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

Implementation Details for STNET

For all baseline models, we use hyper-parameters, as mentioned in their respective papers. For the proposed model, we optimize training using the AdamW optimizer (Kingma and Ba 2015) with a learning rate of 2e-5, a warmup for 2000 steps from 1e-7, and a batch size of 200 with a gradient accumulation value of 2. The model is trained for about 120K steps on four NVIDIA A100 GPUs, which takes ∼48 hours to complete the training process.

We employ the ViT-B/16 vision transformer as the sketch-encoder. To adapt the ImageNet pretrained ViT model to the sketch modality, we fine-tune it for the sketch classification task. We train it for 2 epochs on sketch images from the train split of CSTBIR. We freeze the sketch encoder during the training of STNET since fine-tuning it with the entire model did not yield improvements, and this approach allows for faster training. For the text and image encoders, we use the pretrained CLIP (ViT-B/16) implementation and checkpoint provided by the original authors [1].

The CSTBIR Dataset Analysis

For further data analysis of CSTBIR, we performed part of speech tagging on text descriptions using the NLTK library (Bird, Klein, and Loper 2009). Fig. 1 shows word clouds for the top few adjectives (object attribute indicating words), verbs (action indicating words), and prepositions (position indicating words) for the text descriptions, respectively (left to right) in the CSTBIR dataset.

In Fig. 2, we present a few sketches used in our dataset for a selection of object categories. We observe that the hand-drawn objects may take multiple visual forms (e.g., markhor, paw paw, marimba), orientations (e.g., froe, sugar glider, skycar), and have varying levels of detail (e.g., bouzouki).

Finally, we also present some more selections of examples from the CSTBIR in Fig. 3.

Annotations for Two-stage (Desc) model

We present a few examples of object-description annotations (a mixture of both ChatGPT [2] and human annotated descriptions are provided) for the Two-Stage (desc) model:

- **Jerboa**:
  1. “Tiny desert rodent with large hind legs.”
  2. “Hopping desert creature with long tail.”
  3. “Small kangaroo-like desert mouse.”
- **Blobfish**:
  1. “Gelatinous deep-sea fish with droopy face.”
  2. “Sad-looking, squishy underwater creature.”
  3. “Jelly-like fish with downturned mouth.”
- **Titan Arum**:
  1. “Large foul-smelling flowering plant.”
  2. “Tall plant with big stinky flower.”
  3. “Huge plant with flower that smells like decay.”
- **Zeppelin**:
  1. “Huge cigar-shaped flying airship.”
  2. “Large blimp-like flying vessel.”
  3. “Giant floating airship with rigid frame.”
- **Dulcimer**:
  1. “Stringed instrument played on the lap.”
  2. “Wooden musical box with strings.”
  3. “Fretted string instrument for gentle melodies.”
- **Sun Bear**:
  1. “Small bear with crescent chest mark.”
  2. “Short-haired bear with moon-like patch.”
  3. “Tropical forest bear with golden neck.”
- **Okapi**:
  1. “Forest animal, half zebra, half giraffe.”
  2. “Striped-legged, giraffe-like jungle creature.”
  3. “Long-necked forest dweller with zebra stripes.”
- **Thorny Devil**:
  1. “Spiky lizard from Australian deserts.”
  2. “Desert reptile covered in sharp spikes.”
  3. “Dragon-like lizard with thorn-like protrusions.”
- **Proboscis Monkey**:

