CS Katha Barta | ସଂଗଣକ ବିଜ୍ଞାନ କଥା ବାର୍ତା
NISER Bhubaneswar

Drawing as a means of Communication: Towards Sketch-guided Visual Understanding

Anand Mishra

CS Katha Barta | ସଂଗଣକ ବିଜ୍ଞାନ କଥା ବାର୍ତା
NISER Bhubaneswar

Drawing as a means of Communication: Towards Sketch-guided Visual Understanding

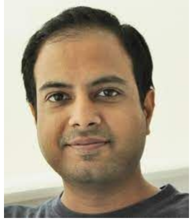
Anand Mishra

Indian Institute of Technology Jodhpur

Joint work with







While hiking in your Australia trip, you saw this animal.

A few days later …

You want to search that animal! How will you search?

Option-A

Describe the query in natural language

Option-A

Describe the query in natural language

What if

- You do not the name?
- Your linguistic skills are weak?

Option-A

Describe the query in natural language

What if

- You do not the name?
- Your linguistic skills are weak?

Option-B

Draw the query

Option-A

Describe the query in natural language

What if

- You do not the name?
- Your linguistic skills are weak?

Option-B

Draw the query

Drawing everything is non-trivial, e.g., activities, color, etc.

We take a middle option

Composite Sketch+Text Based Image Retrieval

You can …

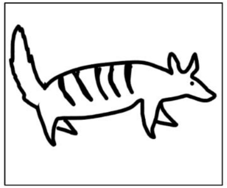
Search Query

Retrieved Images

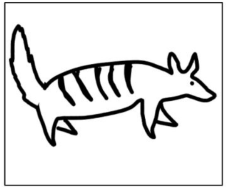
Search via **text**

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

Search via **sketch**



Search via both
Sketch + Text
(Ours)



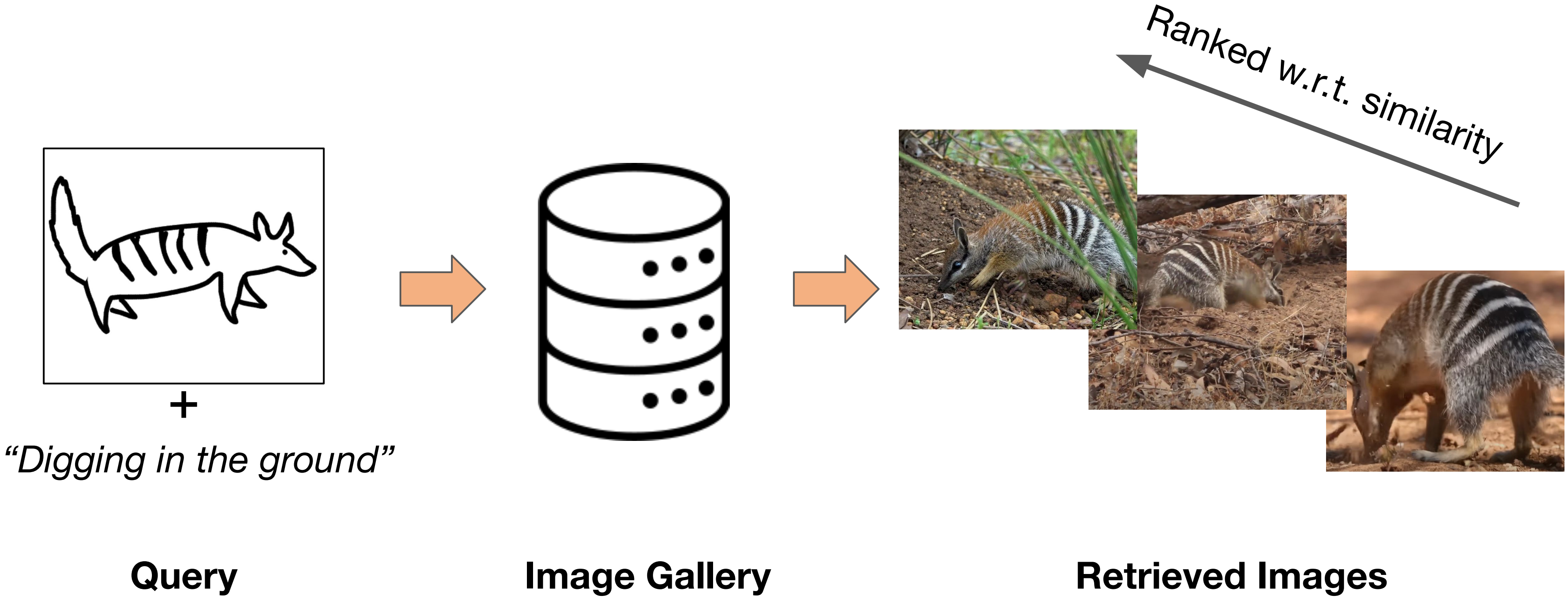
+

"Digging in the ground"

Sketches: from Stone Age to Tablet Age

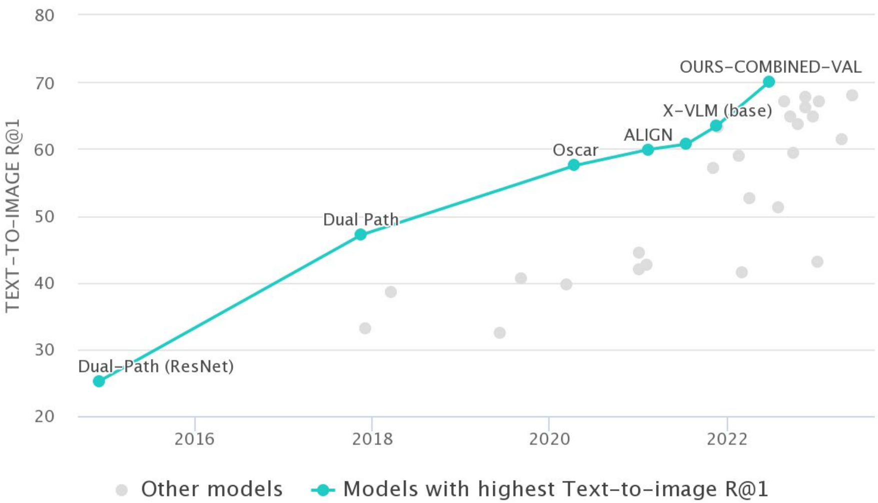
Cave hyena (*Crocuta crocuta spelaea*) painting found in the Chauvet cave (Source: Gutenberg.org) ; now known to be 32,000 year old.

The Problem: CSTBIR



Related Work: Text-based Image Retrieval

Johnson et al., CVPR'15;
Faghri et al., BMVC'18;
Zhang et al., CVPR'19,
and many more..



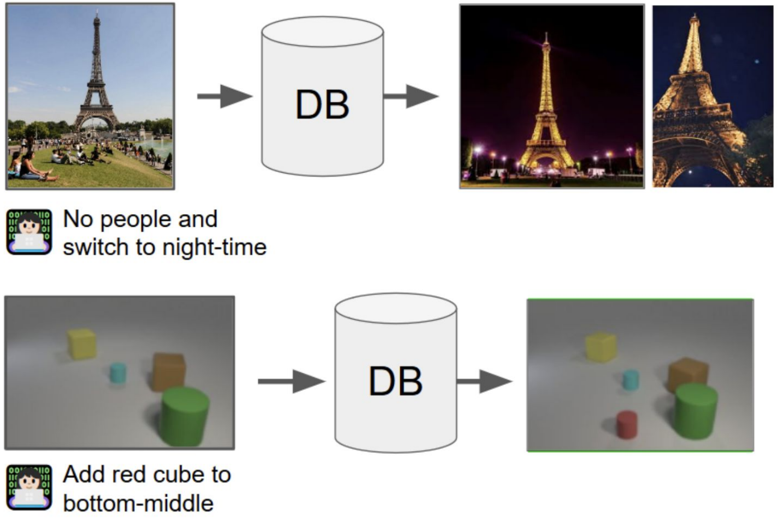
MS COCO Benchmark

T2I is a well-studied problem!

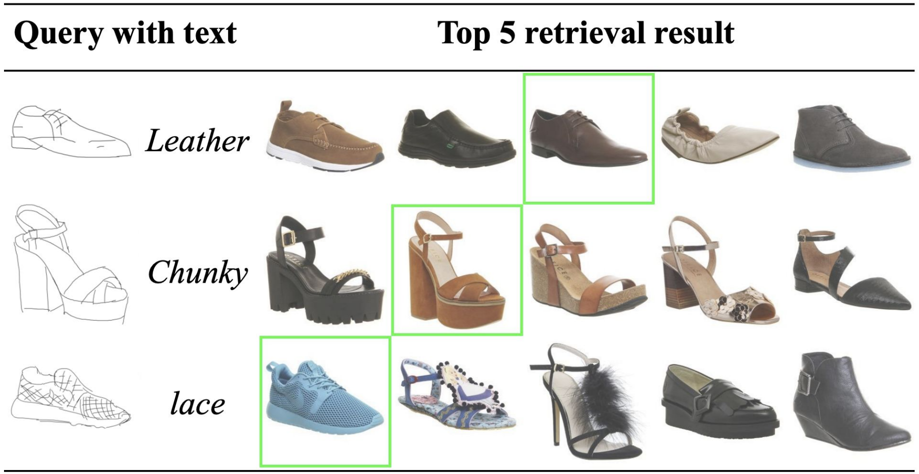
Related Work: Sketch-based Image Retrieval

Sangkloy et al., SIGGRAPH'16

Related Work: Multimodal Query for Image Retrieval



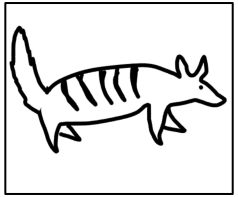
Image+Text to Image Retrieval
(Vo et al., CVPR'18)



Sketch+Tag to Image Retrieval
(Song et al., BMVC'17)

The CSTBIR Dataset

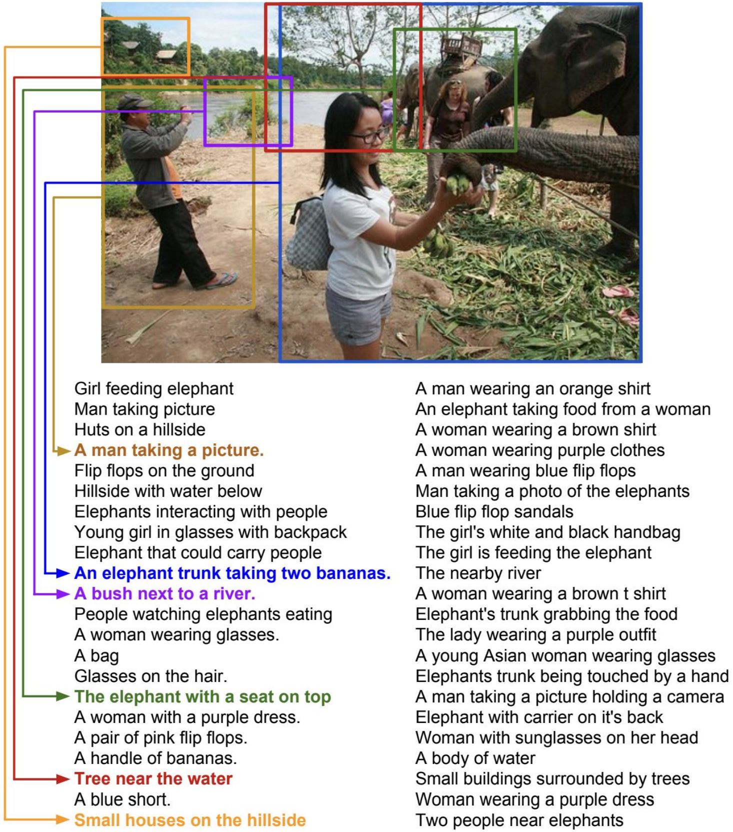
We require a dataset with tuples of the form:

( , *"Digging in the ground"* , )

