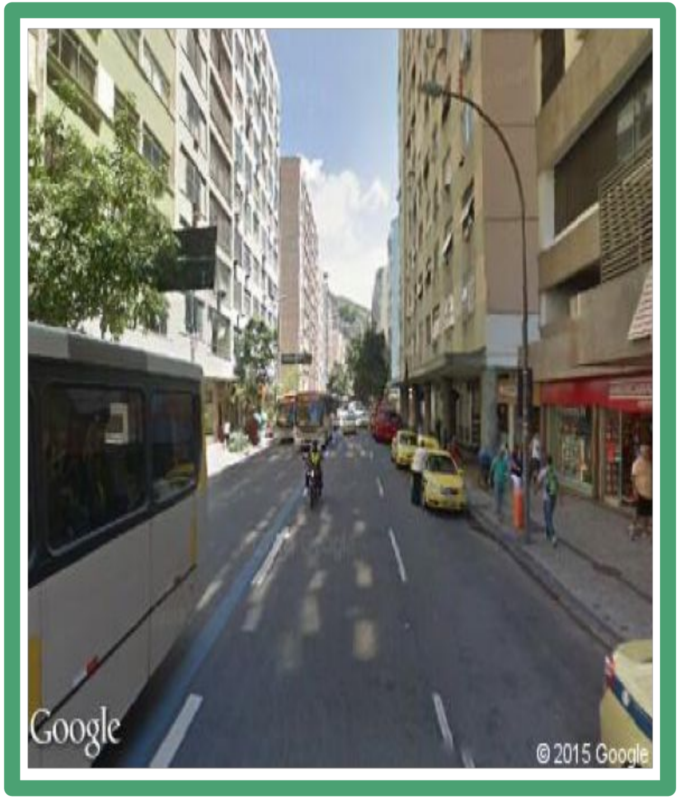
What Makes a Place Feel Safe? Analyzing Street View Images to Identify Relevant Visual Elements

Felipe A. Moreno-Vera, Bruno Brandoli, Jorge Poco

Motivation

Which one looks safer?

Bangú (RJ)



City Center (RJ)

Motivation

By understanding how people perceive and experience cities, we can create more inclusive, attractive, and functional urban solutions that meet the needs and aspirations of their diverse populations.

Context

Urban perception is shaped by a complex interplay of factors. Such as physical design, architectural styles, street layouts, landmarks, and the quality of infrastructure all contribute to the visual characteristics that define a city's identity.

Place Pulse

Place Pulse

Which place looks **livelier** ?

For this question: **362,708** clicks collected

Goal: **500,000** clicks

SEE REAL-TIME RANKINGS

RANK	CITY	CLICKS	TREND
1	Washington DC	6296	
2	London	17982	
3	New York	22424	

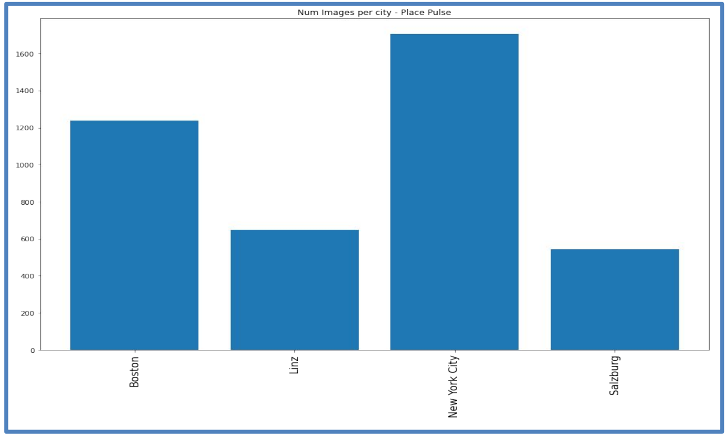
RANK	CITY	CLICKS	TREND
54	Cape Town	16228	
55	Belo Horizonte	12728	
56	Gaborone	4717	

http://pulse.media.mit.edu/

* Comparisons were made using two random images from random cities.