[1] https://github.com/openai/CLIP

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

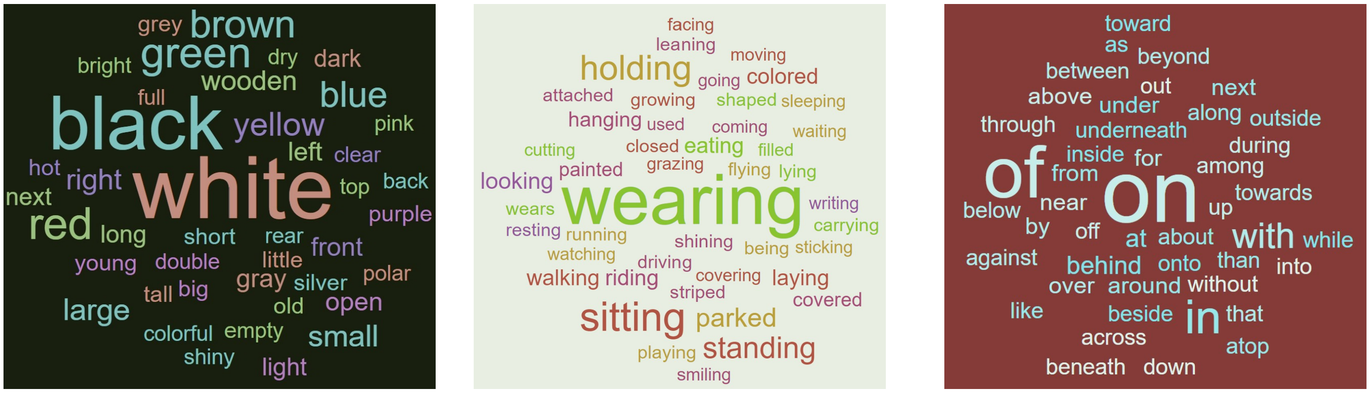

Figure 1: Word clouds for top few adjectives (attributes), verbs (action words), and prepositions (position indicating words) for the text descriptions respectively (left to right) in the CSTBIR dataset.

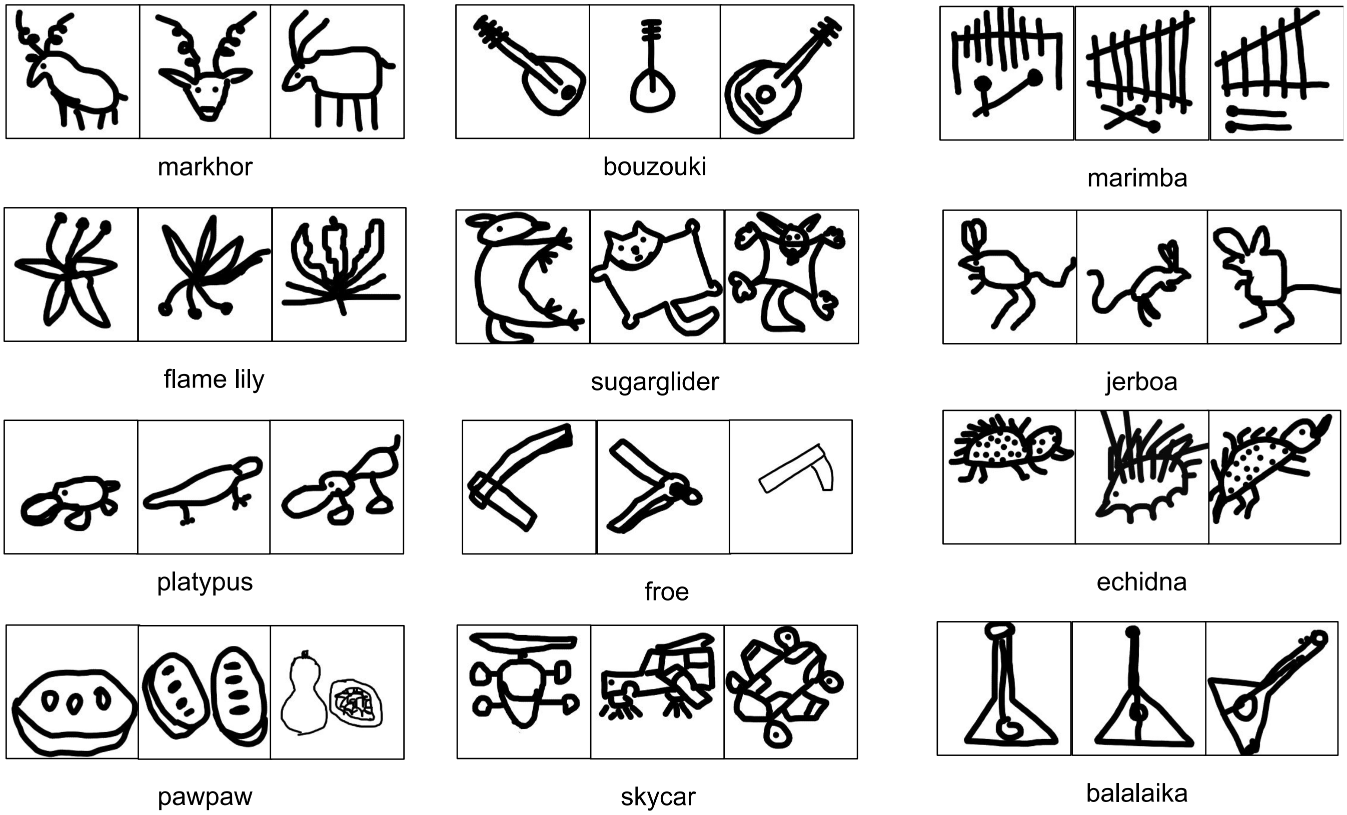

Figure 2: Sketch examples from the CSTBIR dataset, showing that the represented objects may take multiple visual forms (e.g., markhor, paw paw, marimba), orientations (e.g., froe, sugar glider, skycar), and have varying levels of detail (e.g., bouzouki).

