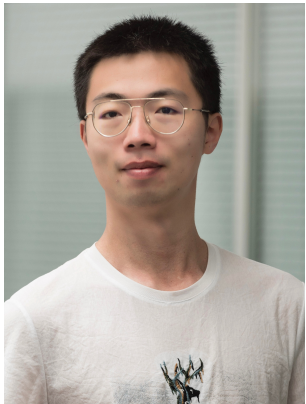


# SketchXAI: A First Look at Explainability of Human Sketches

CVPR 2023: THU-PM-260



Zhiyu Qu<sup>1,3</sup>



Yulia Gryaditskaya<sup>1</sup>



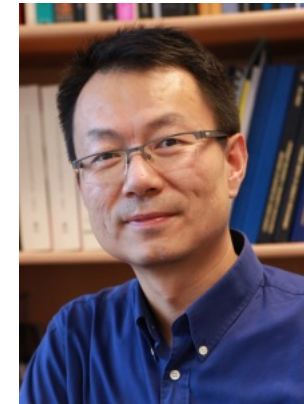
Ke Li<sup>1,2</sup>



Kaiyue Pang<sup>1</sup>



Tao Xiang<sup>1,3</sup>



Yi-Zhe Song<sup>1,3</sup>



<sup>1</sup> SketchX, CVSSP, University of Surrey

<sup>2</sup> Beijing University of Posts and Telecommunications

<sup>3</sup> iFlyTek-Surrey Joint Research Centre



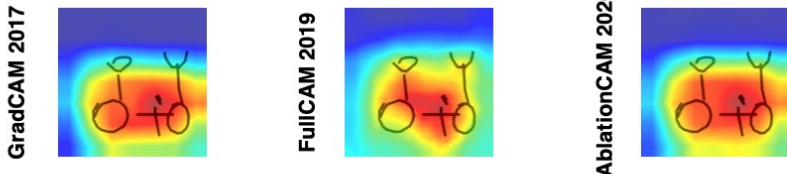
# 1 Minute Brief Introduction



Can I trust this prediction?  
Can AI explain itself?

## Progress so far...

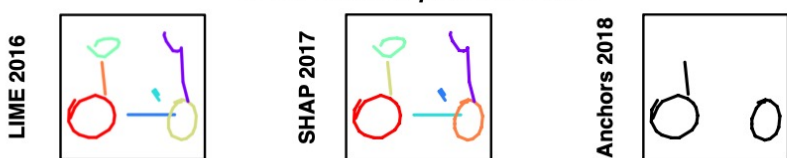
### Saliency map



### Self-attention



### Rule-based perturbation

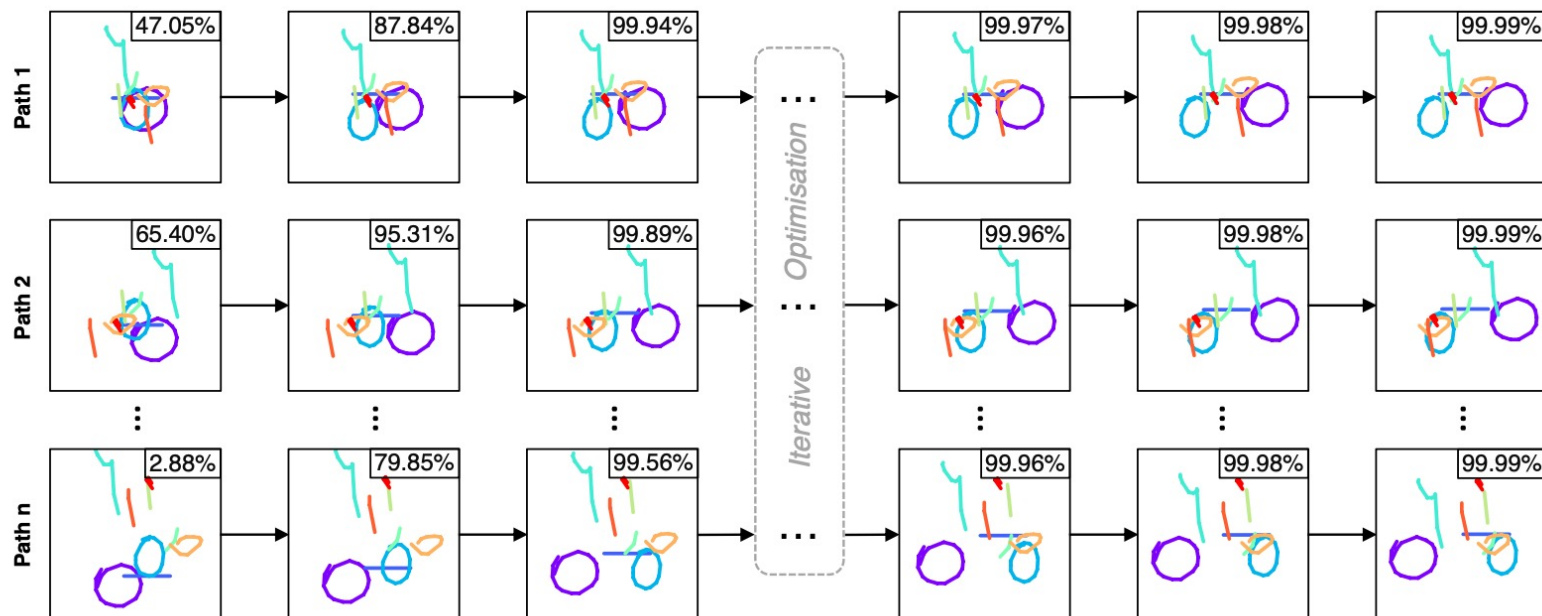


## Our proposal: Bicycle-Informed Stroke Inversion

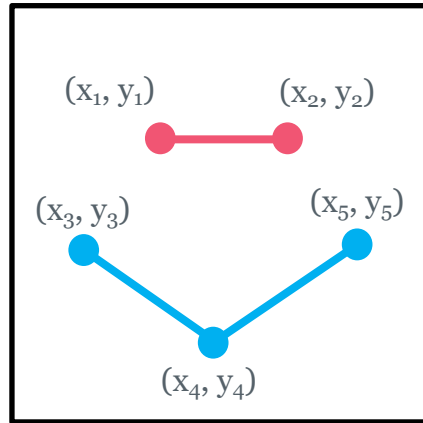
Step 1: Randomise stroke locations first