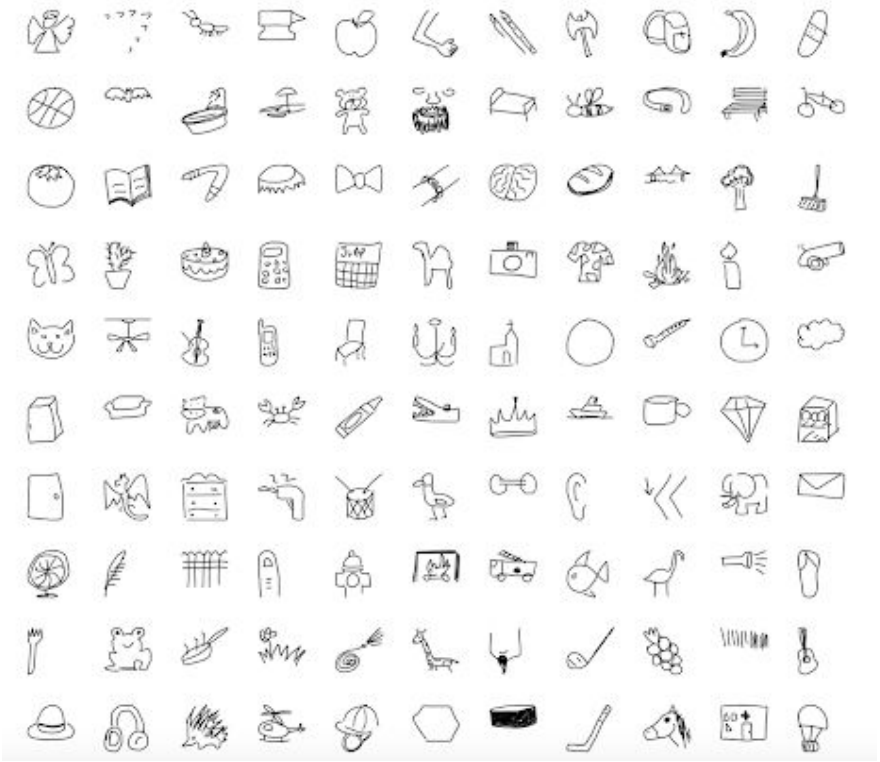
Object Sketch — Complementary text — Target Image

No suitable dataset available…
But can we adopt existing datasets?

The CSTBIR Dataset



Visual Genome (Krishna et al., 2017)

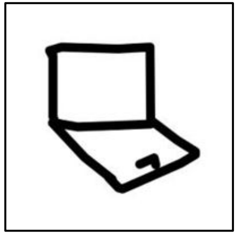


Google's Quick, Draw! (Ha et al., 2017)

The CSTBIR Dataset

Multimodal Query

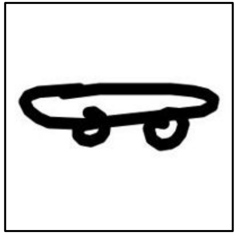
Target Image

A silver  is on the desk

(Object: **laptop**)

The CSTBIR Dataset

Multimodal Query

A man performing a  trick

(Object: **Skateboard**)

Target Image

The CSTBIR Dataset

Multimodal Query

Target Image

A picture hanging above the 

(Object: **Fireplace**)

The CSTBIR Dataset in Numbers

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 % Relevant Area in Target Image	36.7

The CSTBIR vs Other Datasets

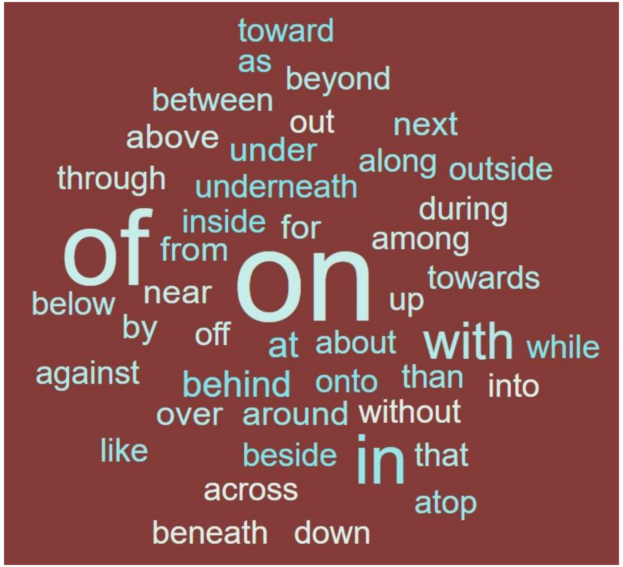
Query	Dataset	# Instances	Sketch	Text	Target Image
Sketch	TU-Berlin	20K	Object	None	Focused Object
Sketch	QMUL-Shoe-V2	6.7K	Object	None	Focused Object
Text	MS COCO	567K	None	Complete	Complete Scene
Text	Flickr-30K	158K	None	Complete	Complete Scene
Sketch+Text	FS COCO	10K	Scene	Complete	Complete Scene
Sketch+Text	CSTBIR (Ours)	2M	**Object**	**Complementary**	**Complete Scene**

The CSTBIR Dataset

Word Clouds for the text descriptions

adjectives
(*attributes*)

verbs
(*action words*)



prepositions
(*position indicating words*)

What about Rare Objects?

Multimodal Query

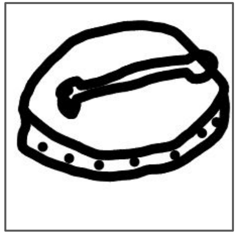
Target Image

Pair of  climbing cliffs on a sunny day.

(Object: **Markhor**)

What about Rare Objects?

Multimodal Query

People admiring a  displayed on a table.

(Object: **Bodhran**)

Target Image

What about Rare Objects?

Multimodal Query

Person dressed in a suit standing beside a  .

(Object: **Penny Farthing**)

Target Image

What about Rare Objects?

Multimodal Query

Students observing an  in a Desert Classroom.

Target Image

(Object: **echidna**)

What about Rare Objects?

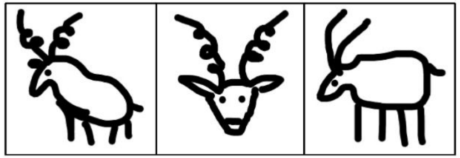
Multimodal Query

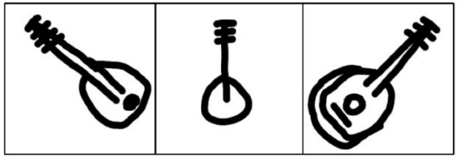
Police officers riding their  across a busy street

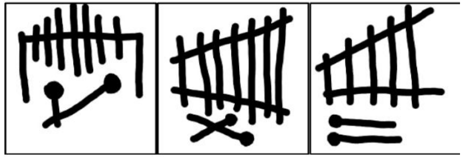
(Object: **segway**)

Target Image

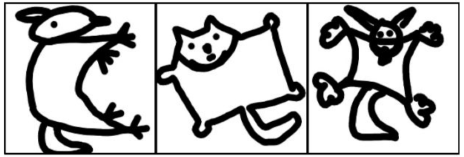
What about Rare Objects?