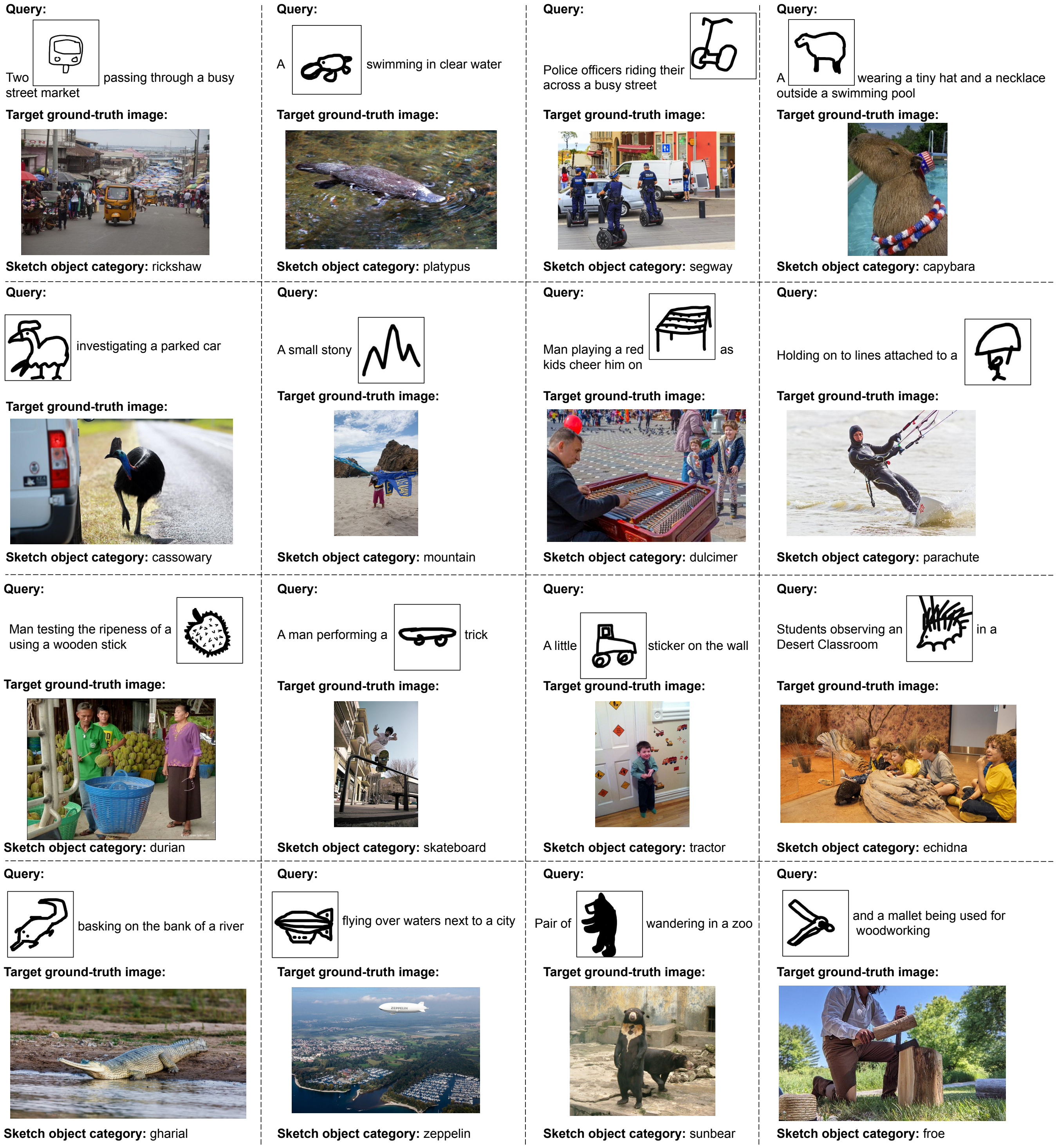
Query:
Two passing through a busy street market
Target ground-truth image:
Sketch object category: rickshaw
Query:
A swimming in clear water
Target ground-truth image:
Sketch object category: platypus
Query:
Police officers riding their across a busy street
Target ground-truth image:
Sketch object category: segway
Query:
A wearing a tiny hat and a necklace outside a swimming pool
Target ground-truth image:
Sketch object category: capybara
Query:
investigating a parked car
Target ground-truth image:
Sketch object category: cassowary
Query:
A small stony
Target ground-truth image:
Sketch object category: mountain
Query:
Man playing a red as kids cheer him on
Target ground-truth image:
Sketch object category: dulcimer
Query:
Holding on to lines attached to a
Target ground-truth image:
Sketch object category: parachute
Query:
Man testing the ripeness of a using a wooden stick
Target ground-truth image:
Sketch object category: durian
Query:
A man performing a trick
Target ground-truth image:
Sketch object category: skateboard
Query:
A little sticker on the wall
Target ground-truth image:
Sketch object category: tractor
Query:
Students observing an in a Desert Classroom
Target ground-truth image:
Sketch object category: echidna
Query:
basking on the bank of a river
Target ground-truth image:
Sketch object category: gharial
Query:
flying over waters next to a city
Target ground-truth image:
Sketch object category: zeppelin
Query:
Pair of wandering in a zoo
Target ground-truth image:
Sketch object category: sunbear
Query:
and a mallet being used for woodworking
Target ground-truth image:
Sketch object category: froe