Step 2: Invert via iterative optimisation



Raster Image



CNN

Vector Image

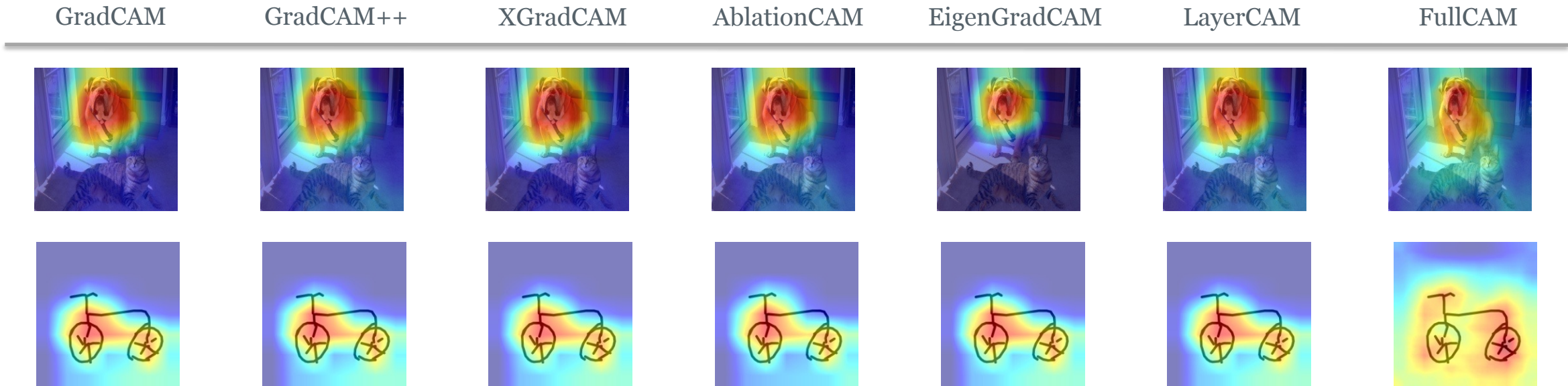
$[x_1, y_1, 0]$ ,  
 $[x_2, y_2, 1]$ ,  
 $[x_3, y_3, 0]$ ,  
 $[x_4, y_4, 0]$ ,  
 $[x_5, y_5, 1]$



RNN/Transformer/GNN



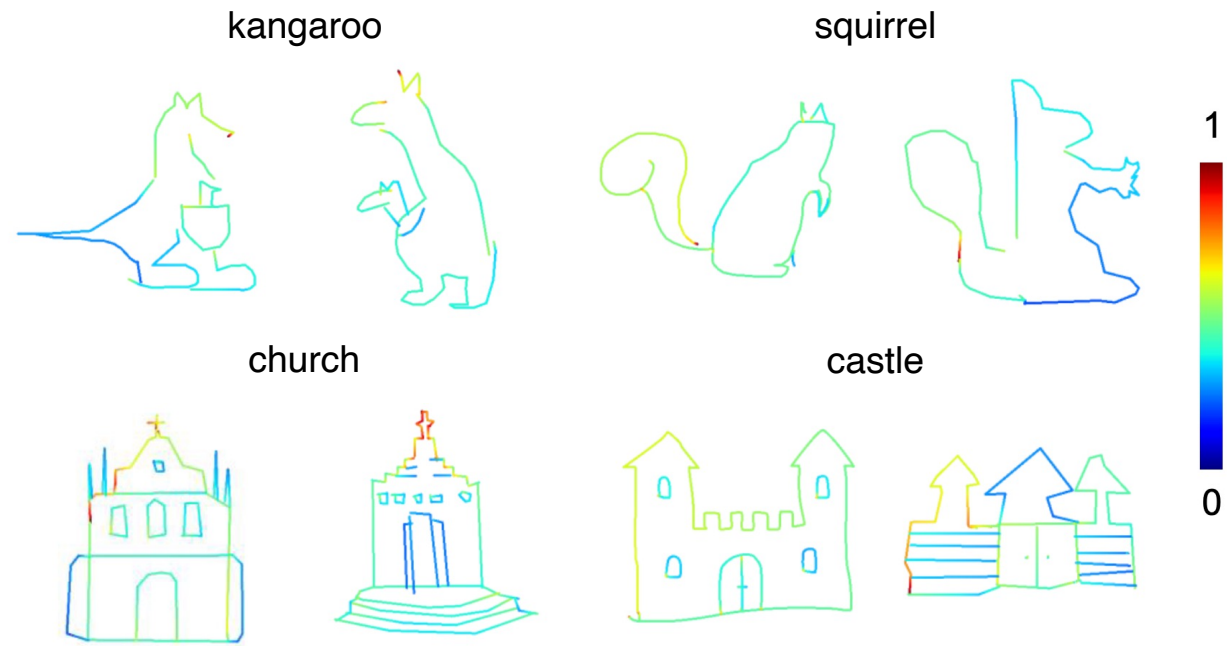
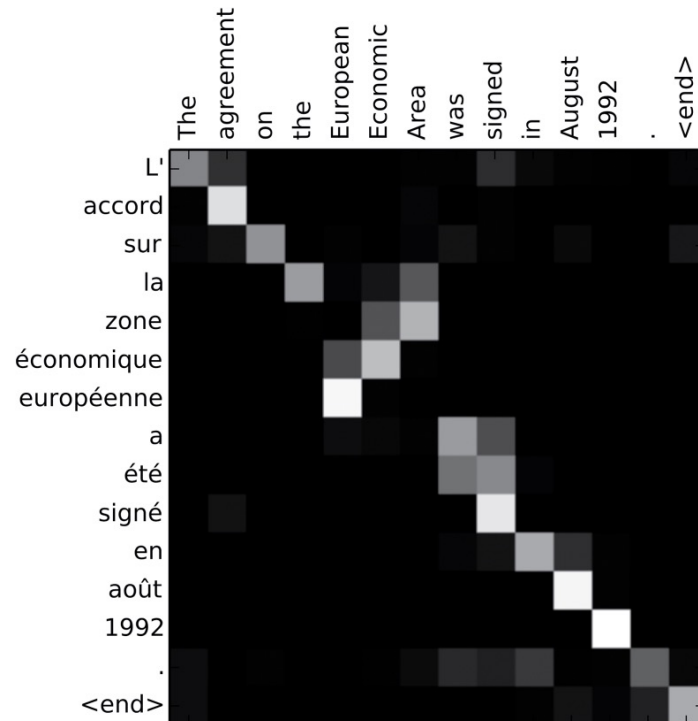
## Image-based Explainable Methods



- Information of images is redundant and continuous, but information of sketches is refined and discrete.

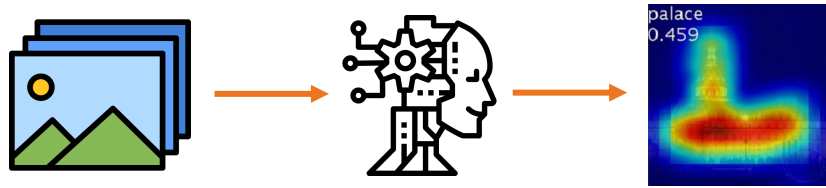


## Text-based Explainable Methods

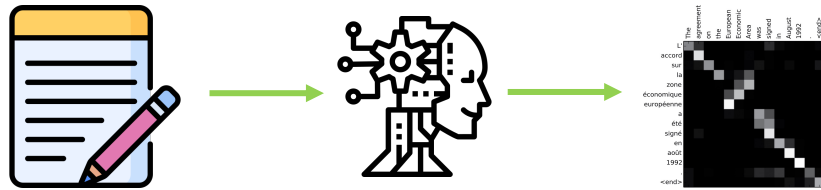


- *Sketches's high information density is similar to that of text. But words are naturally familiar to human, while points are not.*