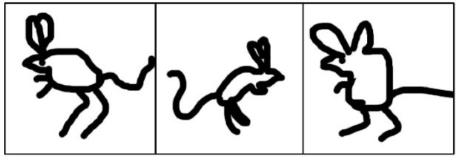

markhor

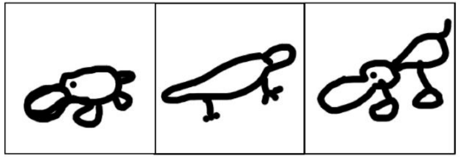

bouzouki

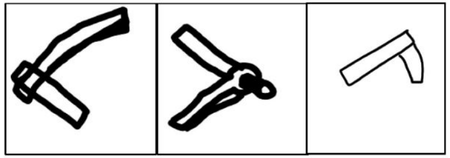

marimba

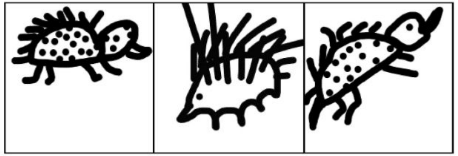

flame lily

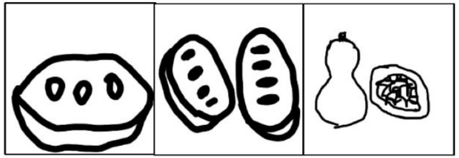

sugarglider


jerboa

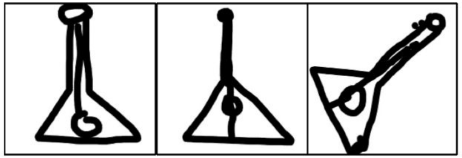

platypus


froe

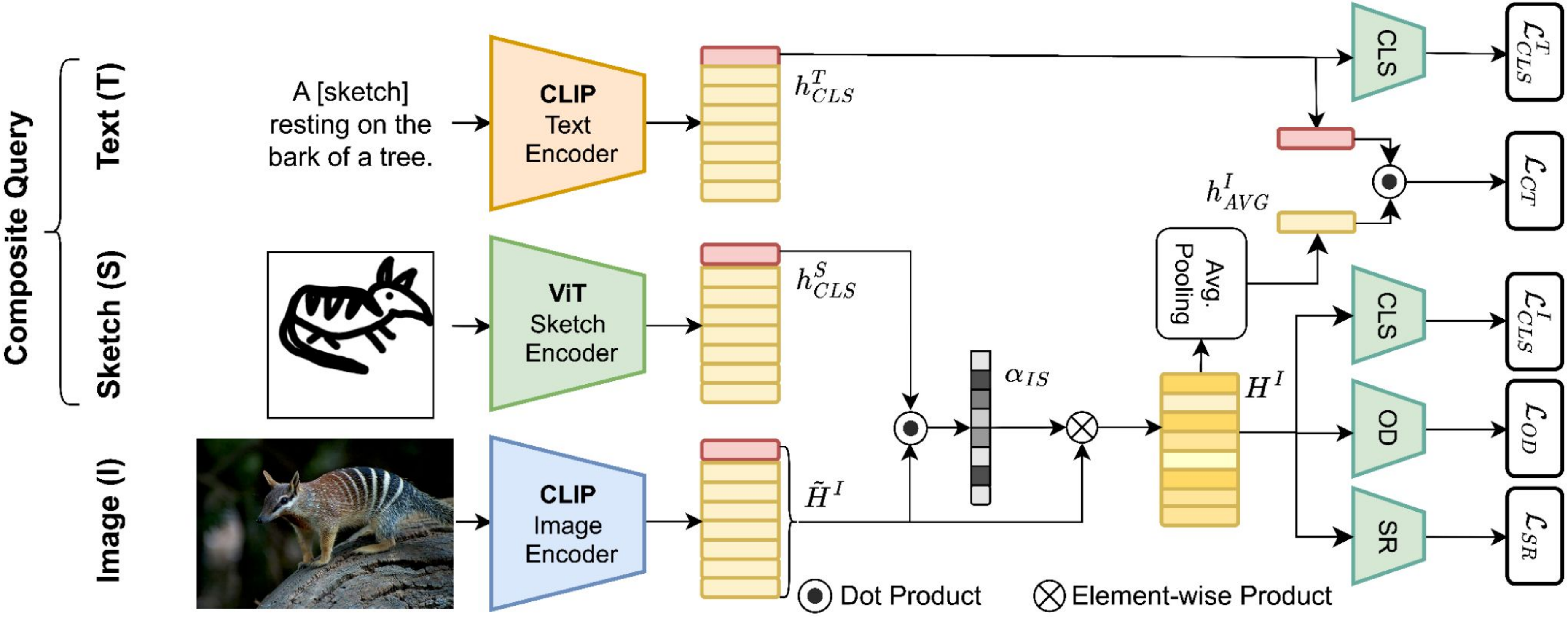

echidna


pawpaw


skycar

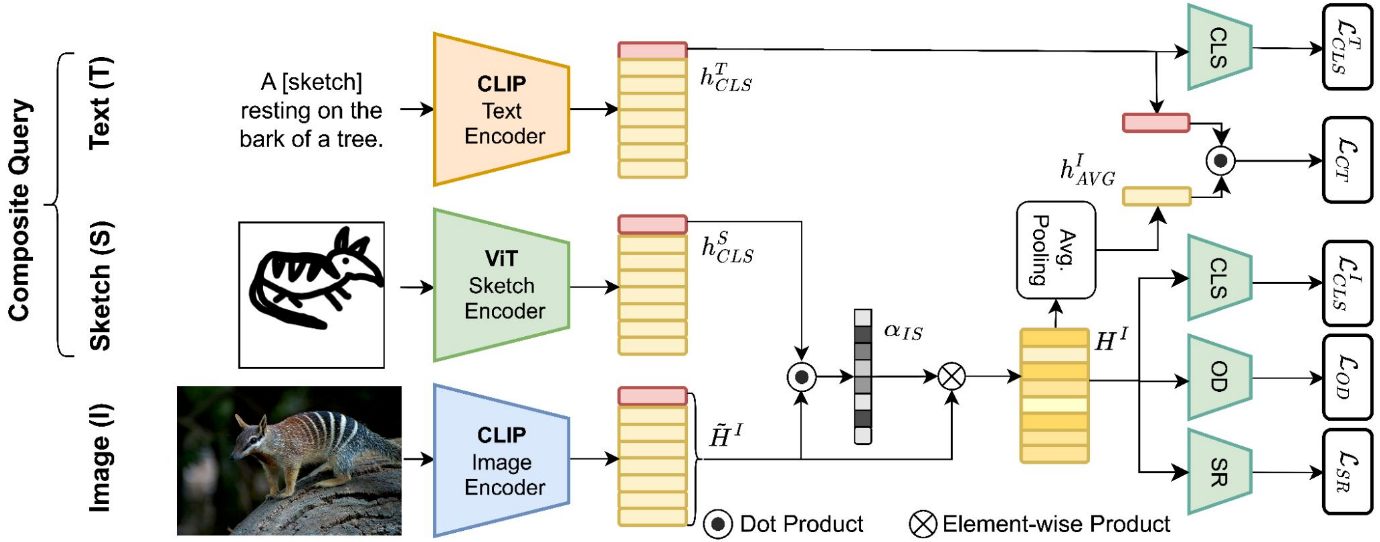

balalaika

STNet: Sketch+Text based Image Retrieval Network



[Gatti et al., 2023 Under Review]

Training Objectives for STNet



We follow the InfoNCE objective of CLIP.

Additionally, we introduce **three new task-specific objectives**:

1. Object Classification (*CLS*)
2. Sketch Object Detection (*OD*)
3. Sketch Reconstruction (*SR*)

Baselines

Based on Modality:

1. **Text-only**
   a. VisualBERT (Li et al., 2019)
   b. ViLT (Kim et al., 2021)
   c. CLIP (Radford et al., 2021)

2. **Sketch-only**
   a. Doodle2Search
   b. DeepSBIR
   c. ViT-Siamese (*our vision transformer-based baseline*)

Baselines

3. **Multimodal (Sketch + Text) baselines**
   a. TIRG (Vo et al., CVPR 2019)

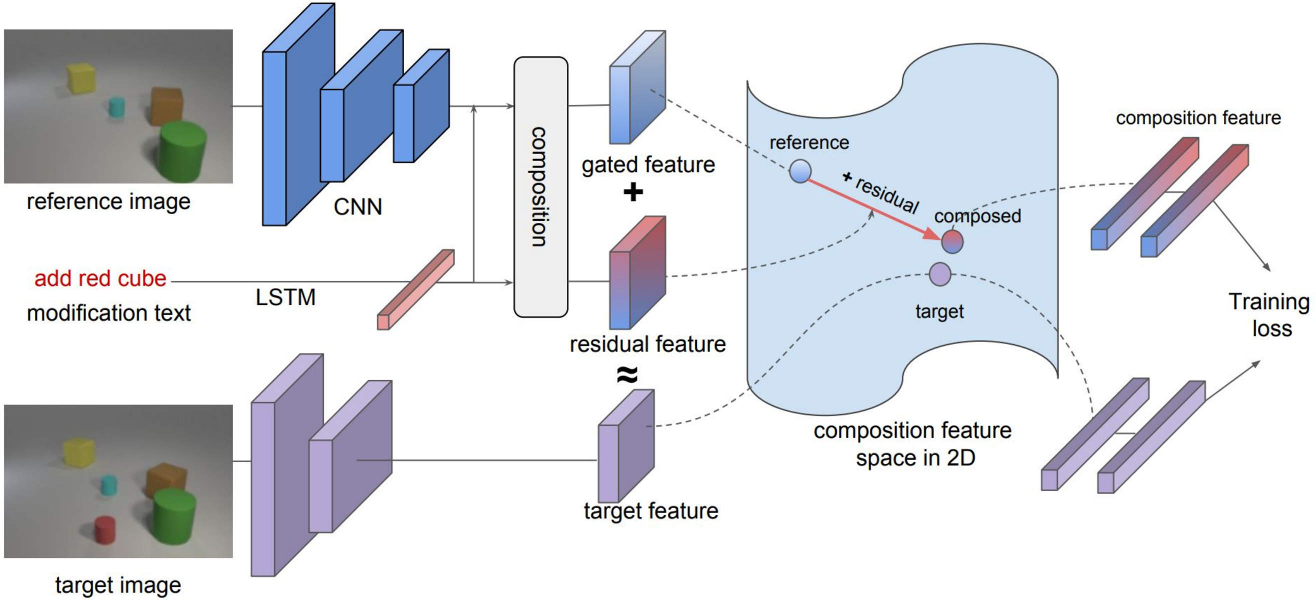


Figure: [Vo et al. CVPR 2019]

Baselines

4. Multimodal (Sketch + Text) baselines

b. Taskformer (Sangkloy et al., ECCV'22)

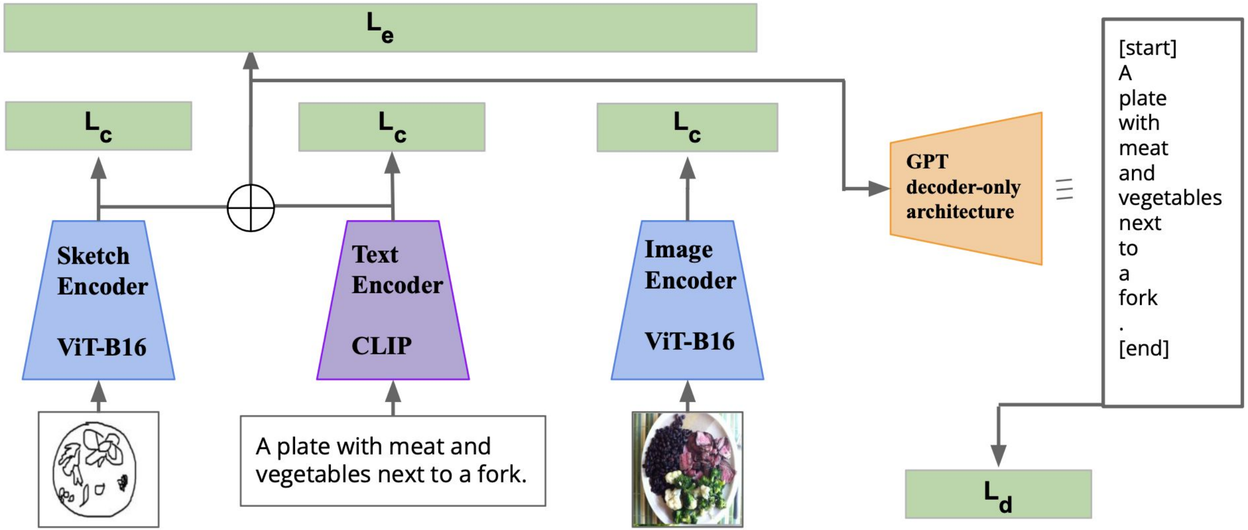
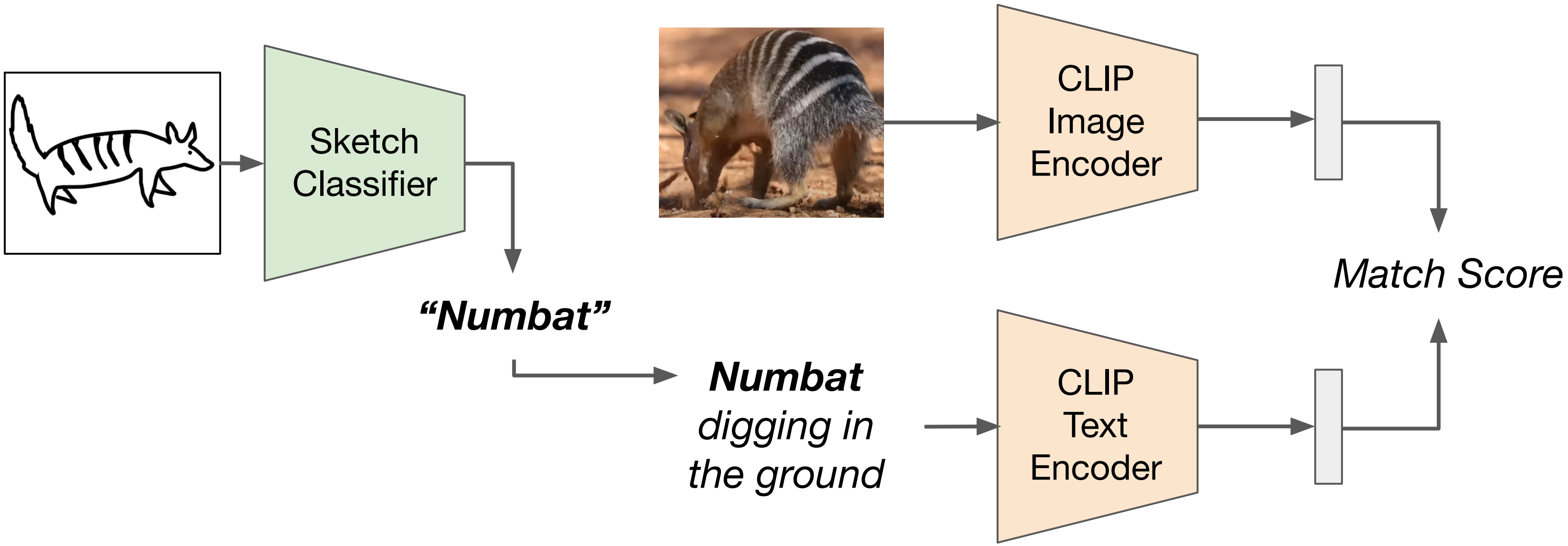


Figure: [Sangkloy et al., ECCV'22]