Figure 3: A selection of examples from the CSTBIR dataset.

1. “Monkey with an unusually large nose.”
2. “Long-nosed primate from Borneo forests.”
3. “Monkey with distinctive snout and potbelly.”

- **Penny Farthing**:
  1. “Old bicycle with one huge front wheel.”
  2. “Historical bike with mismatched wheel sizes.”
  3. “Antique cycle with large front tire.”
- **Segway**:
  1. “Two-wheeled electric personal transporter.”
  2. “Stand-up battery-powered travel device.”
  3. “Self-balancing personal electric vehicle.”

Qualitative Results

Fig. 4 shows the top five retrieved images using the proposed method, STNet, and four other representative baselines for the query “black [sketch] with a yellow outline”, with a sketch of a five-pointed star. The baselines include ViT-Siamese (sketch only), CLIP (text only), Two-Stage Model, and Two-Stage (Desc) Model. We also show the ground truth image on the right. We observe that the sketch-only baseline ignores the text aspects (“black”, “yellow outline”) of the query. The text-only baseline, in general, comes up with black objects without considering the object (sketch). The two-stage model got confused and retrieves images of celebrities rather than the celestial body. The two-stage description model also does not return good results, perhaps because usually a star is “luminous,” but the query asks for a black star, thus “black luminous celestial body” cannot be easily linked to a star. Overall, our proposed method, STNet, retrieves the ground truth image nicely at the top position.

Error Analysis

To better understand the limitations of STNet, we perform an error analysis of its results on the Test-1K set and compare it with the Two-Stage method, our best baseline. We first select 100 random composite queries and their predictions and group error patterns into three categories: (i) Missing labels: the top-10 retrieved images correctly match the search query, but the labeled ground truth image is not ranked within the top-10 (ii) Misrecognized sketch category: the object class is misrecognized and (iii) Object ambiguity: the top result does not contain the object due to ambiguity in the object name (for example: ‘mouse’ refers to both a computer mouse and the animal, and ‘star’ refers to both the shape and a talented or famous entertainer or sports player). Table 1 shows that our STNet model rightly handles object ambiguity and makes fewer mistakes in recognizing the sketch category compared to the two-stage model.

We present a few failure cases of STNet corresponding to the two popular error buckets in Fig. 5. In the first example, we observe that the model fails to recognize the correct sketch object categories as it mistakes “capybara” for “bear” as evidenced by the top-5 retrievals. In the last example, we see that although the top retrievals correctly match the query, it does not contain the ground-truth target image. STNet fails to identify them as correct images due to missing annotations in the Visual Genome dataset.

Method	Missing labels	Misrecognized sketch category	Object Ambiguity
Two-stage	22	12	2
STNet (Ours)	31	9	0

Table 1: Error analysis of predictions on 100 randomly chosen samples from the CSTBIR Test-1K set.

