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]

{pgatti, parikh.2, paul.4, mishra}@iitj.ac.in, gmanish@microsoft.com

[1]Indian Institute of Technology, Jodhpur; [2]Microsoft

Problem of Complex Queries

- How to find this? chipmunk, badger, weasel, mongoose, or skunk?

- Complex queries
  - "difficult-to-name but easy-to-draw" objects.
  - "difficult-to-sketch but easy-to-verbalize" object's attributes or interaction with the scene.
  - Query: "numbat digging in the ground"

Prajwal Gatti, Kshitij Parikh, Dhriti Paul, Manish Gupta, Anand Mishra. Composite Sketch+Text Queries for Retrieving Objects with Elusive Names and Complex Interactions. AAAI 2024.

Related Work

- Sketch-Based Image Retrieval (SBIR)
  - Methods: CNNs, Transformer-based methods, Deep Siamese models with triplet loss
  - Specialized forms: Zero Shot-SBIR, Finegrained SBIR, Category-level SBIR
- Text-Based Image Retrieval (TBIR)
  - Alignment of (query text, images) using VisualBERT, ViLT
  - Cross-attention-based models
  - Object tags in images
  - Contrastive learning methods, zero-shot learning methods
- Multimodal Query Based Image Retrieval
  - Reference images and category text for image retrieval.
  - speech and mouse traces as the query
  - Detailed sketch and text input
    - e-commerce product images using CNNs and LSTMs
    - scene images using CLIP

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

Prajwal Gatti, Kshitij Parikh, Dhriti Paul, Manish Gupta, Anand Mishra. Composite Sketch+Text Queries for Retrieving Objects with Elusive Names and Complex Interactions. AAAI 2024.

CSTBIR Problem and Dataset

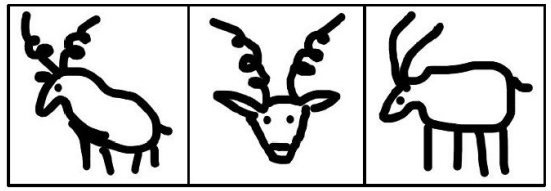
- Given: a hand-drawn sketch S, a complementary text T and a database D of N natural scene images with multiple objects
- Rank: N images according to relevance to composite ⟨S, T⟩ query.
- Natural images and text descriptions from Visual Genome.
- Sketches from Quick, Draw!
- Train (~1.89M queries, ~97K images)
- Validation (~5K images, ~97K queries)
- Test
  - Test-1K: 1K queries, 1K images
  - Test-5K: 4K queries, 5K images
  - Open-Category set: 750 queries, 70 objects, 1K images.

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

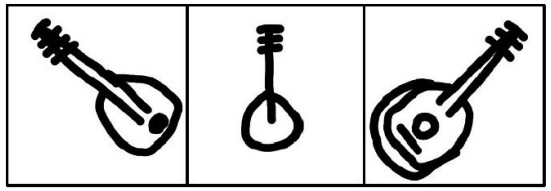
markhor, bodhran, and penny-farthing

Prajwal Gatti, Kshitij Parikh, Dhriti Paul, Manish Gupta, Anand Mishra. Composite Sketch+Text Queries for Retrieving Objects with Elusive Names and Complex Interactions. AAAI 2024.