Baselines

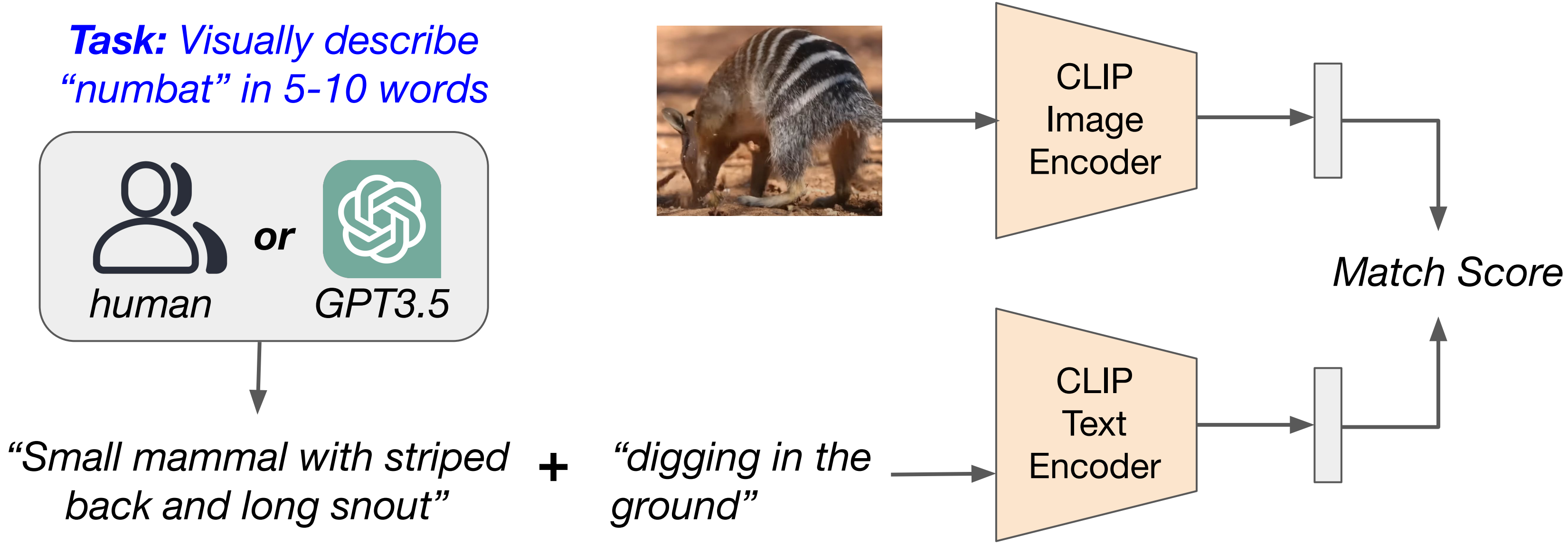
5. **Multimodal (Sketch + Text) baselines**
   c. Two-step Model (Categorize-then-Retrieve)



Baselines

6. Multimodal (Sketch + Text) baselines

d. Two-step (Description-based baseline)



Results

Comparison with baselines on the CSTBIR Test-1K set.

Input Modality	Method	Test-1K				
		R@10 ↑	R@20 ↑	R@50 ↑	R@100 ↑	MdR ↓
Sketch	Doodle2Search [10]	14.3	24.5	36.2	45.7	129.0
	DeepSBIR [63]	5.2	8.8	18.9	27.4	258.5
	ViT-based Siamese Network	20.4	34.2	51.0	62.6	48.0
Text	VisualBERT [32]	23.3	35.9	40.8	54.0	46.0
	ViLT [24]	28.1	42.7	60.2	74.3	30.0
	CLIP [47]	50.6	63.1	78.8	86.7	10.0
Sketch+Text	TIRG [61]	31.9	44.2	62.8	73.2	27.5
	Taskformer [53]	22.4	35.6	42.3	53.8	48.0
	Two-stage Model	67.0	77.4	88.6	**93.7**	5.0
	Two-stage Model (desc)	60.1	73.7	85.5	91.6	7.0
	STNet (Ours)	**73.7**	**80.6**	**89.4**	93.5	**3.0**

STNet has better overall performance on CSTBIR

Ablation Study

Modality and loss ablation on CSTBIR Test-1K split.

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

- Without either sketch or text inputs, STNet performance drops.
- Object Classification loss is the most effective additional loss.

Selected Visual Results

Search Query

 on a slide being fed red ice cream

Top-5 Retrieved Images



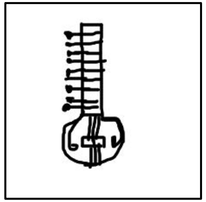


(Object: **capybara**)

Selected Visual Results

Search Query

Bearded man on the bank of a river playing besides a man playing tabla.



Top-5 Retrieved Images

(Object: **sitar**)

Selected Visual Results

Search Query

Person dressed in a suit standing beside a  .

Top-5 Retrieved Images

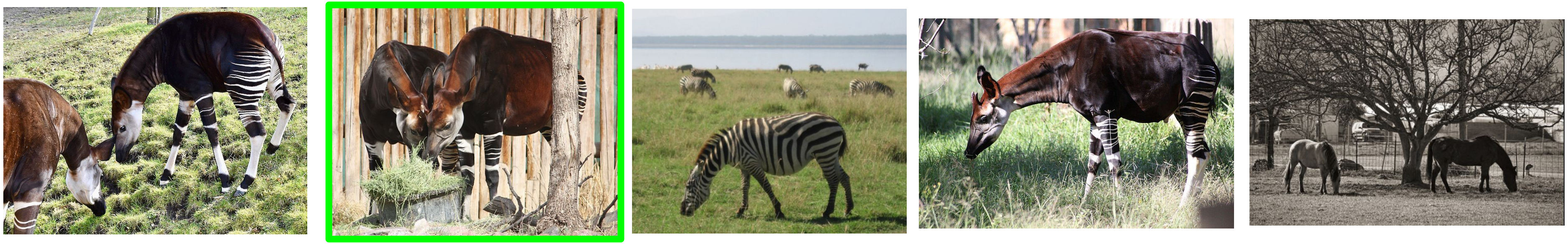
(Object: **penny farthing**)

Selected Visual Results

Search Query

Pair of  feeding on green grass.

Top-5 Retrieved Images



(Object: **okapi**)

Where do the errors come from?

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

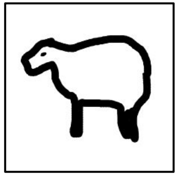
Method	Missing labels	Misrecognized sketch category	Object Ambiguity
Two-stage	22	12	2
STNET (Ours)	31	9	0

- **Missing labels**: query matches multiple images; missing annotation in VG
- **Misrecognized sketches**: sketches may be misrecognized

Where do the errors come from?

Search Queries

Top-3 Retrieved Results

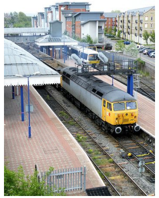
A  (sketch: capybara) resting in green grass

Error Type: Misrecognized sketch

Platform next to  yard

Error Type: Missing labels

Summary so far …

- **Deeper exploration into** composite modality/ sketch+text based image retrieval
- **Object localization based transformer framework**
- Moving towards **retrieval for open world category images**
- **Future directions:** extension of sketch+text retrieval and localization to videos

Work under review, stay tuned for code and datasets.

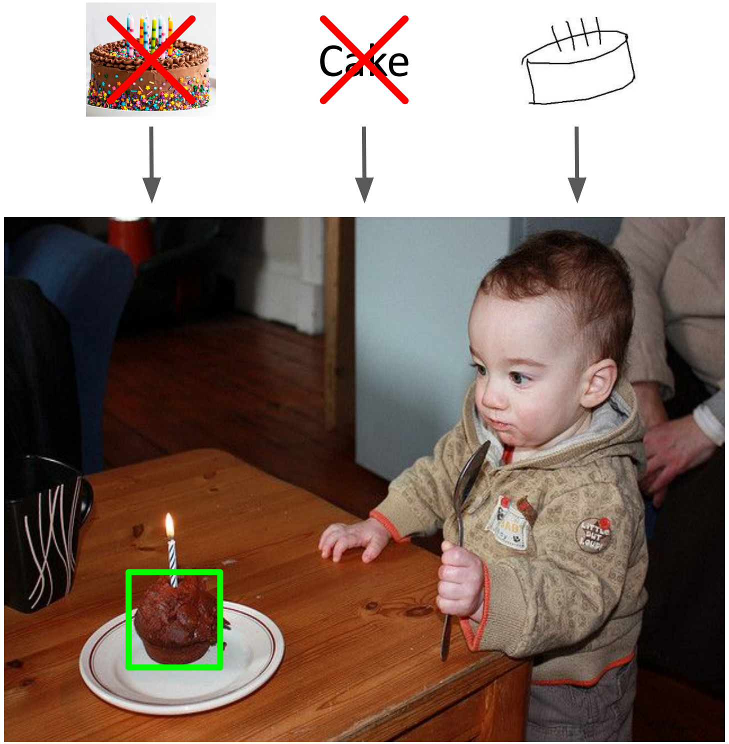
Sketch-Guided Object Localization in Natural Images

Aditay Tripathi[1], Rajath R Dani[1], Anand Mishra[2] and Anirban Chakraborty[1]

[1]Indian Institute of Science, Bengaluru

[2]Indian Institute of Technology, Jodhpur

Query-Guided Object Localization



Using natural image as query

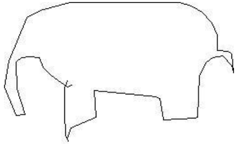
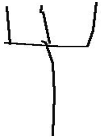
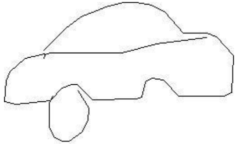
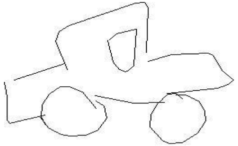
[Hsieh et al., NeurIPS 2019]

Using text as query

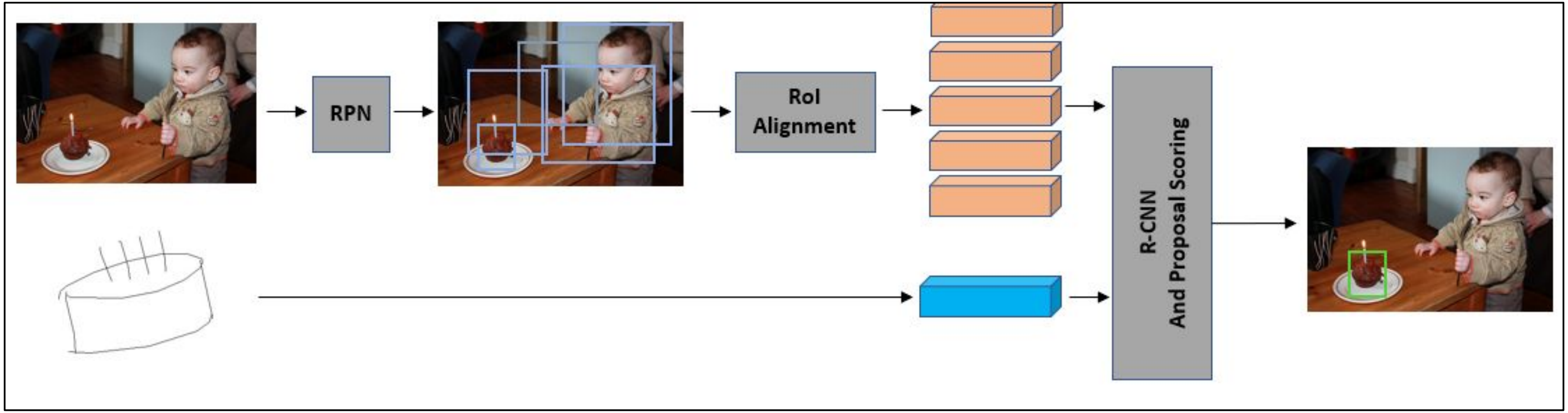
[Wang et al., TPAMI 2017]

Sketch-guided object localization (this work)

Challenges

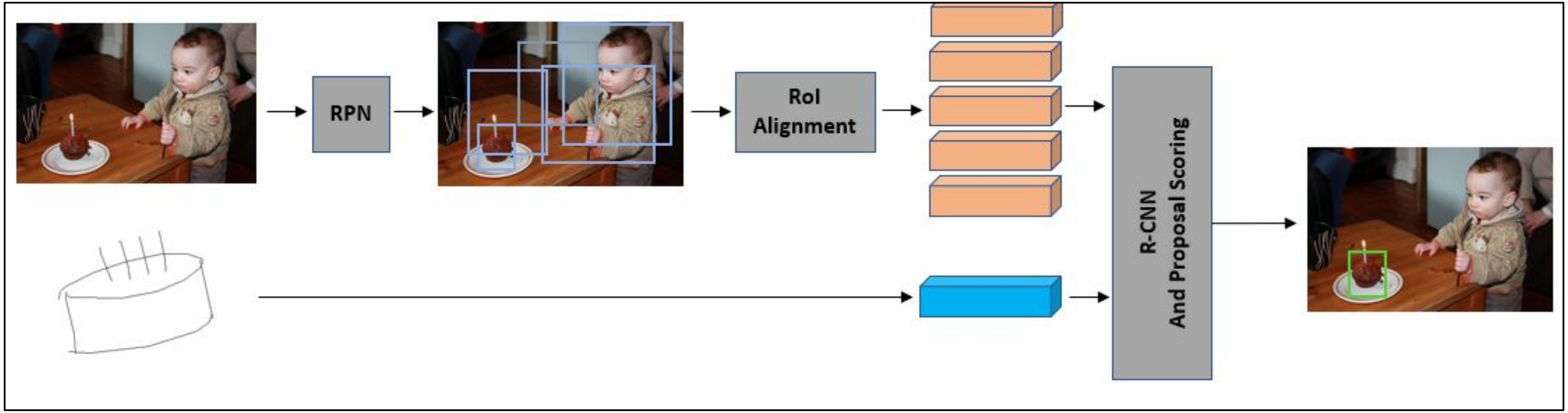
Domain Gap		
Unseen object		
Significant Variability		

Plausible Solution: Modified Faster R-CNN ?