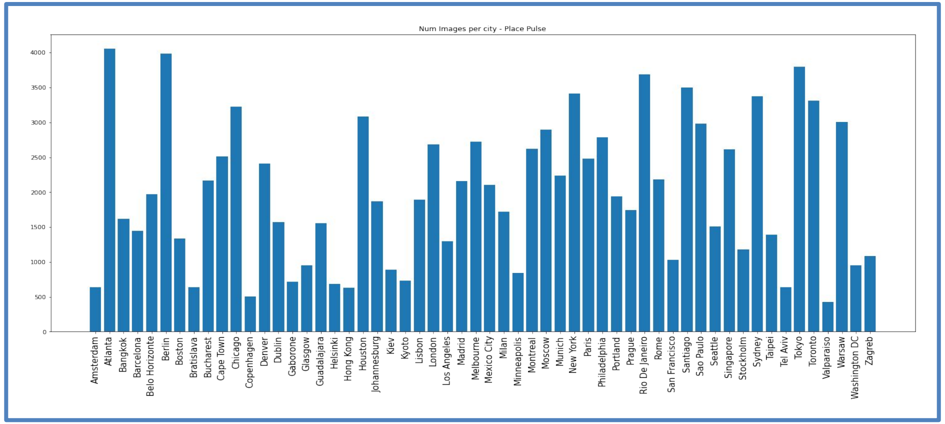
Place Pulse 1.0

- Release date: 2013
- 73 806 Comparisons
- 4 136 images
- 2 Countries
- 4 cities
- 3 categories



Place Pulse 2.0

- Release date: 2016
- 1 223 649 Comparisons
- 111 390 images
- 32 countries
- 56 cities
- 6 categories



Data Preparation

Data samples

left-id	right-id	winner	left-lat	left-long	right-lat	right-long	category
513d7e23fdc9f	513d7ac3fdc9f	equal	40.744156	-73.93557	-33.52638	-70.591309	depressing
513f320cfdc9f	513cc3acfdc9f	left	52.551685	13.416548	29.76381	-95.394621	safety
513e5dc3fdc9f	5140d960fdc9f	right	48.878382	2.403116	53.32932	-6.231007	lively

Perceptual Scores

$$W_i = \frac{w_i}{w_i + d_i + l_i}$$

$$L_i = \frac{l_i}{w_i + d_i + l_i}$$

*

$$q_{i,k} = \frac{10}{3}(W_{i,k} + \frac{1}{n^w_{i,k}}(\sum_{j_1} W_{j_1,k}) - \frac{1}{n^l_{i,k}}(\sum_{j_2} L_{j_2,k}) + 1)$$

*Nassar et al, “The evaluative image of the city”, 1990
Salesse et. al, “The Collaborative Image of The City: Mapping the Inequality of Urban Perception”, 2013

Rank Scores

$$\mu_x \longleftarrow \mu_x + \frac{\sigma_x^2}{c} \cdot f\left(\frac{(\mu_x - \mu_y)}{c}, \frac{\varepsilon}{c}\right)$$

$$\mu_y \longleftarrow \mu_y - \frac{\sigma_y^2}{c} \cdot f\left(\frac{(\mu_x - \mu_y)}{c}, \frac{\varepsilon}{c}\right)$$

$$\sigma_x^2 \longleftarrow \sigma_x^2 \cdot \left[1 - \frac{\sigma_x^2}{c} \cdot g\left(\frac{(\mu_x - \mu_y)}{c}, \frac{\varepsilon}{c}\right)\right]$$

$$\sigma_y^2 \longleftarrow \sigma_y^2 \cdot \left[1 - \frac{\sigma_y^2}{c} \cdot g\left(\frac{(\mu_x - \mu_y)}{c}, \frac{\varepsilon}{c}\right)\right]$$