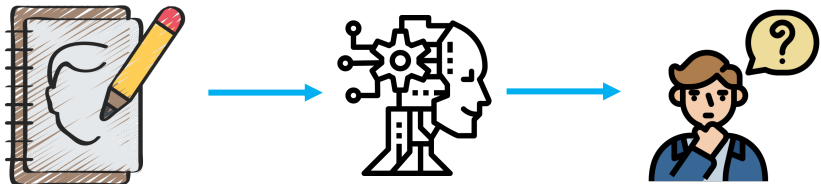




images → pixels / super pixels / patches → semantics



text → word

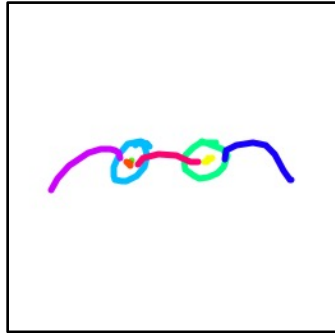


sketches → pixels / points → semantics

strokes

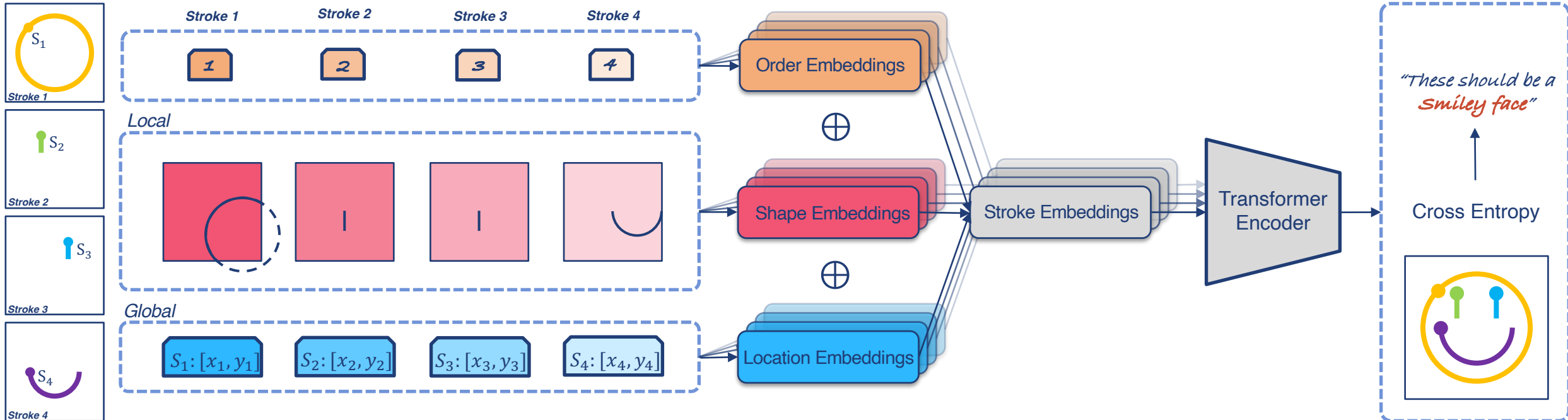


# A Stroke-based Framework – SketchXAI Net



## Strokes Attributes

- Order
- Shape
- Location

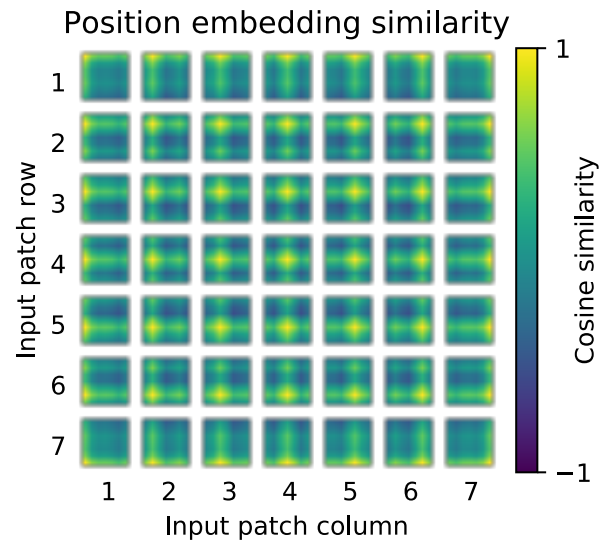


Methods	Acc. (%)	Params (M)
ResNet-50	78.76	24.2
Sketch-a-Net	68.71	8.5
SketchMate	80.51	64.7
ViT-Base	77.90	86.6
Swin-Base	78.71	87.8
SketchFormer	78.34	13.1
SketchAA	81.51	26.7
Sketch-R2CNN (ResNet-50)	84.81	32.7
Sketch-R2CNN (ResNet-101)	85.30	51.7
-----		
SketchXAINet-Tiny (No Shape)	31.04	-
SketchXAINet-Tiny (No Location)	81.41	-
SketchXAINet-Tiny (No Order)	83.66	-
-----		
SketchXAINet-Tiny	86.10	6.1
<b>SketchXAINet-Base</b>	<b>87.21</b>	<b>91.7</b>