- Modified to allow Query-guided object localization
- Score RoI features with sketch features

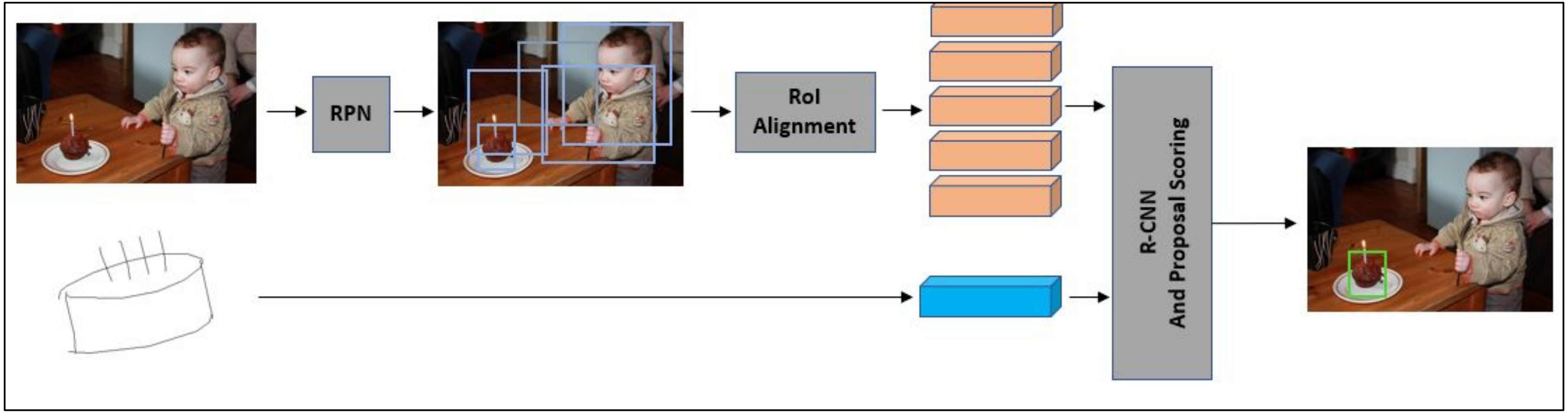
Plausible Solution: Modified Faster R-CNN ?



- Modified to allow Query-guided object localization
- Score RoI features with sketch features

Is the problem solved?

Plausible Solution: Modified Faster R-CNN ?



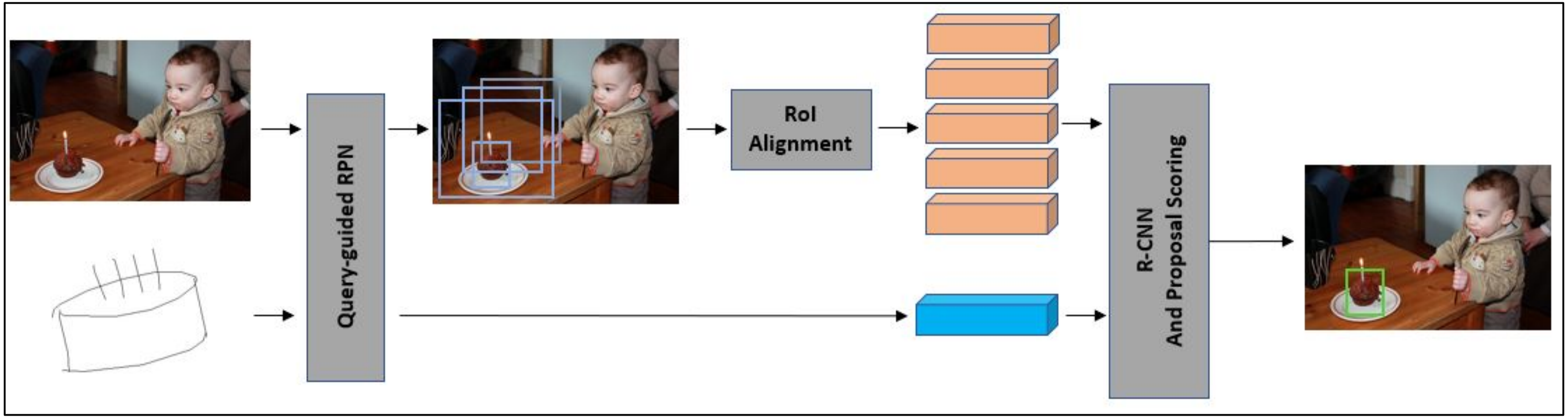
- Modified to allow Query-guided object localization
- Score RoI features with sketch features

Is the problem solved?: NO

Vanilla RPN may not generate proposals relevant to the object of interest

Proposed Model

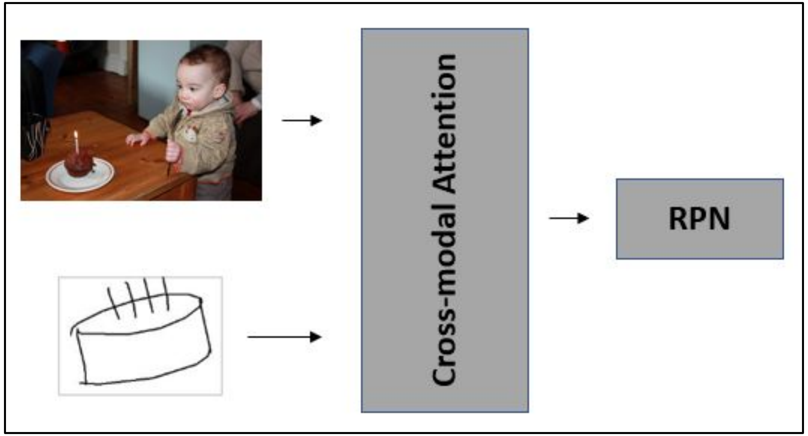
- **Query-guided RPN**
- Proposed ***Cross-modal attention*** to incorporate query information in RPN



Query-guided RPN

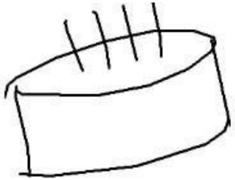
- Cross-Modal Attention learns a spatial compatibility between global sketch features and local image features

- The attended features are passed through RPN



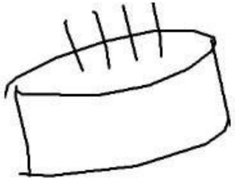
Cross-Modal Attention

- Find locations in image that are similar to the sketch query



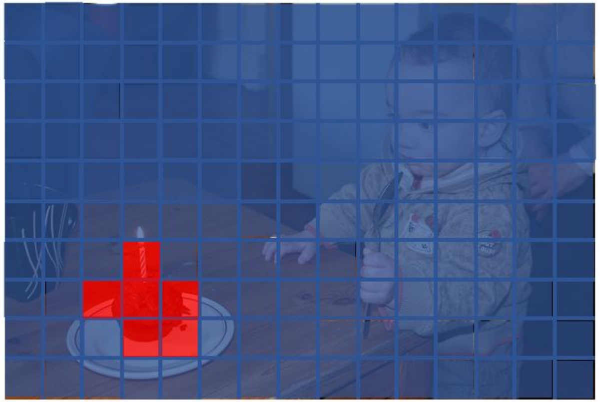
Cross-Modal Attention

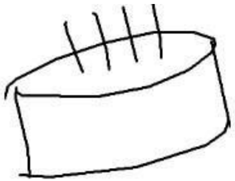
- Assigns **high score** to the locations **compatible** to the sketch



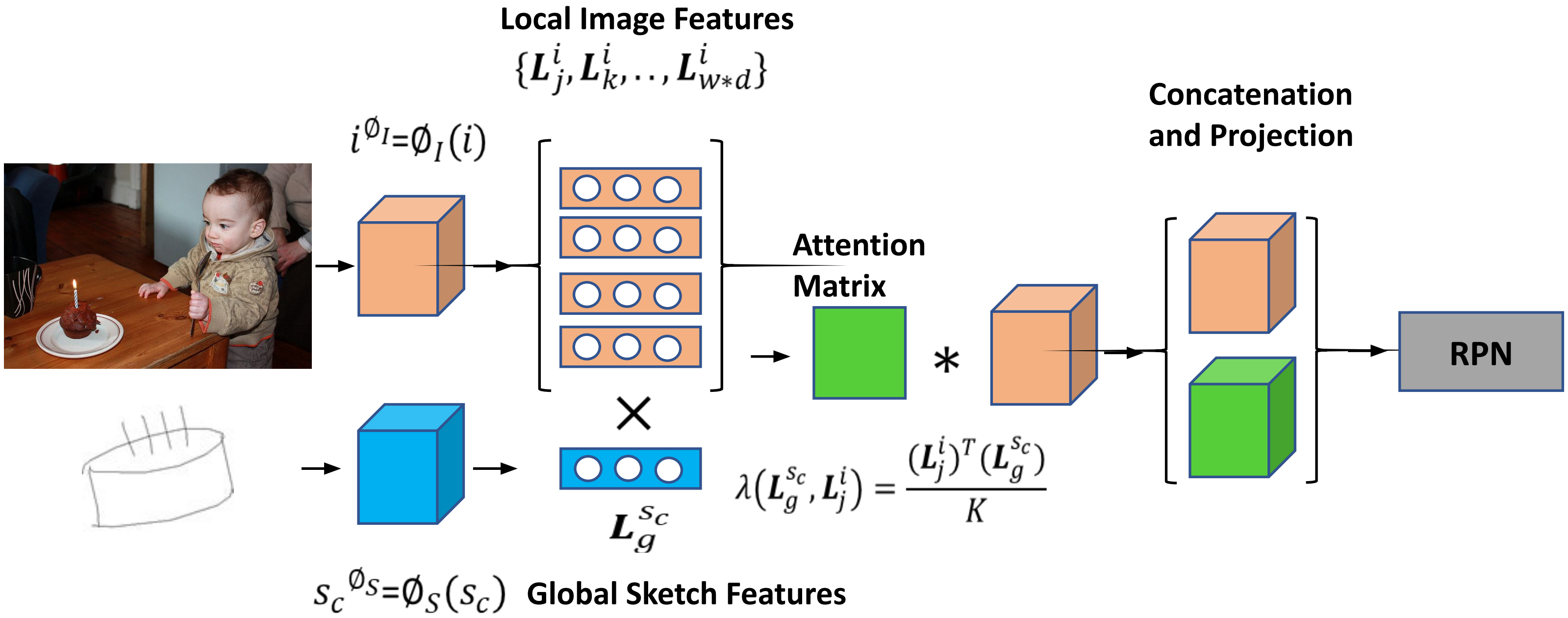
Cross-Modal Attention

- Assigns **high score** to the locations **compatible** to the sketch
- Low score to **incompatible** locations





Cross-Modal Attention - in detail



*http://visual-computing.in/sketch-guided-object-localization/

Proposal Scoring

- **Subsets** of **proposals** are pooled from proposals generated by query-guided RPN.
- Proposals are labelled **1(or 0)** based on **IoU** of the proposal with GT bounding box.
- A **margin-ranking** loss between the pooled proposals and sketch query is minimized.