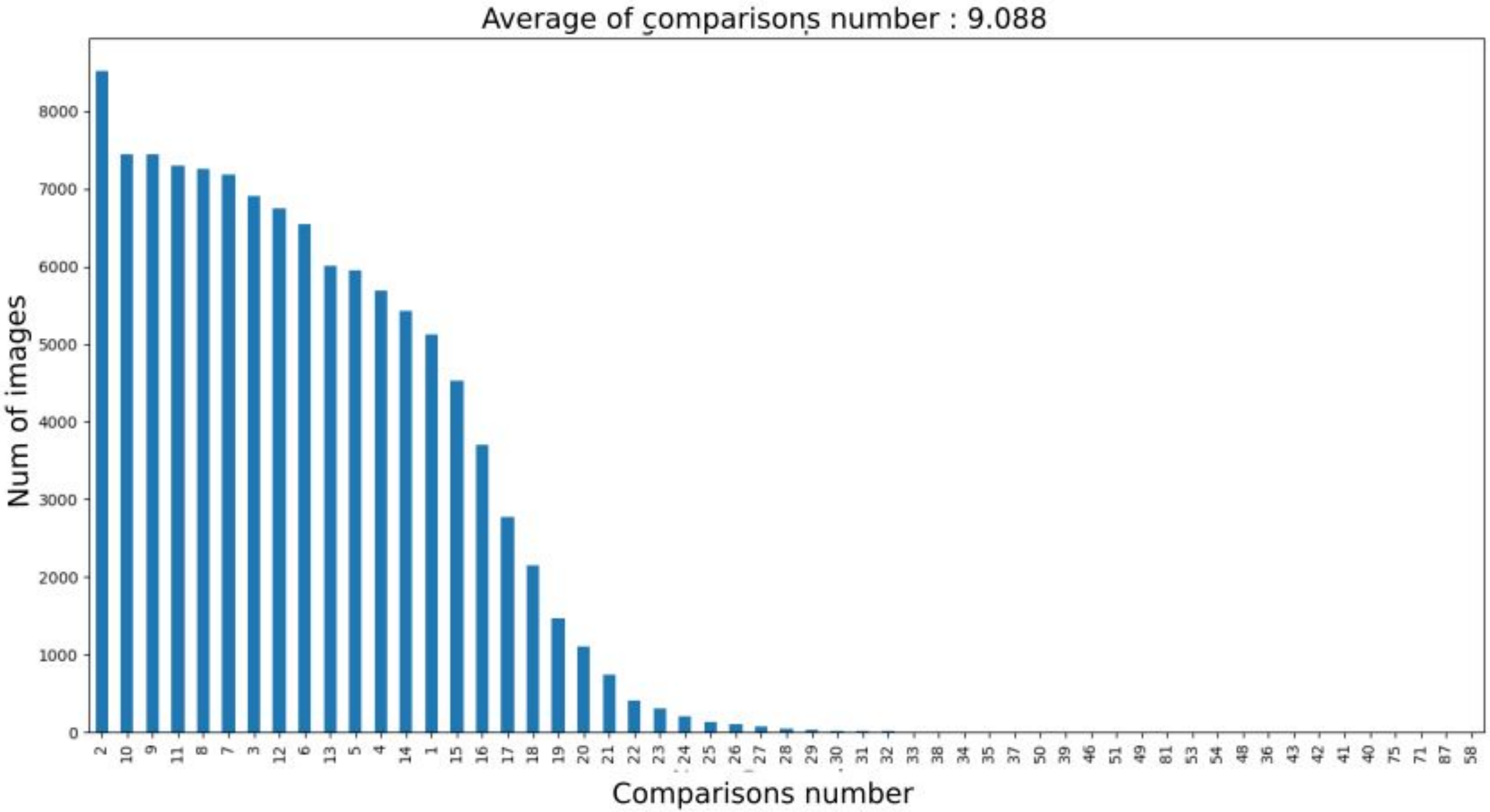
$$c^2 = 2\beta^2 + \sigma_x^2 + \sigma_y^2$$

**

$$q_{i,k} = \frac{10}{c_{max,k}}(c_{i,k})$$

**Minka et al, “TrueSkill 2: An improved Bayesian skill rating system”, 2018
Dubey et. al, “Deep Learning the City : Quantifying Urban Perception At A Global Scale”, 2016

Number of comparisons



Perceptual Scores

$$W_i = \frac{w_i}{w_i + d_i + l_i}$$

$$L_i = \frac{l_i}{w_i + d_i + l_i}$$

$$q_{i,k} = \frac{10}{3}^{*}(W_{i,k} + \frac{1}{n^w_{i,k}}(\sum_{j_1} W_{j_1,k}) - \frac{1}{n^l_{i,k}}(\sum_{j_2} L_{j_2,k}) + 1)$$

*Nassar et al, "The evaluative image of the city", 1990
Salesse et. al, "The Collaborative Image of The City: Mapping the Inequality of Urban Perception", 2013

Rank Scores

$$\mu_x \longleftarrow \mu_x + \frac{\sigma_x^2}{c} \cdot f\left(\frac{(\mu_x - \mu_y)}{c}, \frac{\varepsilon}{c}\right)$$