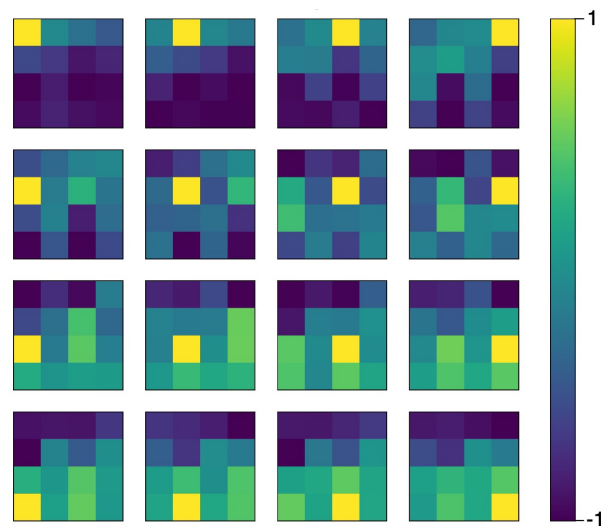




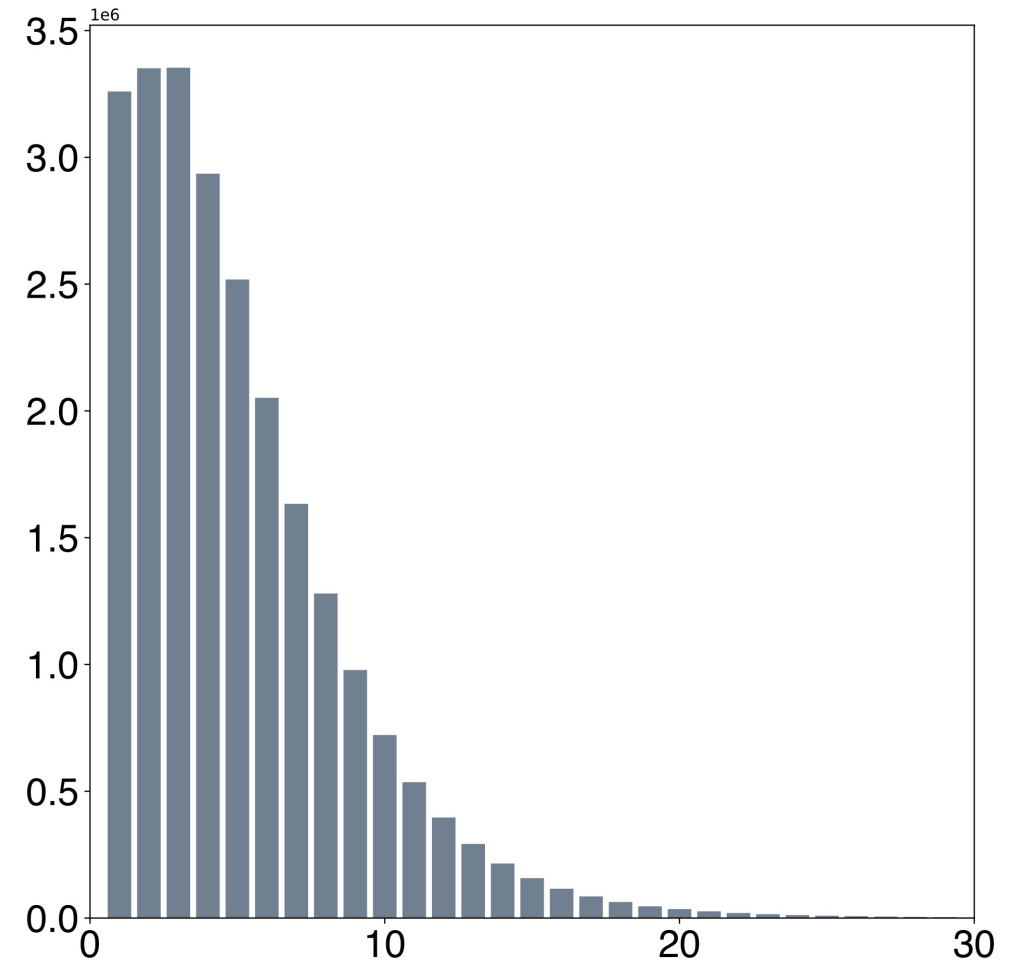
ViT



Ours



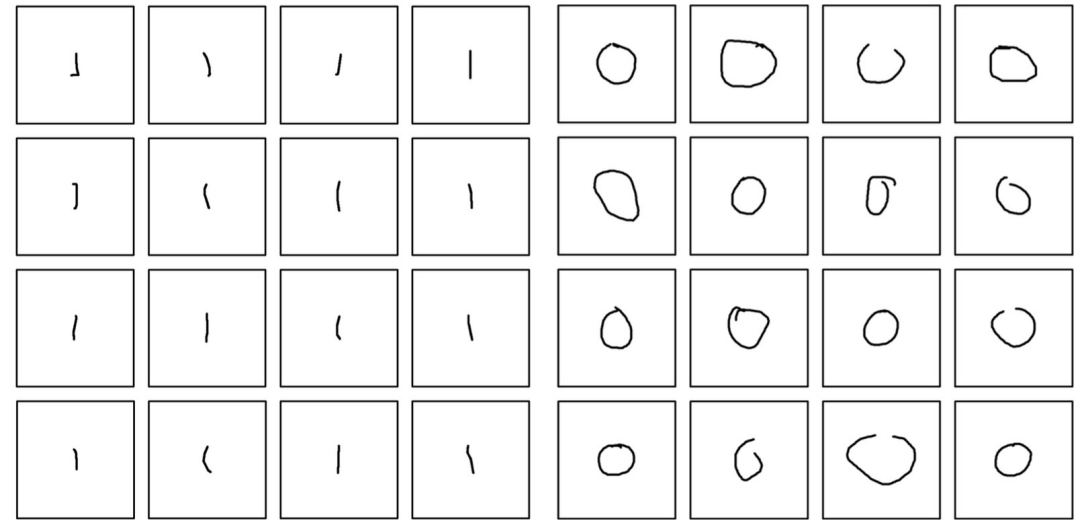
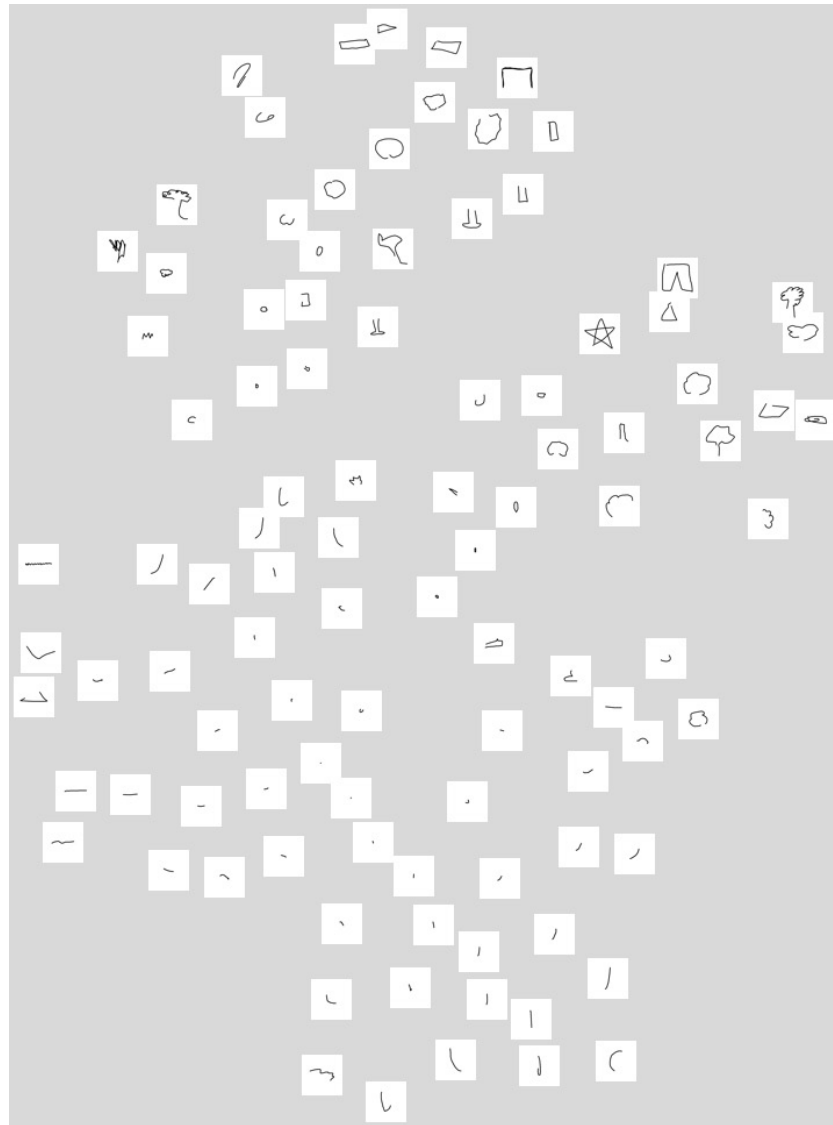
## Number of strokes statistics



91.44% of sketches have no more than 10 strokes.