Proposal Scoring

- Let Ɵ be the **scoring function** that scores the sketch query to an object proposal.

$$a_k = \theta(g_m(p_k); g'_m(s))$$

- The margin ranking loss is given as follows:

$$L(\mathbf{R}, s) = \sum_{k} \{y_k max(m^+ - a_k, 0) + (1 - y_k)\max(a_k - m^-, 0) + L_{MR}^k\}$$

$$L_{MR}^k = \sum_{l=k+1} \{1_{[y_l = y_k]} \max(|a_k - a_l| - m^-, 0) + 1_{[y_l \neq y_k]} \max(m^+ - |a_k - a_l|, 0)\}$$

Proposal Scoring

- Let Θ be the **scoring function** that scores the sketch query to an object proposal.

$$a_k = \theta(g_m(p_k); g'_m(s))$$

- The margin ranking loss is given as follows:

$$L(\mathbf{R}, s) = \boxed{\sum_k \{y_k max(m^+ - a_k, 0) + (1 - y_k)\max(a_k - m^-, 0)} + L_{MR}^k\}$$

$$L_{MR}^k = \sum_{l=k+1} \{1_{[y_l = y_k]} \max(|a_k - a_l| - m^-, 0) + 1_{[y_l \neq y_k]} \max(m^+ - |a_k - a_l|, 0)\}$$

Proposal Scoring

- Let ϴ be the **scoring function** that scores the sketch query to an object proposal.

$$a_k = \theta(g_m(p_k); g'_m(s))$$

- The margin ranking loss is given as follows:

$$L(\mathbf{R}, s) = \sum_k \{y_k max(m^+ - a_k, 0) + (1 - y_k)\max(a_k - m^-, 0) + \boxed{L_{MR}^k}\}$$

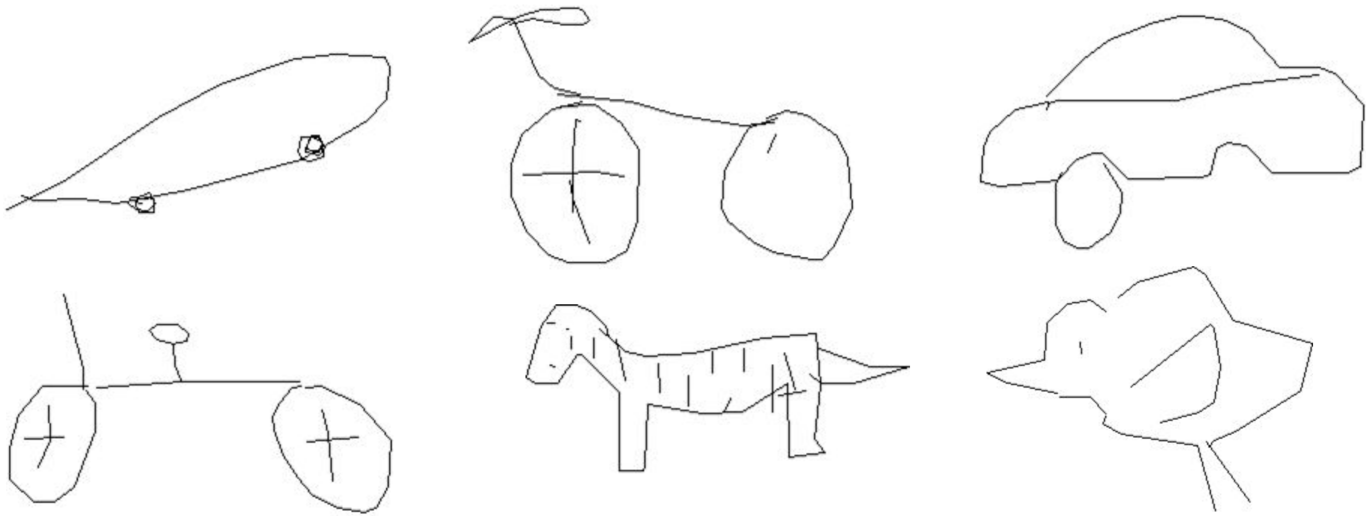
$$\boxed{L_{MR}^k = \sum_{l=k+1} \{1_{[y_l = y_k]} \max(|a_k - a_l| - m^-, 0) + 1_{[y_l \neq y_k]} \max(m^+ - |a_k - a_l|, 0)\}}$$

Proposal Scoring

- Along with the proposal scoring additional losses are used in training.
- Cross-entropy loss on the labeled (**background** or **foreground**) feature vectors of the region proposals
- Regression loss on the predicted bounding box locations with respect to the ground truth bounding box.

Dataset

- Sketches from QuickDraw dataset are used in our experiments.
- Consists of 50 M drawing across 345 categories.
- Selected a subset of 800k sketches for our experiments.



Dataset

- Images are chosen from MS-COCO*[Lin et al. ECCV2014]* dataset and Pascal-VOC*[Everingham et al. ICCV 2010]* datasets.
- MS-COCO: 330K images across 80 categories.
- Pascal-VOC: 9,963 images across 20 categories.
- Commonly used datasets in Object detection research.
- MS-COCO has 56 categories common with QuickDraw.
- Pascal-VOC has 9 categories common with QuickDraw.

One-shot Common Train-Test Categories

- Both "**seen**" and "**unseen**" categories used in training.
- **Single sketch** as a query.

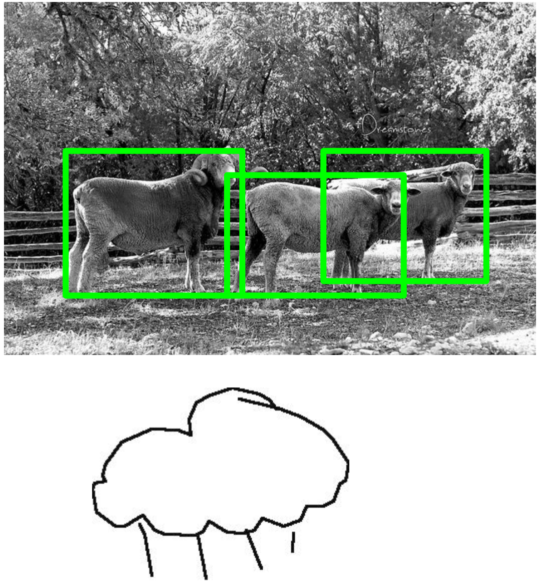
Model	COCO val2017		VOC test2007
	mAP	%AP@50	mAP
Modified Faster-RCNN	0.18	31.5	**0.65**
Matchnet *[Hsieh et al., NeurIPS 2019]*	0.28	48.5	0.61
Cross-Modal Attention	**0.3**	**50**	**0.65**

One-shot Disjoint Train-Test Categories

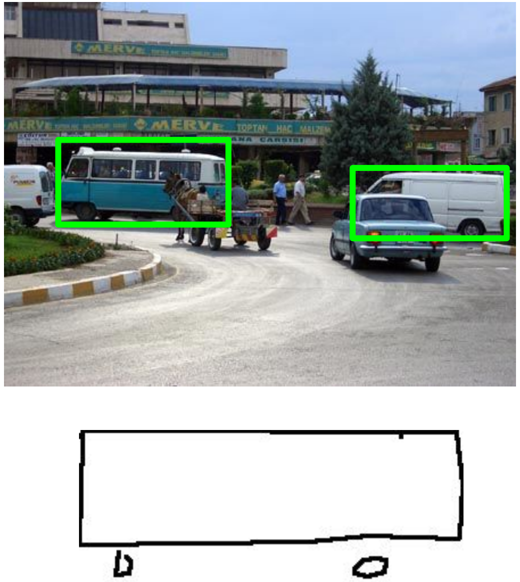
- “**Unseen classes**” not used during training.
- **Single** sketch as a query.

Model	%AP@50	
	unseen classes	seen classes
Modified Faster-RCNN	7.4	34.5
Matchnet *[Hsieh et al., NeurIPS 2019]*	12.4	**49.1**
Cross-Modal Attention	**15**	48.8