Unseen Category Experiment Details

To create a dataset for the unseen category experiment, we picked 70 novel objects from 9 broad types listed below. Among the 70, we chose 50 “difficult-to-name” objects and the remaining 20 from Visual Genome/Quick Draw.

To obtain hand-drawn sketches for the 50 difficult-to-name objects, we obtained diverse sketches (3 per object) hand-drawn by 11 human annotators (2 of which are authors of this paper). The annotators were instructed to draw the sketches capturing unique visual aspects of the object and to draw in <20 seconds to obtain them in the style of Quick, Draw! Further, we then obtain scene images for these objects from the publicly available CC12M dataset as well as images from Wikimedia Commons [3].

The results of this experiment are provided in the main paper. Here, we show the performance of baselines on the unseen category experiment for only the 50 “difficult-to-name” objects. We find that there is good scope for improvement in this area of generalizing to novel objects.

The objects and their broad types, used in the unseen category experiment are:

- **Animals**: Okapi, Platypus, Cassowary, Pangolin, Markhor, Proboscis Monkey, Capybara, Numbat, Echidna, Blobfish, Gharial, Sugar glider, Giant Isopod, Thorny Devil, Jerboa, Sun Bear, Monkey, Octopus.
- **Fruits**: Durian, Feijoa, Carambola, Mangosteen, Pawpaw, Canistel, Kaffir Lime, Noni, Buddha’s Hand, Star Fruit.
- **Musical Instruments**: Balalaika, Dulcimer, Autoharp, Bouzouki, Crwth, Glass Armonica, Bodhran, Sitar, Chapman Stick, Marimba.
- **Tools**: Froe, Draw Knife, Blacksmith Tongs, Sundial, Saw, Drill, Dumbbell.
- **Flowers**: Flame Lily, Titan Arum, Gibraltar Campion.
- **Vehicles**: Penny Farthing, Rickshaw, Monowheel, Zeppelin, Skycar, Segway, Firetruck, Sailboat, Airplane.
- **Places**: Hospital, Church.
- **Electronics**: Television, Computer
- **Miscellaneous**: Rosetta Stone, Cake, Cactus, Tree, Fence, Finger, Couch, Envelope, Bathtub.

[3] https://commons.wikimedia.org/

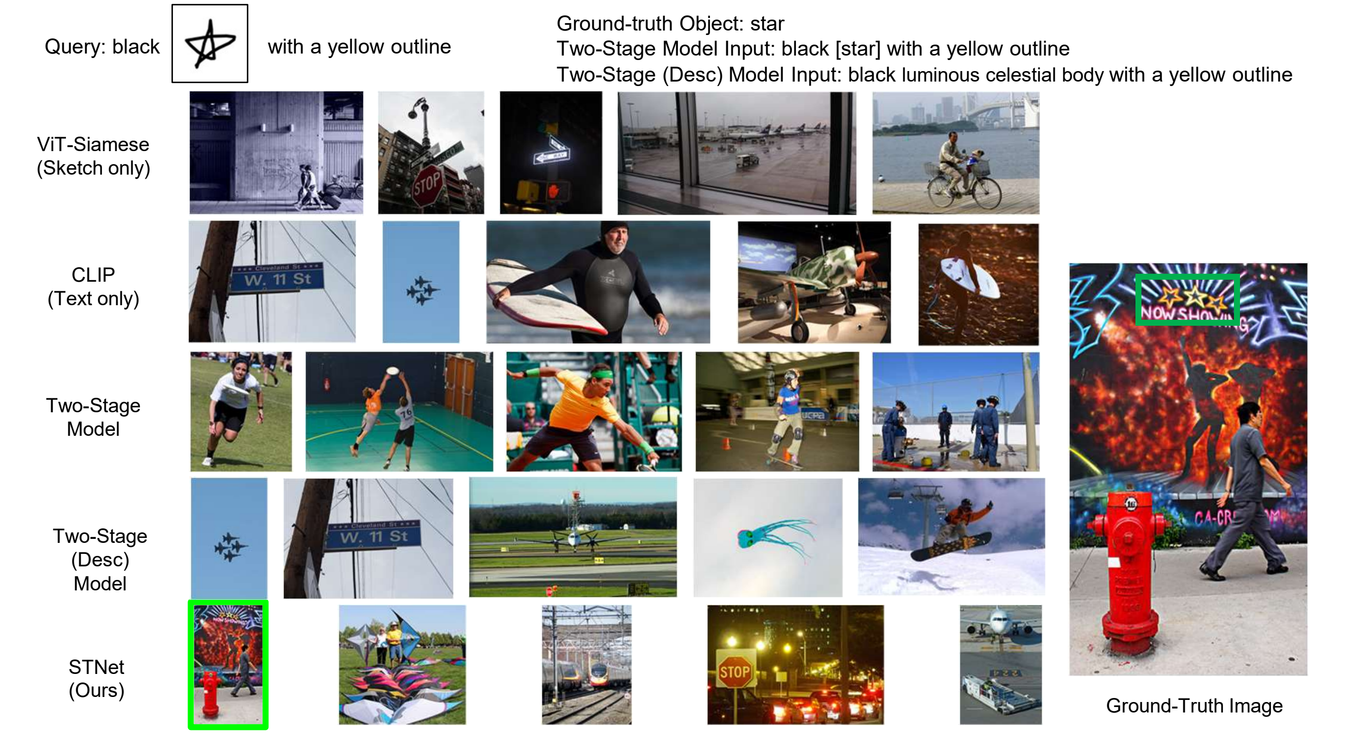
Query: black
with a yellow outline
Ground-truth Object: star
Two-Stage Model Input: black [star] with a yellow outline
Two-Stage (Desc) Model Input: black luminous celestial body with a yellow outline
ViT-Siamese (Sketch only)
CLIP (Text only)
Two-Stage Model
Two-Stage (Desc) Model
STNet (Ours)
Ground-Truth Image