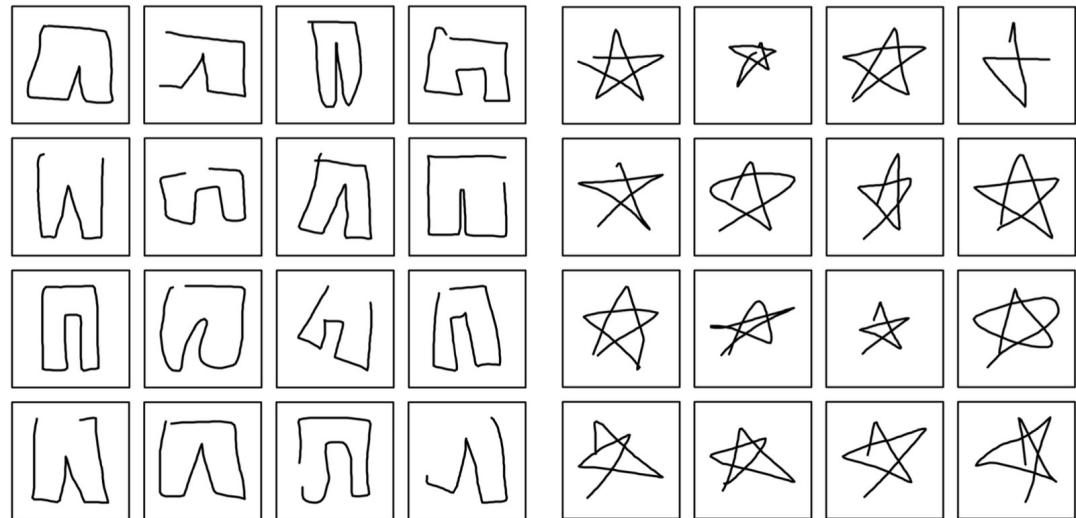


# Shape Analysis



(a) ID=15

(b) ID=67

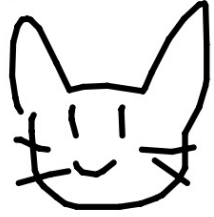
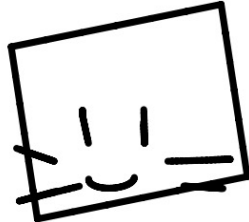
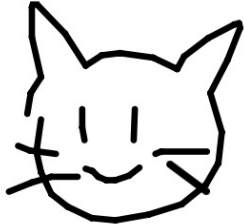

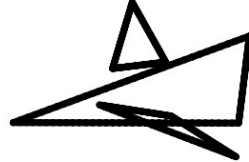
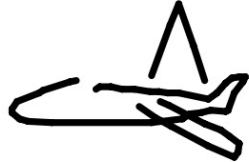




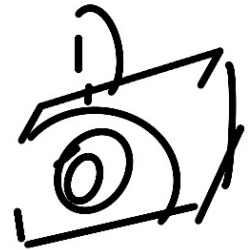
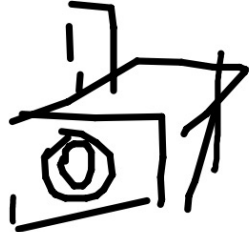
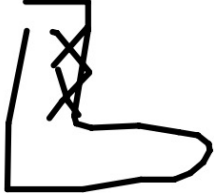
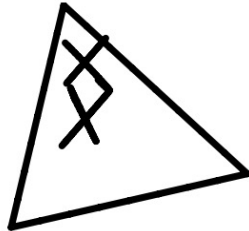
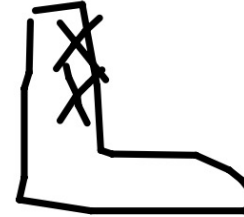


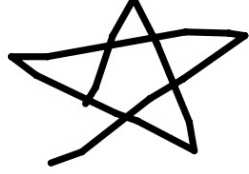


(c) ID=19

(d) ID=14

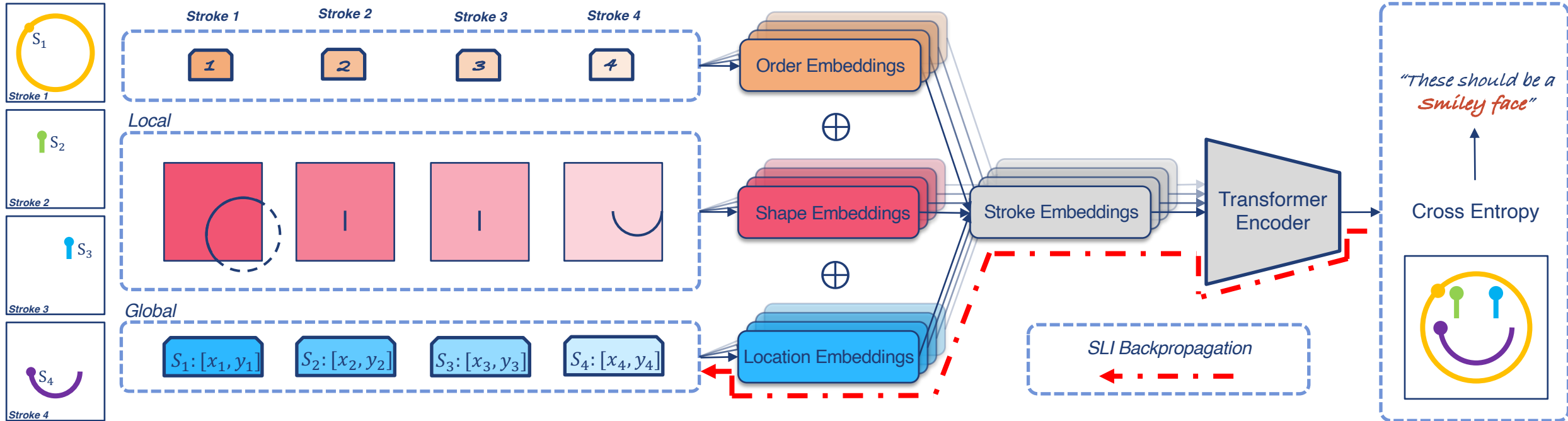


# Shape Analysis

Input Sketch	Primitives	Ours	Input Sketch	Primitives	Ours
<p>cat</p> 			<p>airplane</p> 		
<p>firetruck</p> 			<p>camera</p> 		
<p>shoe</p> 			<p>star</p> 		



# Stroke Location Inversion



## Recovery

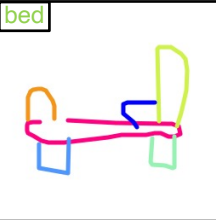
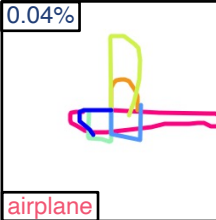