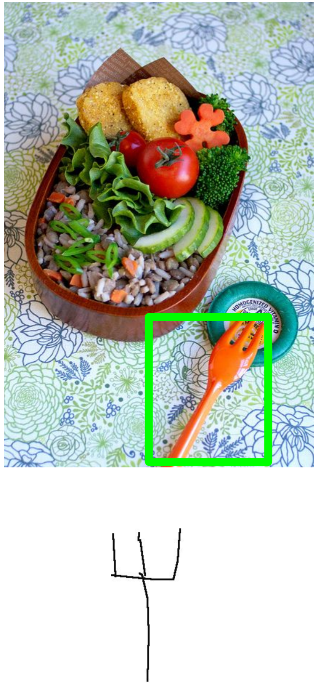
Selected Results



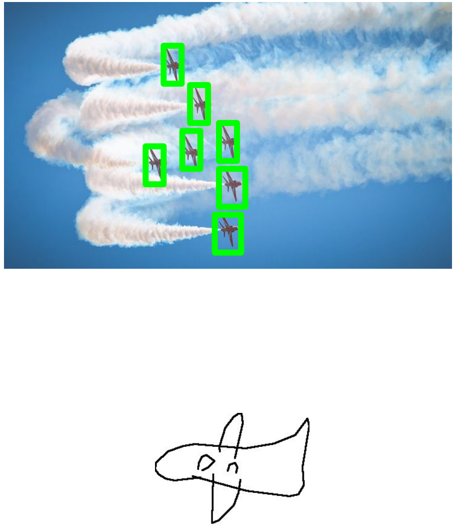
Multiple instance



Occluded object

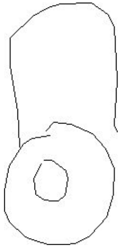


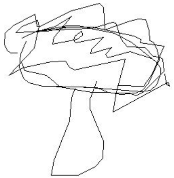
Unseen Object

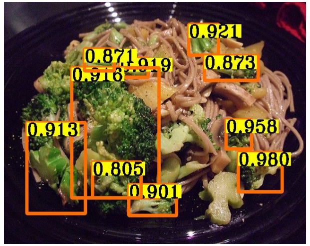


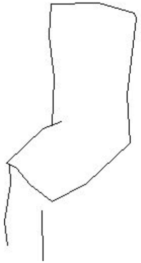
Small Object

More Results

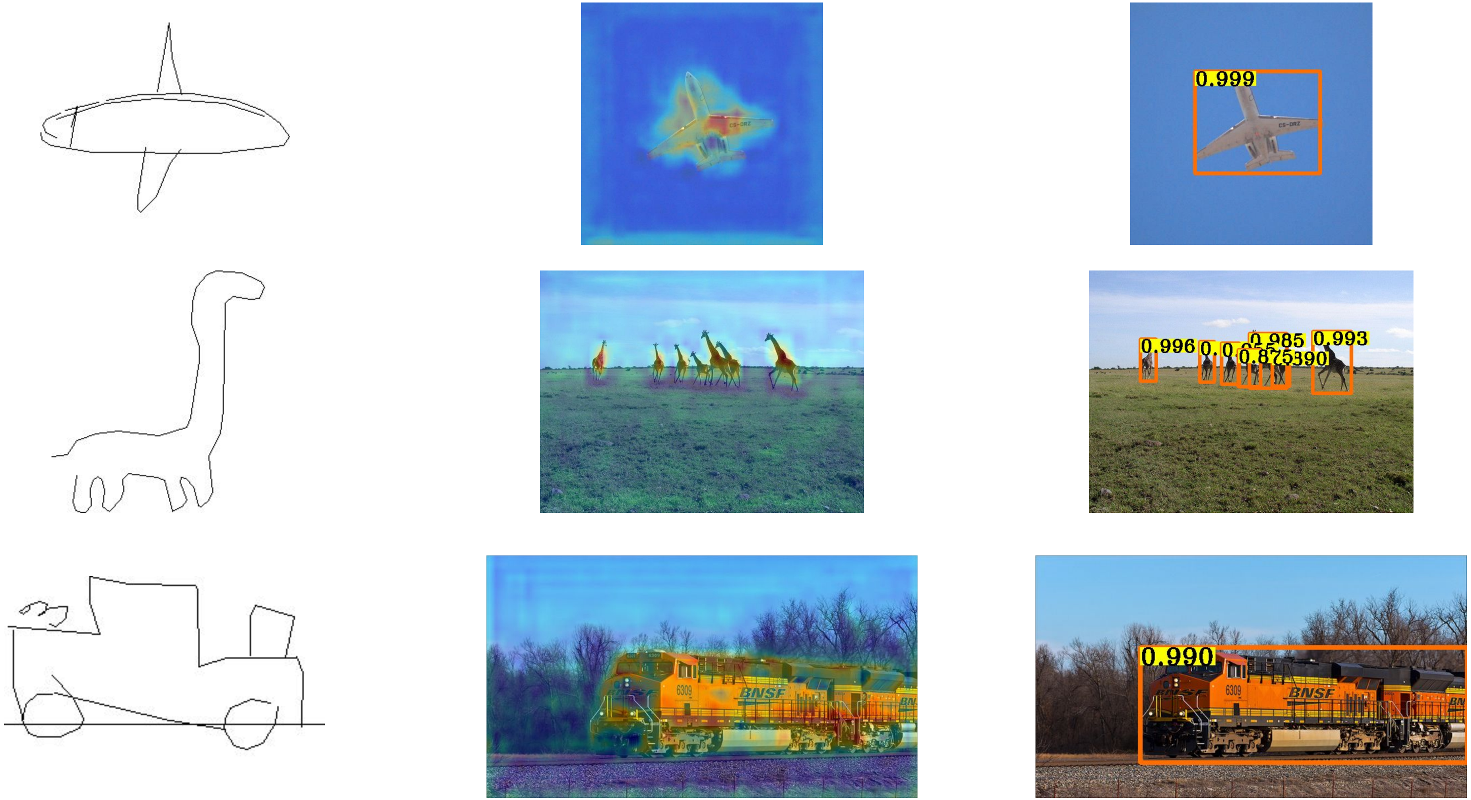






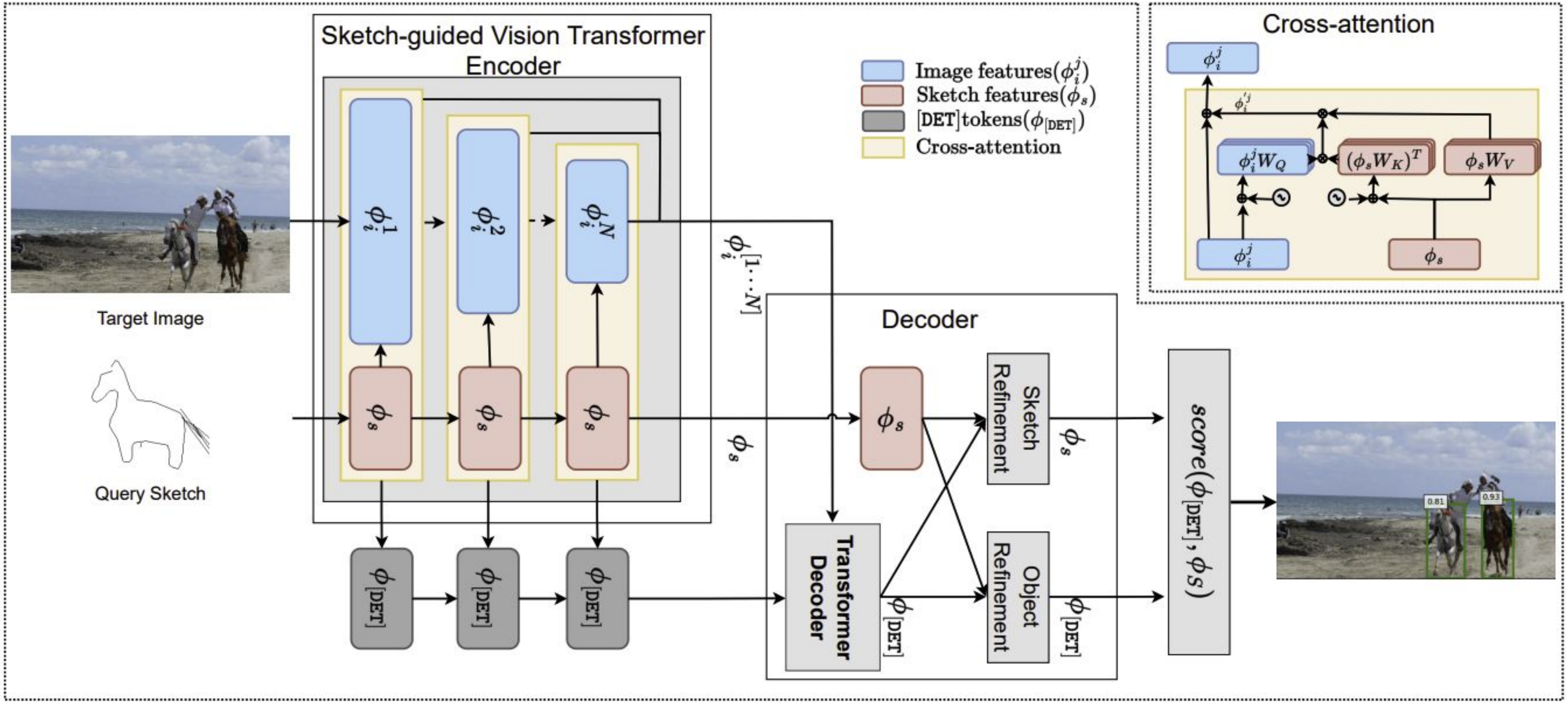


More Results



What about modern architectures for this task?

Sketch-guided Vision Transformer (under review)



Much better results, but far from being solved!

Models	mAP	AP@50	AP^L
Modified FasterRCNN	3.3	7.4	6.2
CoAT (Hsieh et al. 2019)	5.9	12.4	10.6
CMA (Tripathi et al. 2020)	7.5	15.0	12.4
Ours	**12.2**	**18.3**	**24.6**

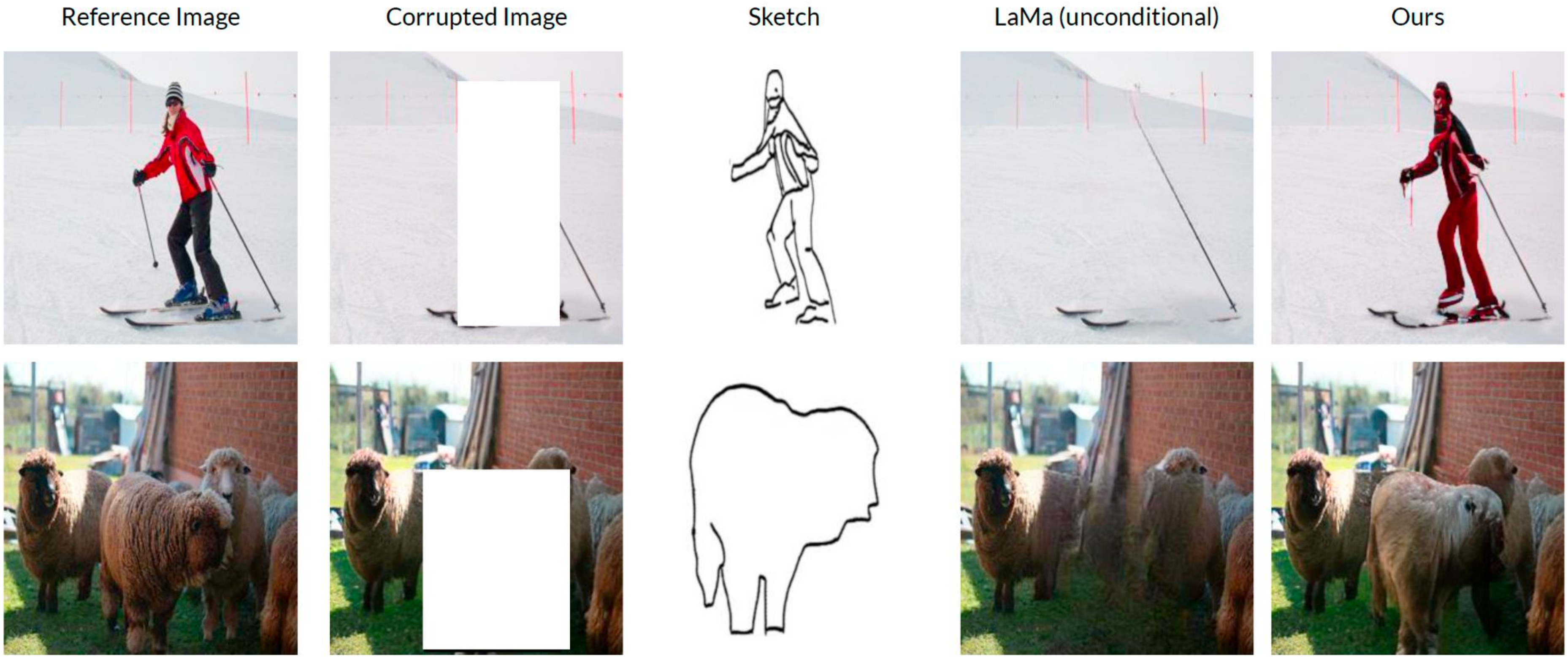
Summary so far …

- **Novel Task:** Sketch-Guided Object Localization
- **Query-Guided Region Proposal Network**
- **Cross-Modal Attention**
- A step towards **open-world object localization**
- **Future direction:** bridging sketch and language

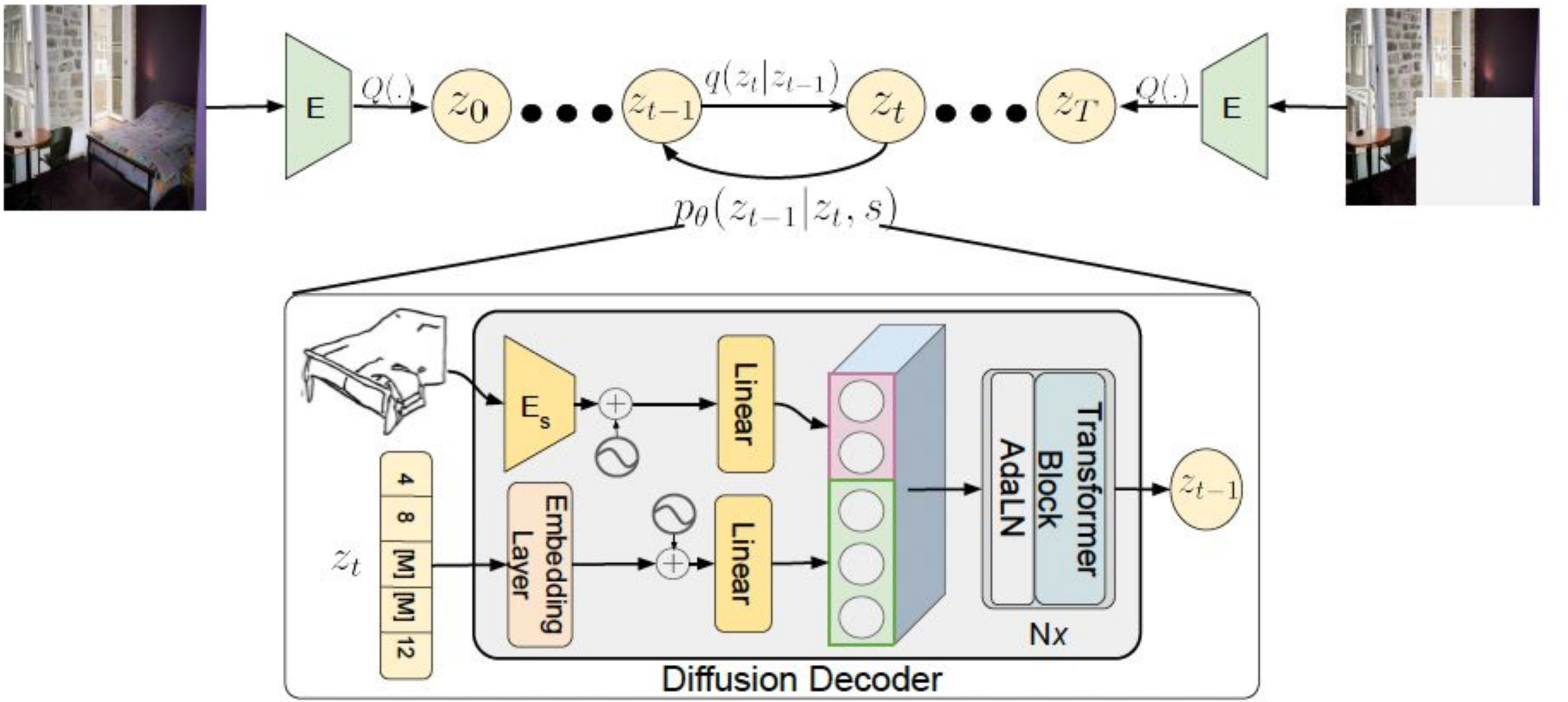
Code Available!