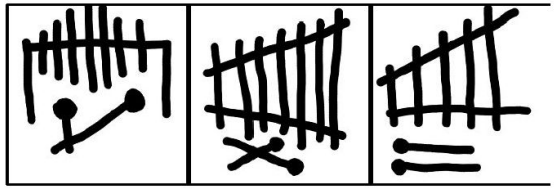
Sketches



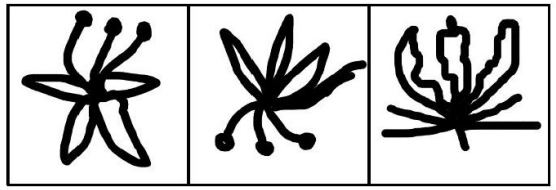
markhor



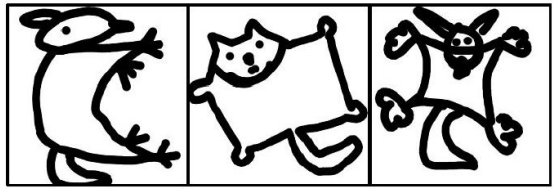
bouzouki



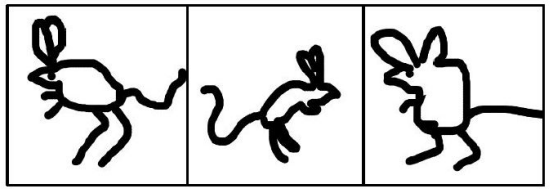
marimba



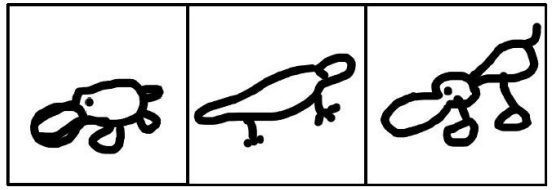
flame lily



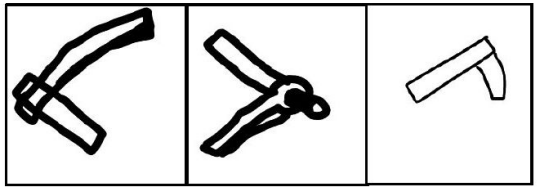
sugarglider



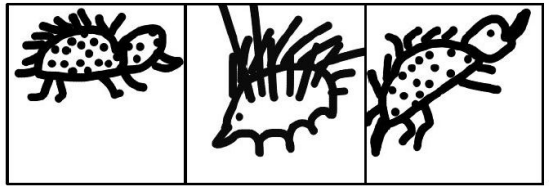
jerboa



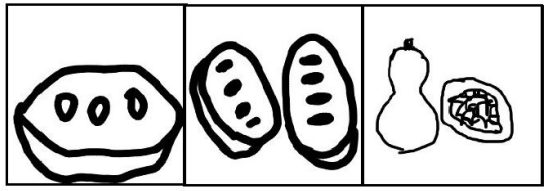
platypus



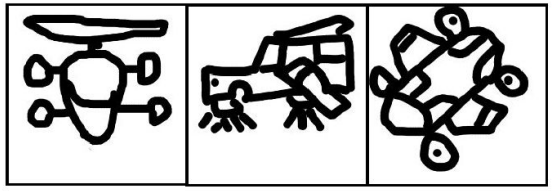
froe



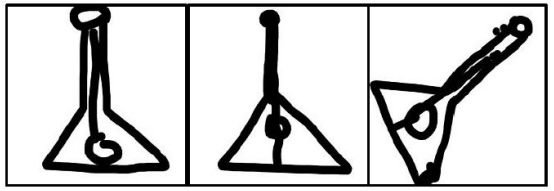
echidna



pawpaw



skycar

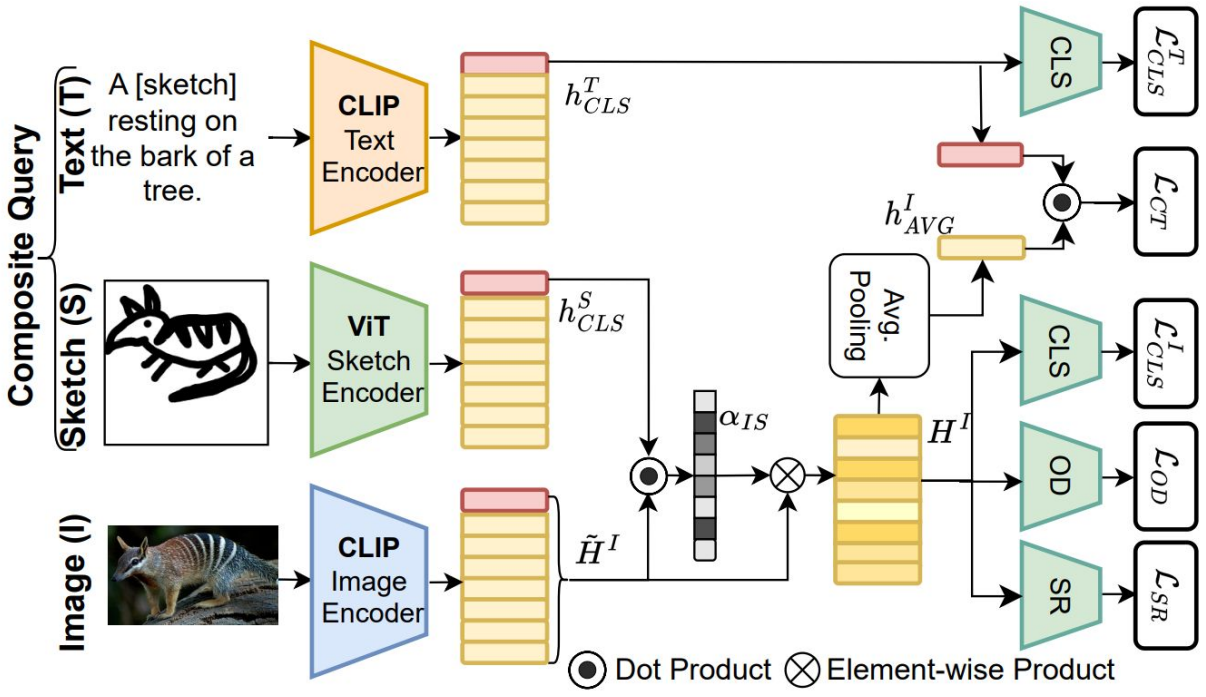


balalaika

Prajwal Gatti, Kshitij Parikh, Dhriti Paul, Manish Gupta, Anand Mishra. Composite Sketch+Text Queries for Retrieving Objects with Elusive Names and Complex Interactions. AAAI 2024.

STNet Model for CSTBIR

- Query (Sketch+Text) Encoding
  - pretrained CLIP text encoder
  - pretrained Vision Transformer (ViT) finetuned on sketches.
- Image Encoding
  - pretrained CLIP-ViT
- STNet Training
  - L_{CT}: Contrastive Training (query, image)
    - InfoCNE(h^T_{CLS}, h^I_{AVG})
  - L^T_{CLS} and L^I_{CLS}: Object Classification
  - L_{OD}: Sketch-Guided Object Detection
    - intersection over union (IoU)
  - L_{SR}: Sketch Reconstruction
    - eight blocks of Convolution-BatchNormReLU
    - Binary Cross Entropy loss and the DICE loss