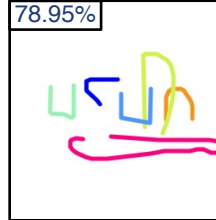
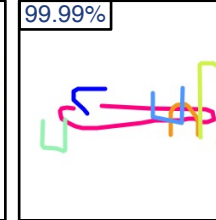
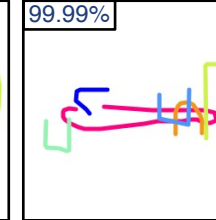
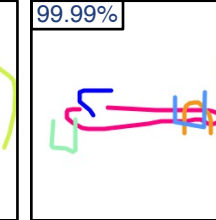
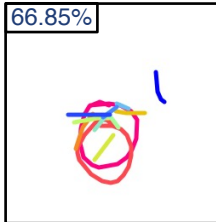
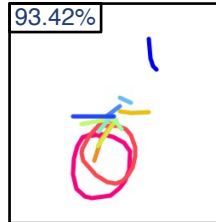
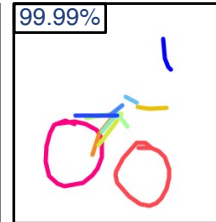
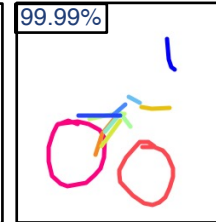
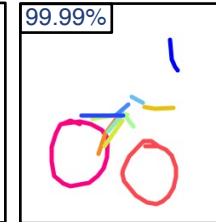


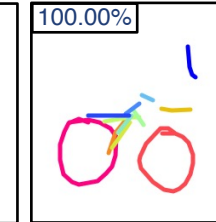
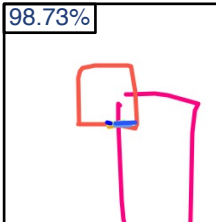
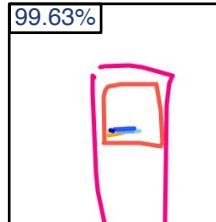
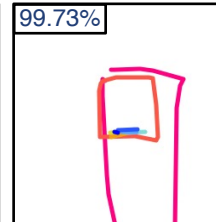
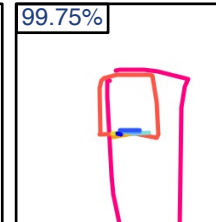
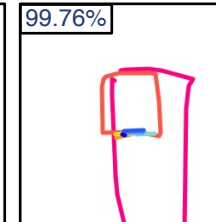
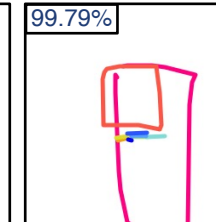
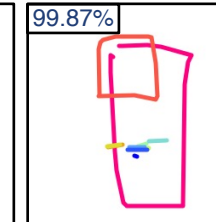
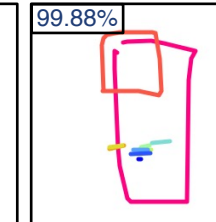
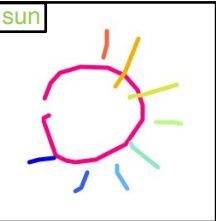
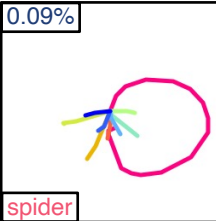
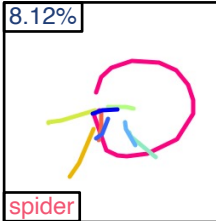
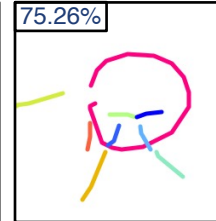
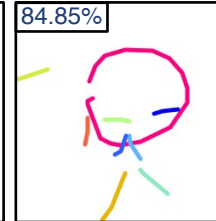
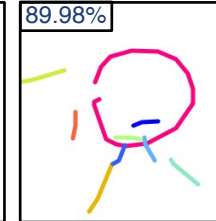
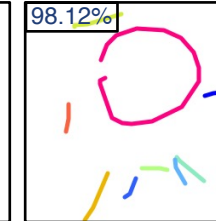
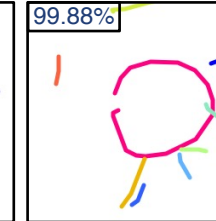
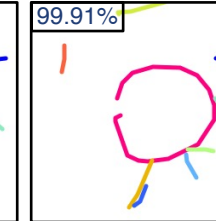

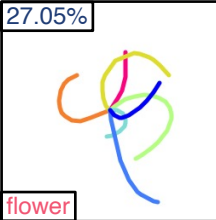
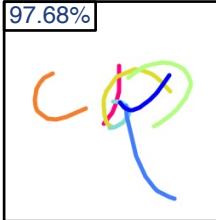
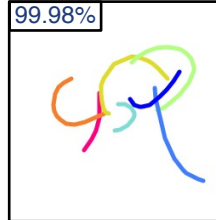
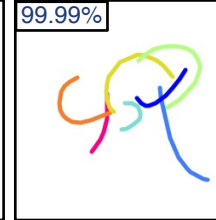
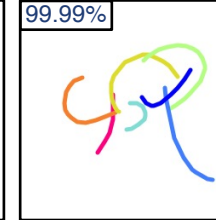
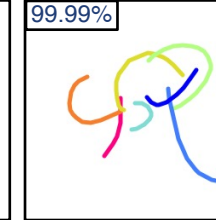
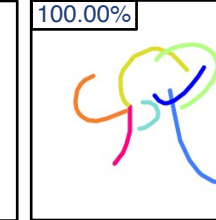
set locations all randomly and the label to original category, see if our model can be recover the sketch.

## Transfer

set the label to another category, see if our model can reorganise the strokes.

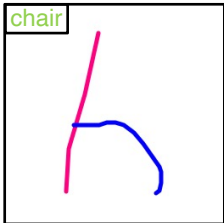
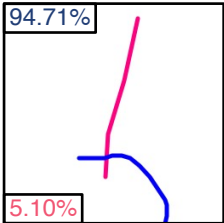
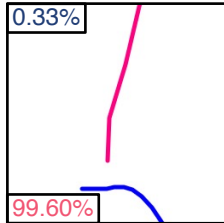
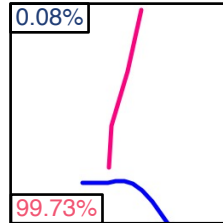
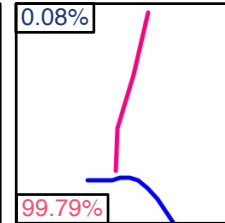
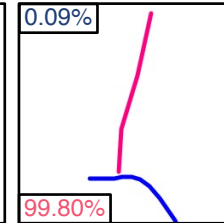
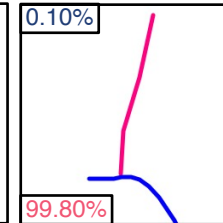
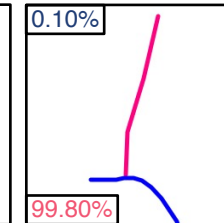
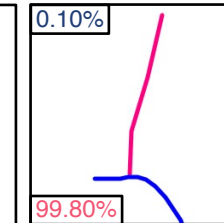





