NeurIPS Track on Datasets and Benchmarks*.

Sain, A.; Bhunia, A. K.; Potlapalli, V.; Chowdhury, P. N.; Xiang, T.; and Song, Y.-Z. 2022. Sketch3t: Test-time training for zero-shot sbir. In *CVPR*.

Sain, A.; Bhunia, A. K.; Yang, Y.; Xiang, T.; and Song, Y.-Z. 2021. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *CVPR*.

Sangkloy, P.; Jitkrittum, W.; Yang, D.; and Hays, J. 2022. A Sketch is Worth a Thousand Words: Image Retrieval with Text and Sketch. In *ECCV*.

Song, C. H.; Yoon, J.; Choi, S.; and Avrithis, Y. 2023. Boosting vision transformers for image retrieval. In *WACV*.

Song, J.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2017a. Fine-Grained Image Retrieval: the Text/Sketch Input Dilemma. In *BMVC*.

Song, J.; Yu, Q.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2017b. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*.

Sudre, C. H.; Li, W.; Vercauteren, T.; Ourselin, S.; and Jorge Cardoso, M. 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *DLMIA/ML-CDS@MICCAI*.

Sun, Z.; Wang, C.; Zhang, L.; and Zhang, L. 2012. Sketch2Tag: automatic hand-drawn sketch recognition. In *ACM-MM*.

Tian, Y.; Newsam, S.; and Boakye, K. 2023. Fashion Image Retrieval With Text Feedback by Additive Attention Compositional Learning. In *WACV*.

Tripathi, A.; Dani, R. R.; Mishra, A.; and Chakraborty, A. 2020. Sketch-guided object localization in natural images. In *ECCV*.

Tripathi, A.; Dani, R. R.; Mishra, A.; and Chakraborty, A. 2023. Multimodal query-guided object localization. *Multimedia Tools and Applications*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*.

Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.-J.; Fei-Fei, L.; and Hays, J. 2019. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*.

Yu, Q.; Liu, F.; Song, Y.-Z.; Xiang, T.; Hospedales, T. M.; and Loy, C.-C. 2016. Sketch Me That Shoe. In *CVPR*.

Zhang, Q.; Lei, Z.; Zhang, Z.; and Li, S. Z. 2020. Context-aware attention network for image-text retrieval. In *CVPR*.

Zhou, W.; Li, H.; and Tian, Q. 2017. Recent advance in content-based image retrieval: A literature survey. *arXiv preprint arXiv:1706.06064*.