$$\alpha_{IS} = \text{Softmax}(\tilde{H}^I \times h^S_{CLS})$$

$$H^I = \alpha_{IS} \odot \tilde{H}^I$$

$$h^I_{AVG} = \frac{1}{m}\sum_{i=1}^{m} H^I_i$$

Prajwal Gatti, Kshitij Parikh, Dhriti Paul, Manish Gupta, Anand Mishra. Composite Sketch+Text Queries for Retrieving Objects with Elusive Names and Complex Interactions. AAAI 2024.

Image retrieval results on Test-1K (T1K) and Test-5K (T5K)

- Sketch-based Image Retrieval (SBIR)
  - Doodle2Search and DeepSBIR
  - ViT-based Siamese: 2 ImageNet pre-trained ViT encoders for sketch and image modalities trained using InfoNCE
- Text-based Image Retrieval (TBIR)
  - VisualBERT, ViLT, CLIP
- Composite Query-based Image Retrieval
  - TIRG and Taskformer
  - 2-stage
    - ViT trained for sketch classification to get an object name
    - Insert(object name, incomplete text query) and use pretrained CLIP
  - 2-stage (desc): Insert(object description, incomplete text query)

	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**

- sketch+text > text-only > sketch-only
- STNet>2-stage
  - incomplete semantics in object name
  - Ambiguous objects: mouse, bat, star

Prajwal Gatti, Kshitij Parikh, Dhriti Paul, Manish Gupta, Anand Mishra. Composite Sketch+Text Queries for Retrieving Objects with Elusive Names and Complex Interactions. AAAI 2024.