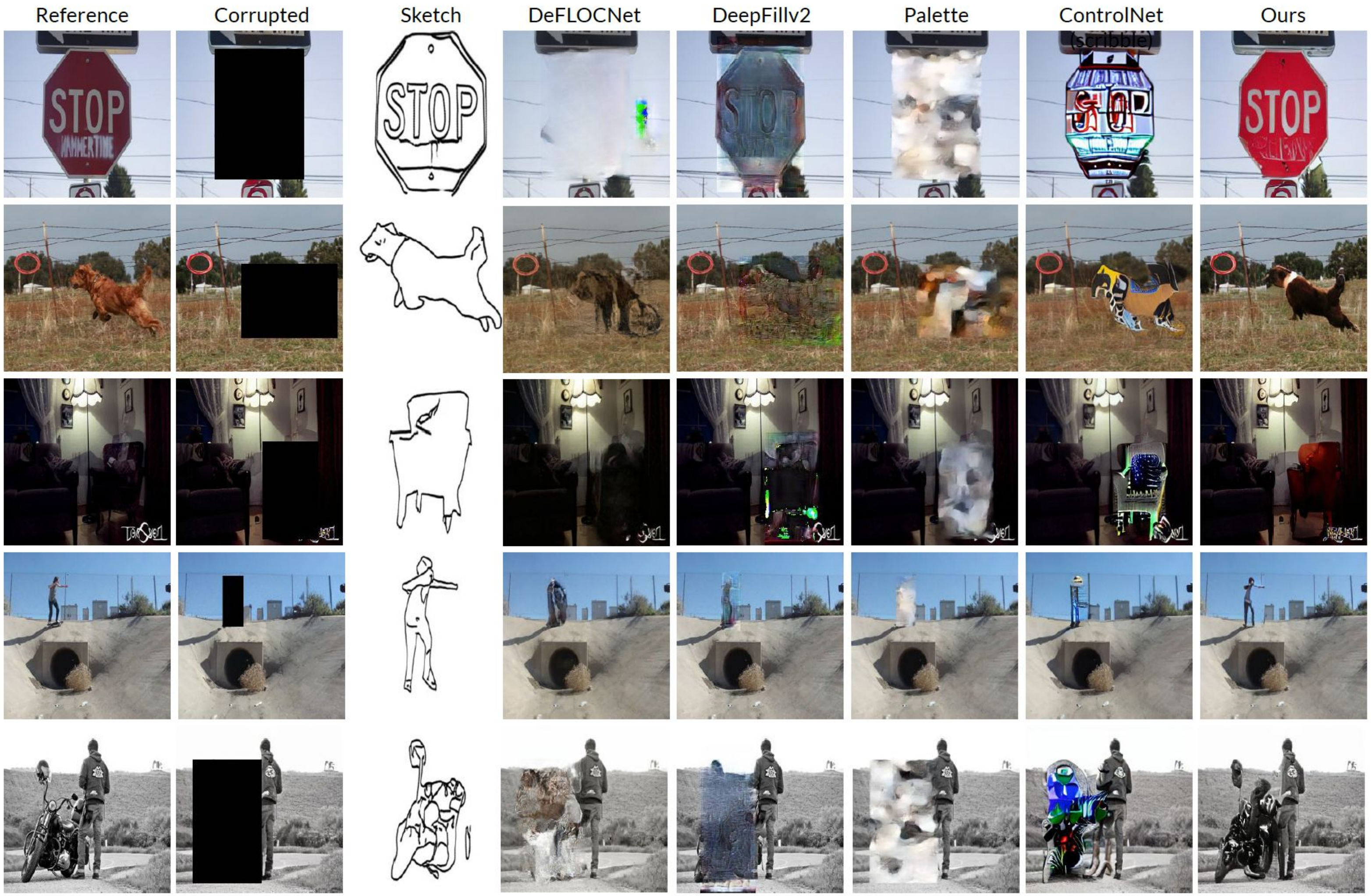
Sketch-Guided Image Inpainting

Sketch-Guided Image Inpainting



Partial Discrete Diffusion



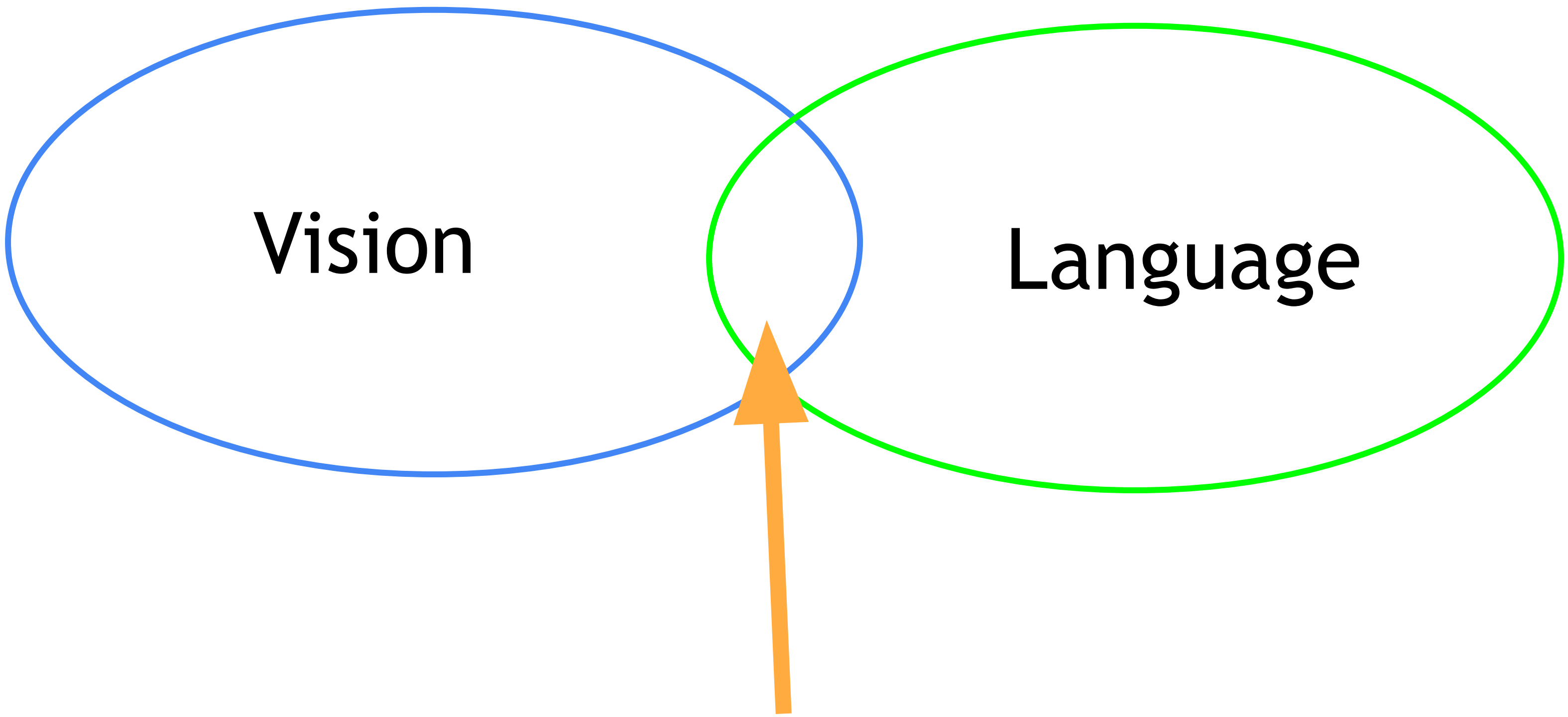

Reference
Corrupted
Sketch
DeFLOCNet
DeepFillv2
Palette
ControlNet (scribble)
Ours
STOP
HAMMERTIME
STOP
STOP

Open Areas

- Large-scale self-supervised models and foundation models using sketchified unlabelled images
- Open-set tasks (detection, segmentation, recognition)
- Creative Sketch Generation
- **Applications:** educational content search and creation

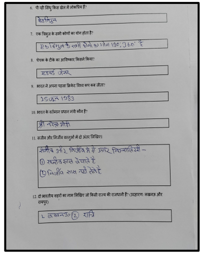
Our group@IITJ - VL2G

Our group@IITJ - VL2G



Our Focus: Vision and Language

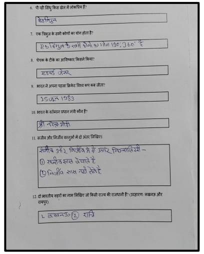
Language inside Images



Multilingual H/W, Scene Text, Visual Translation

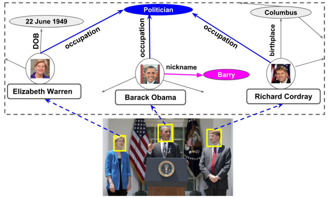
Our Focus: Vision and Language

Language inside Images



Multilingual H/W, Scene Text, Visual Translation

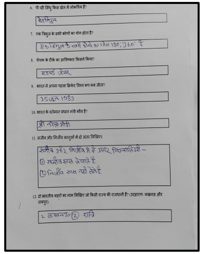
Language outside Images



Integrating vision with World Knowledge for Commonsense and factual reasoning

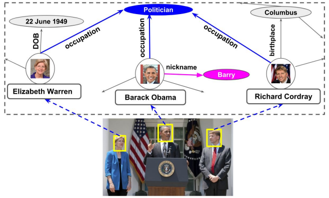
Our Focus: Vision and Language

Language inside Images



Multilingual H/W, Scene Text, Visual Translation

Language outside Images



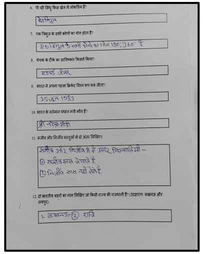
Integrating vision with World Knowledge for Commonsense and factual reasoning

Sketch as a Language

Sketch for localization, retrieval and inpainting

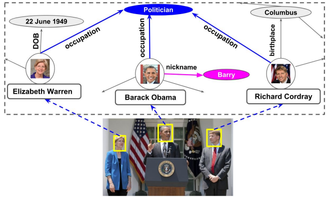
Our Focus: Vision and Language

Language inside Images



Multilingual H/W, Scene Text, Visual Translation

Language outside Images



Integrating vision with World Knowledge for Commonsense and factual reasoning

Sketch as a Language

Sketch for localization, retrieval and inpainting

Video and Language

Twist Shower Head Clockwise

Core Video Tasks, Localization, Grounding

Thank You

Questions/comments/Suggestions?

Full-time RA and PhD Positions available in the group! Consider applying.

Contact: mishra@iitj.ac.in

Group Website: https://vl2g.github.io/