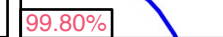


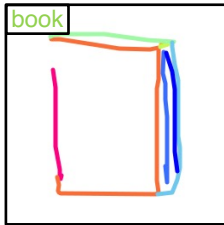
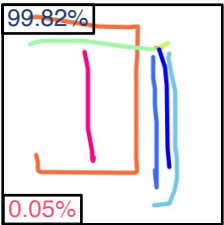
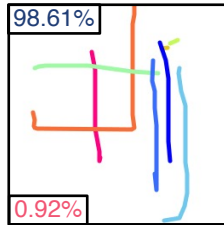
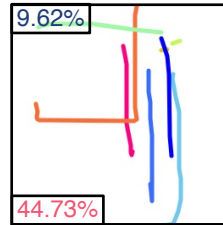
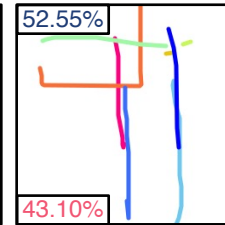
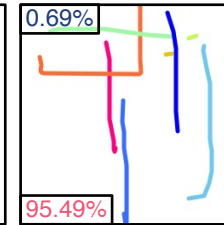
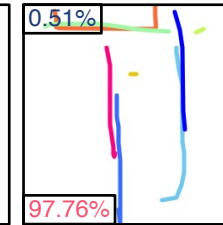
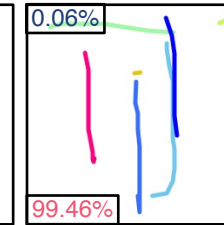
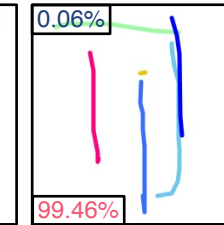








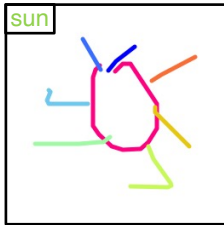
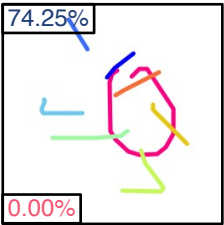
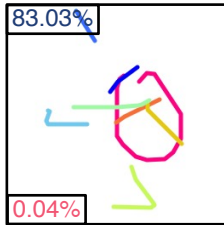
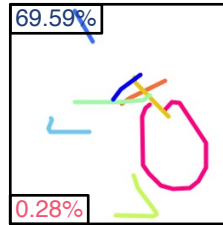
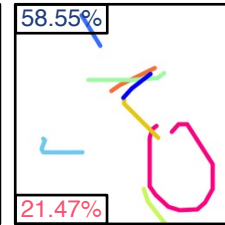
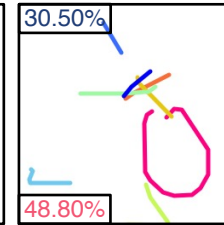
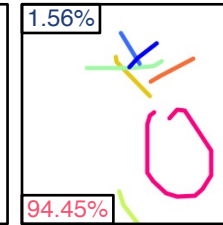
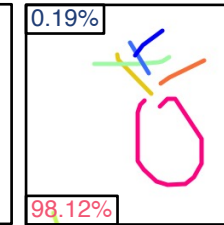
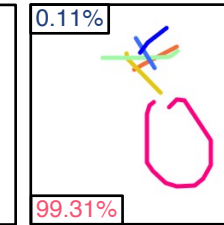








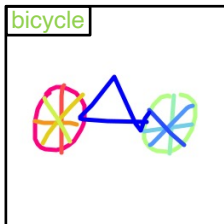
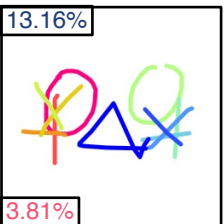
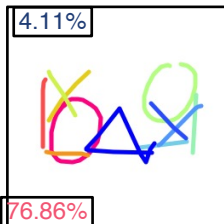
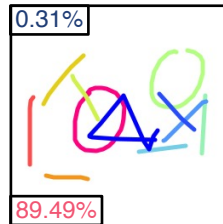
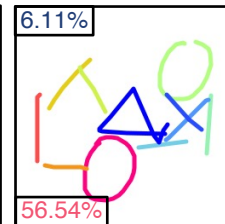
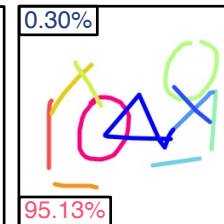
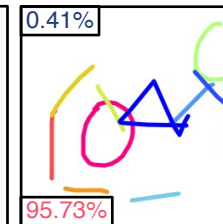
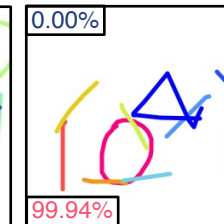
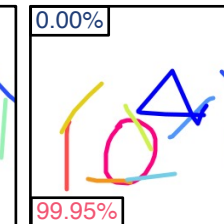









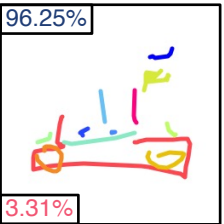
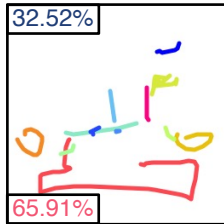
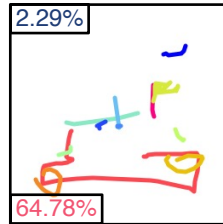
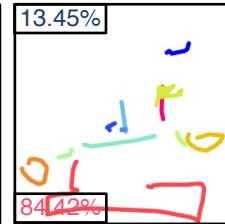
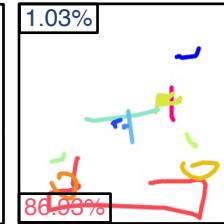
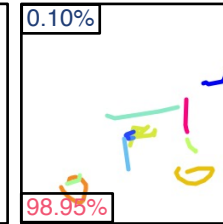
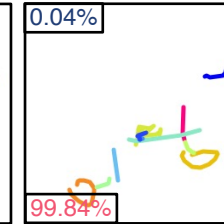
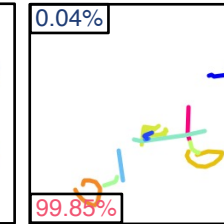










# SLI -- Recovery

Origin	Input	Iter. 1	Iter. 2	Iter. 3	Iter. 4	Iter. 5	Iter. 10	Iter. 50	Iter. 100
 bed	 0.04% airplane	 12.81% airplane	 78.95%	 99.66%	 99.98%	 99.99%	 99.99%	 99.99%	 99.99%
 bicycle	 4.74% eye	 66.85%	 93.42%	 99.99%	 99.99%	 99.99%	 99.99%	 100.00%	 100.00%
 cell phone	 80.13%	 98.73%	 99.63%	 99.73%	 99.75%	 99.76%	 99.79%	 99.87%	 99.88%
 sun	 0.09% spider	 8.12% spider	 45.67% spider	 75.26%	 84.85%	 89.98%	 98.12%	 99.88%	 99.91%
 tree	 27.05% flower	 97.68%	 99.98%	 99.99%	 99.99%	 99.99%	 99.99%	 100.00%	 100.00%