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

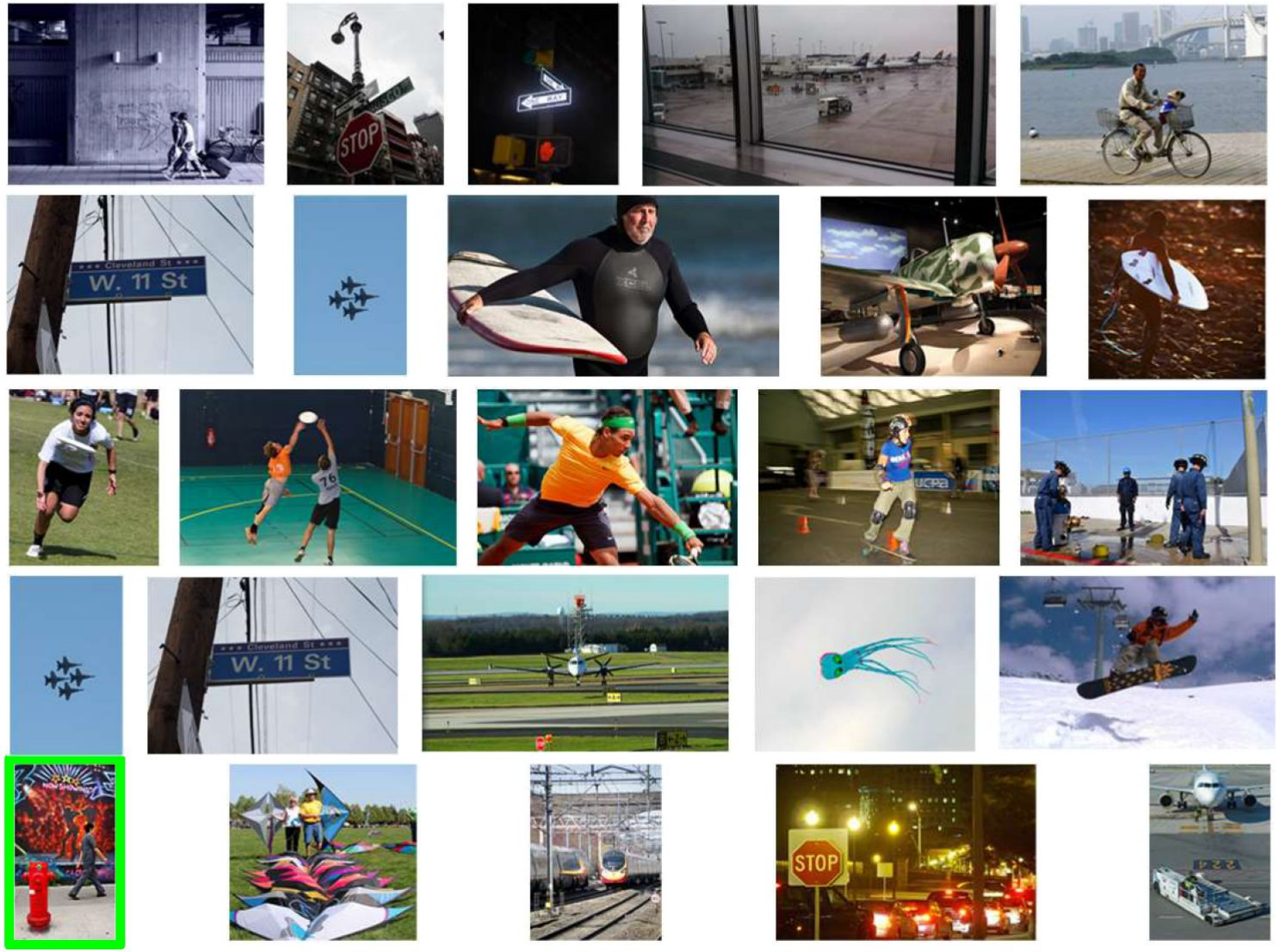
ViT-Siamese (Sketch only)

CLIP (Text only)

Two-Stage Model

Two-Stage (Desc) Model

STNet (Ours)



Ground-Truth Image

Prajwal Gatti, Kshitij Parikh, Dhriti Paul, Manish Gupta, Anand Mishra. Composite Sketch+Text Queries for Retrieving Objects with Elusive Names and Complex Interactions. AAAI 2024.

Further Experiments and Results

- Ablation study on Test-1K
  - Removing any of the 3 losses hurts.
  - Removing L_{CLS} hurts the most.

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**

- Results on Open-Category Test Set
  - Open-Category setting is difficult.
  - STNet is more robust to this complex setting.

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**

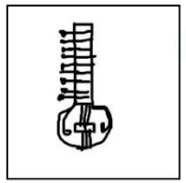
Prajwal Gatti, Kshitij Parikh, Dhriti Paul, Manish Gupta, Anand Mishra. Composite Sketch+Text Queries for Retrieving Objects with Elusive Names and Complex Interactions. AAAI 2024.

capybara, sitar, penny-farthing, and okapi.

Search Queries

on a slide being fed red ice cream

Bearded man on the bank of a river playing



besides a man playing tabla.