Not enough comparisons

$$= 2\beta^2 + \sigma_x^2 + \sigma_y^2$$

$$q_{i,k} = \frac{10}{c_{max,k}}^{**}(c_{i,k})$$

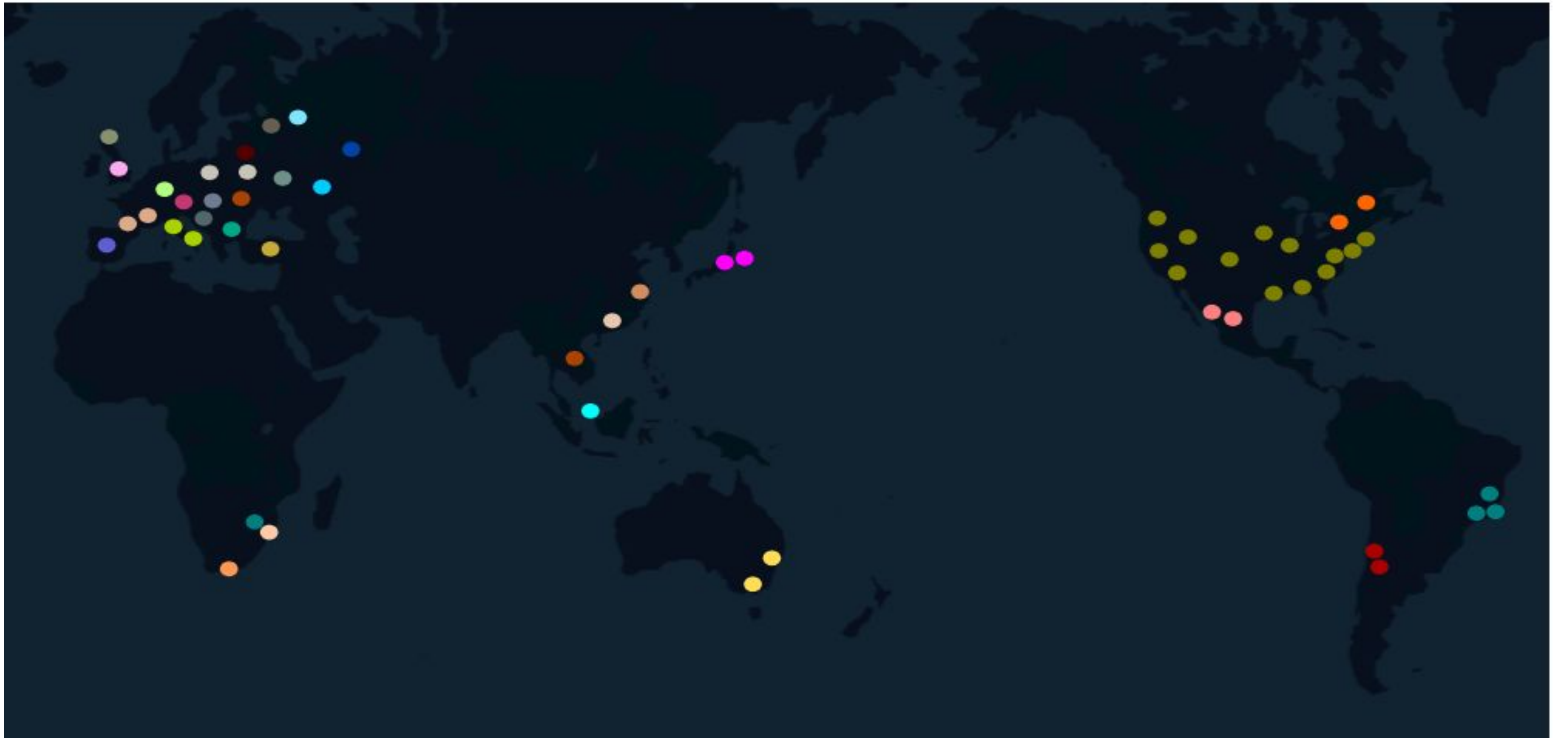
**Minka et al, "TrueSkill 2: An improved Bayesian skill rating system", 2018
Dubey et. al, "Deep Learning the City : Quantifying Urban Perception At A Global Scale", 2016

Processed samples

Image	ID	Safety	Lively	Wealthy	Beauty	Boring	Depressive
	513d7e23fdc9f	7.42	8.58	6.5	7.3	2.64	1.23
	513f320cfdc9f	6.07	4.97	7.13	8.61	1.67	0.86

Exploratory Data Analysis

Geographical city distribution



Note: Same color means same country.