# SLI -- Transfer

Origin	Input	Iter. 1	Iter. 2	Iter. 3	Iter. 4	Iter. 5	Iter. 10	Iter. 50	Iter. 100
	broom	94.71% 	0.33% 	0.08% 	0.08% 	0.09% 	0.10% 	0.10% 	0.10% 
		5.10% 	99.60% 	99.73% 	99.79% 	99.80% 	99.80% 	99.80% 	99.80% 
	pants	99.82% 	98.61% 	9.62% 	52.55% 	0.69% 	0.51% 	0.06% 	0.06% 
		0.05% 	0.92% 	44.73% 	43.10% 	95.49% 	97.76% 	99.46% 	99.46% 
	apple	74.25% 	83.03% 	69.59% 	58.55% 	30.50% 	1.56% 	0.19% 	0.11% 
		0.00% 	0.04% 	0.28% 	21.47% 	48.80% 	94.45% 	98.12% 	99.31% 
	camera	13.16% 	4.11% 	0.31% 	6.11% 	0.30% 	0.41% 	0.00% 	0.00% 
		3.81% 	76.86% 	89.49% 	56.54% 	95.13% 	95.73% 	99.94% 	99.95% 
	bicycle	96.25% 	32.52% 	2.29% 	13.45% 	1.03% 	0.10% 	0.04% 	0.04% 
		3.31% 	65.91% 	64.78% 	84.12% 	86.89% 	98.95% 	99.84% 	99.85% 



## General Training (performance)

$$\arg \min_w \ell(f_w(x_i), y_i)$$

Use data to optimize a model

Q: “What label matches the sample?”

```
data.requires_grad = False
model.requires_grad = True

prediction = model(data)
loss = criterion(prediction, label)
loss.backward()
```

## Model Inversion (explainability)

$$\arg \min_{x_i - part} \ell(f_w(x_i), y)$$

Use a (trained) model to optimize data

Q: “What sample matches the label?”

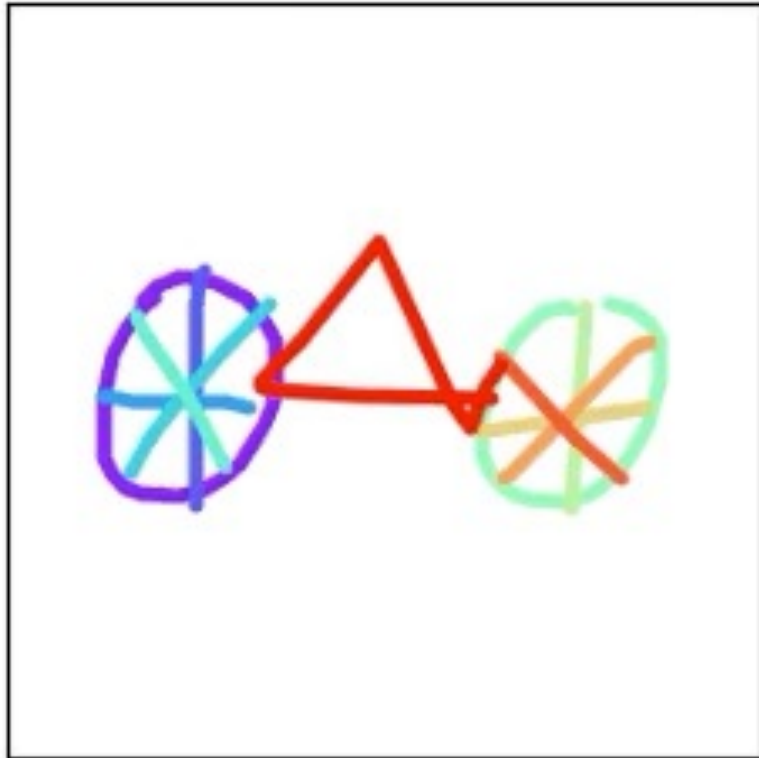
```
data.requires_grad = True
model.requires_grad = False

prediction = model(data)
loss = criterion(prediction, label)
loss.backward()
```

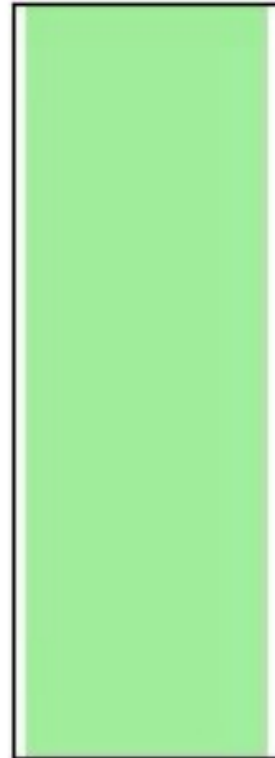


# Dynamic Process of SLI

bicycle-0



bicycle

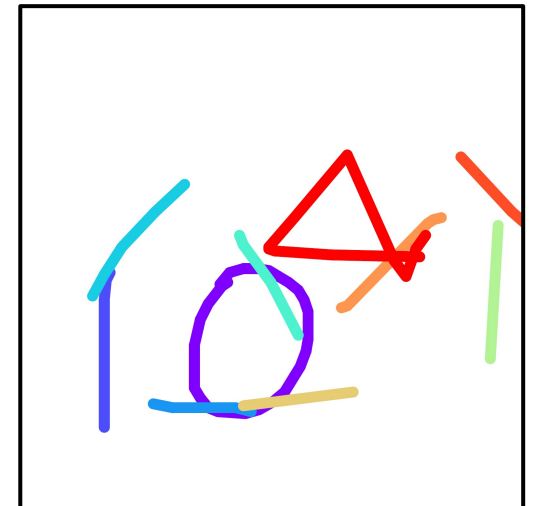
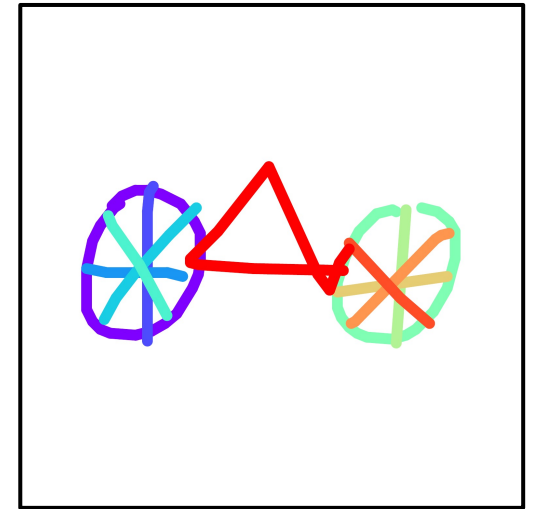


99.968719%

camera



0.001203%





SCAN ME



Thank you! For more details, please visit:  
<https://sketchxai.github.io/>

# SketchXAI: A First Look at Explainability of Human Sketches

CVPR 2023: THU-PM-260