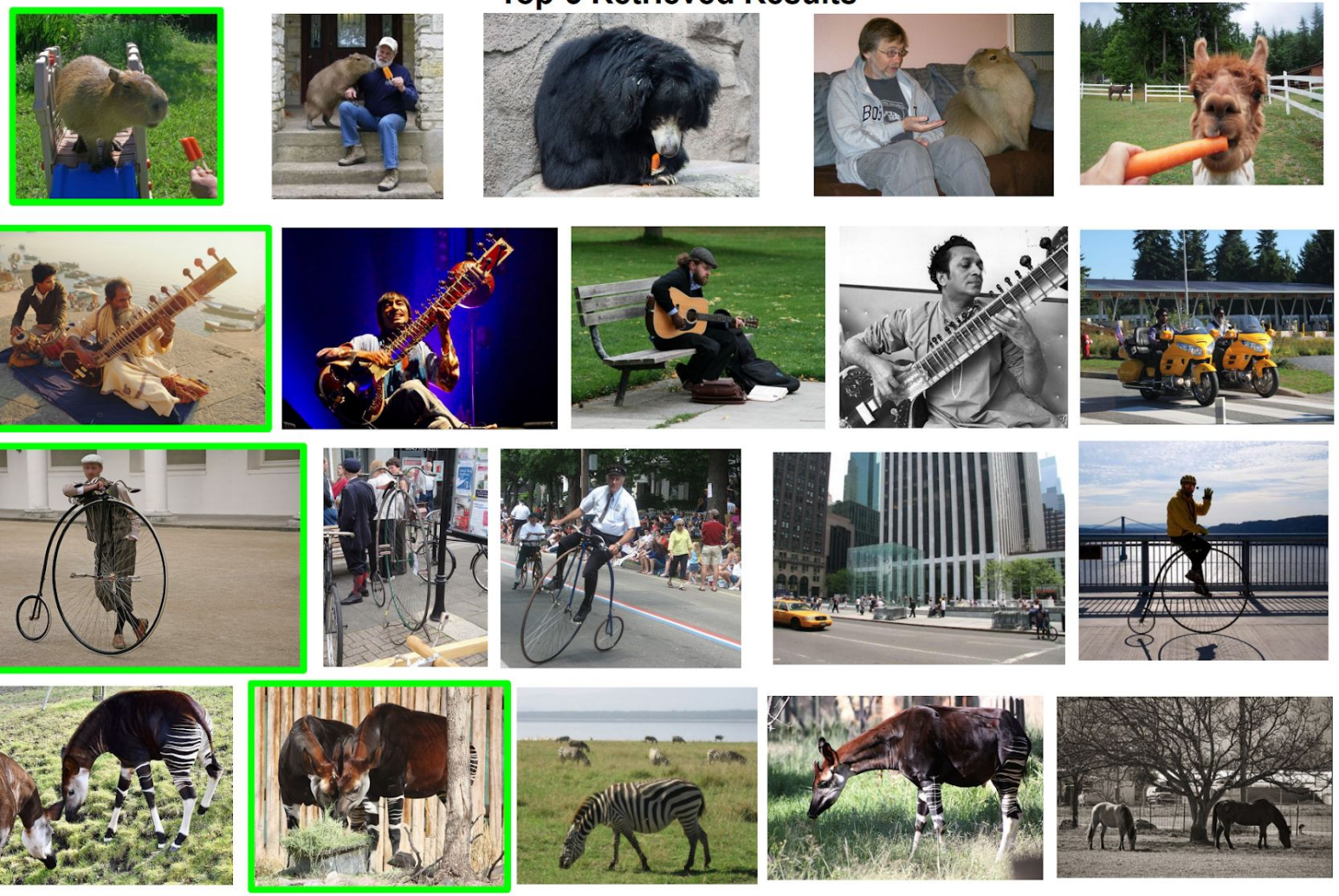
Person dressed in a suit standing beside a

.

Pair of

feeding on green grass.

Top-5 Retrieved Results



Prajwal Gatti, Kshitij Parikh, Dhriti Paul, Manish Gupta, Anand Mishra. Composite Sketch+Text Queries for Retrieving Objects with Elusive Names and Complex Interactions. AAAI 2024.

Conclusion

- CSTBIR (Composite Sketch+Text Based Image Retrieval)
  - New dataset: ~2M queries and ~108K natural scene images.
  - STNet (Sketch+Text Network)
    - Pretrained multimodal transformer
    - Uses a hand-drawn sketch to localize relevant objects in the natural scene image
    - Encodes the text and image to perform image retrieval
    - contrastive loss, object classification loss, sketch-guided object detection loss, and sketch reconstruction loss
- Search for missing people, search for a product in digital catalogs, …

- Thanks!
- LinkedIn: http://aka.ms/manishgupta
- HomePage: https://sites.google.com/view/manishg/

Prajwal Gatti, Kshitij Parikh, Dhriti Paul, Manish Gupta, Anand Mishra. Composite Sketch+Text Queries for Retrieving Objects with Elusive Names and Complex Interactions. AAAI 2024.