Number of images per continent

Number of images per geographical level

Place Pulse 2.0				
Category/Level	City	Country	Continent	Global
safety	20,143	45,640	85,890	111,390
lively	14,803	38,216	79,788	111,349
Beautiful	9,410	28,811	66,792	110,767
Wealthy	7,642	24,326	57,780	107,796
Depressing	6,556	21,171	52,504	105,496
Boring	6,148	20,931	52,031	106,364

Summary

Place Pulse 2.0			
Continent	#countries	#cities	#images
Europe	19	22	38,747
North America	3	17	37504
South America	2	5	12,524
Asia	5	7	11,417
Oceania	1	2	6,097
Africa	2	3	5,101
Total	32	56	111,390

Place Pulse 2.0			
Category	# comparisons	# images	*mean*
Safety	368,926	111,389	5.188
Lively	267,292	111,348	5.085
Beautiful	175,361	110,766	4.920
Wealthy	152,241	107,795	4.890
Depressing	132,467	105,495	4.816
Boring	127,362	106,363	4.810
Total	1,223,649		

Urban Security Perception

Perceptual scores

left	right	winner
		draw
		left
		right
⋮	⋮	⋮
		right
		left

$$\hat{y}_{i,k} = q_{i,k}$$

I: (X,Y)

Image	Perceptual Scores
(	, 8.35)
(	, 7.16)
. . .	
(	, 5.01)
. . .	
(	, 1.29)
(	, 0.55)

Data labeling

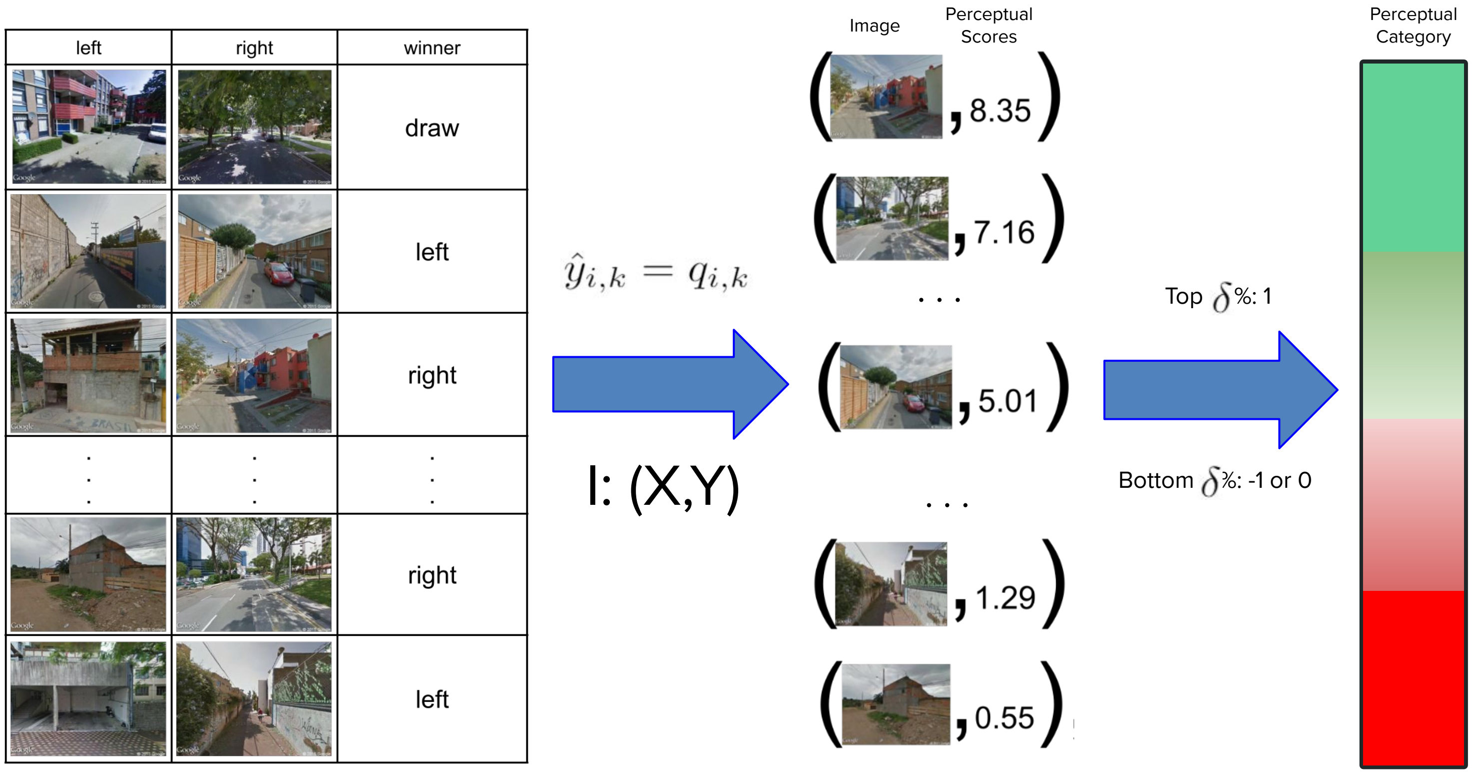
We define a parameter δ which will helps to labeling our data.