Figure 4: Qualitative analysis comparing retrieved images for baselines using the query "black [sketch] with yellow outline" with a 5-pointed star sketch. STNET retrieves the correct image, while the Two-Stage model confuses "star" with famous sports players (Best viewed with zoom).

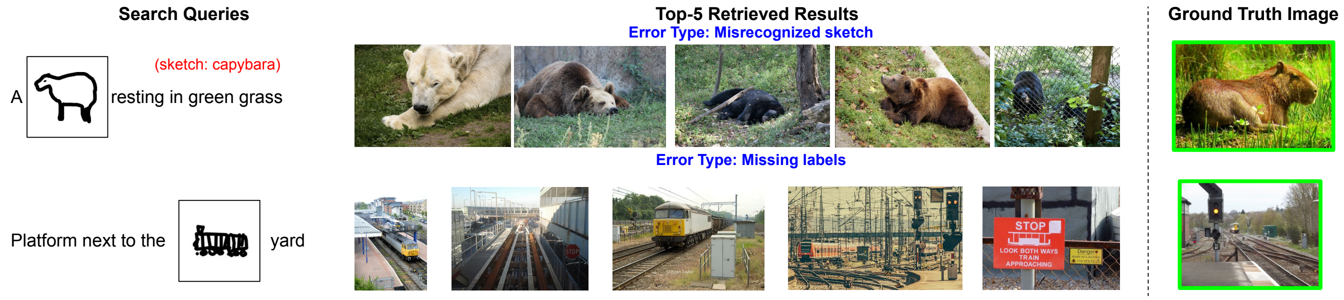

Figure 5: Two common categories of errors in STNET. We observe the model misrecognizes the sketch objects in the first example ("capybara" as a "bear"). In the last example, the model correctly retrieves matching results, but the target ground-truth image is not present among these top retrievals.

Instance-Level Retrieval Experiment Details

While the sketches we use in our work are 'crude', we ask if STNET can extend to sketches that provide more information such as pose, size, and orientation. Particularly, we aim to perform the task with instance-level sketches, i.e., a sketch with an exact visual match of the corresponding object in the target image. To obtain such sketches, we choose an automatic photo-to-sketch generator (Li et al. 2019), using the implementation and checkpoint provided in https://github.com/mtli/PhotoSketch. Sketches are obtained of only the region which focuses on the target object in the image. We create such sketches for both the training and the Test-1K dataset.

Ethical Concerns

No personally identifiable data has been used for this work. The Visual Genome data[4] is licensed under a Creative Commons Attribution 4.0 International License. QuickDraw dataset[5] is made available by Google, Inc. under the Creative Commons Attribution 4.0 International license.

References

Bird, S.; Klein, E.; and Loper, E. 2009. *Natural language processing with Python: analyzing text with the natural lan-*

[4] http://visualgenome.org/

[5] https://github.com/googlecreativelab/quickdraw-dataset

guage toolkit. "O'Reilly Media, Inc.".

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

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