$$\hat{y}_{i,k} = q_{i,k}$$

$$\Downarrow$$

$$y_{i,k} = \begin{cases} 1 & \text{if } (q_{i,k})\text{in the top } \delta\% \\ \text{-1} & \text{if } (q_{i,k})\text{in the bottom } \delta\% \end{cases}$$

Perceptual category

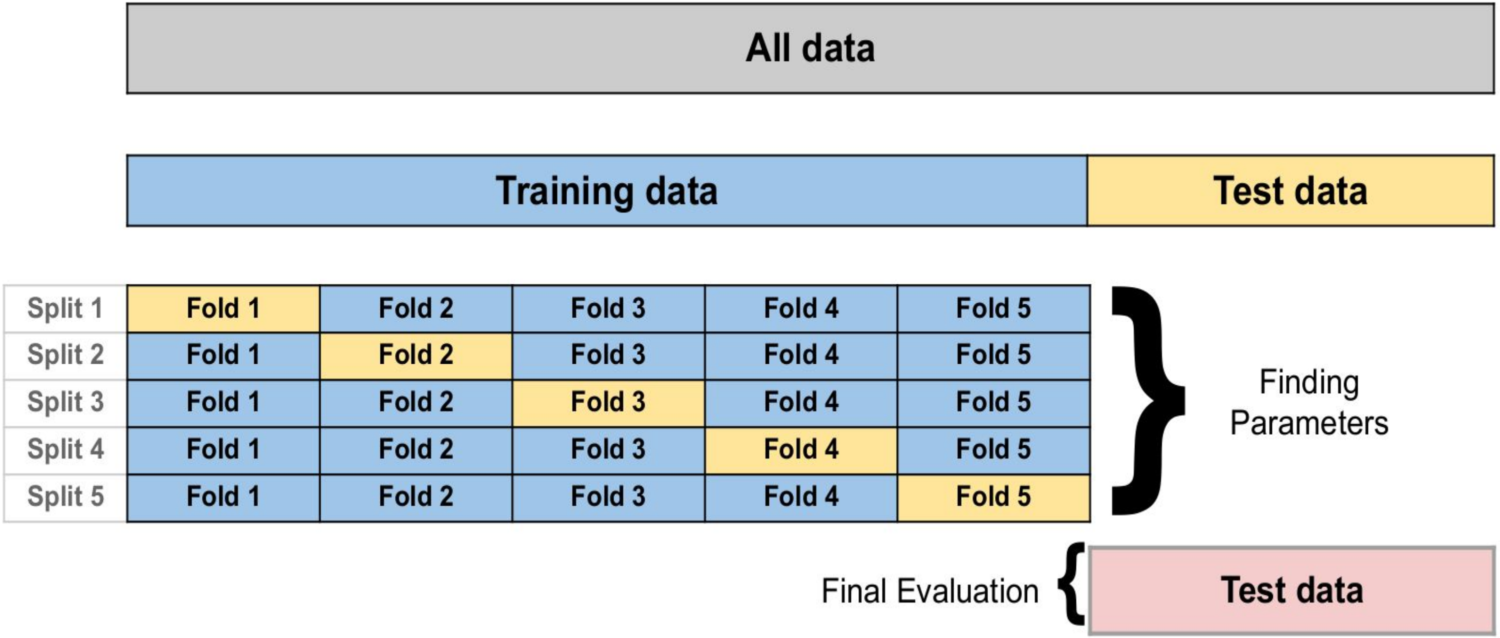


Experiments and Results

Data split

- **Oversampling method to balance classes and split data into 75% and 25%, respectively**

- **Hyperparameters tuning: Grid search using Stratified 5 Cross-Validation**



Classification details

- We fuse ViT B model and OneFormar segmentation model. Then, we use a dense layers to build our classifier

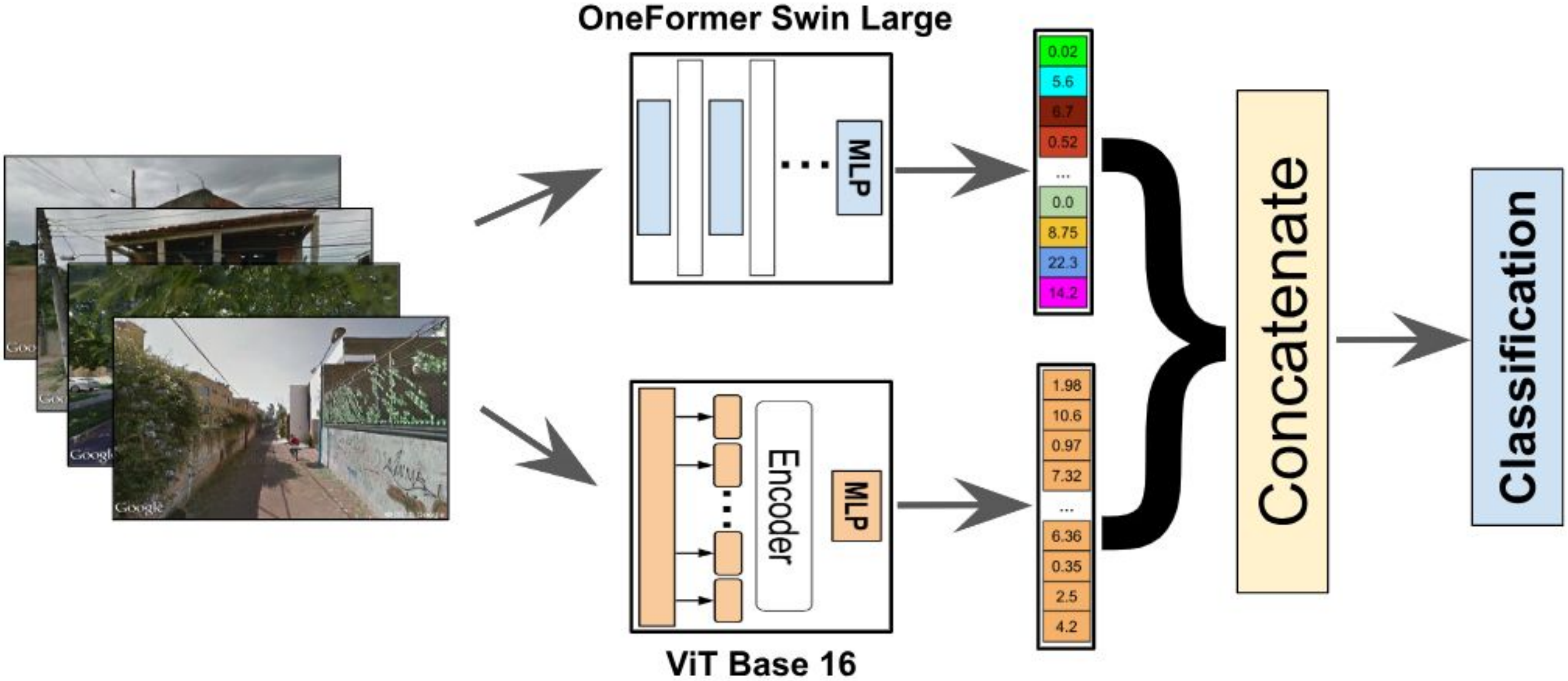
- We perform two tasks: binary classification (between safe and unsafety) and 10-label classification.

- The second one is dividing the range of the scores, e.g., 0-1 is the label 0, 1-2 is label 1, and so on.

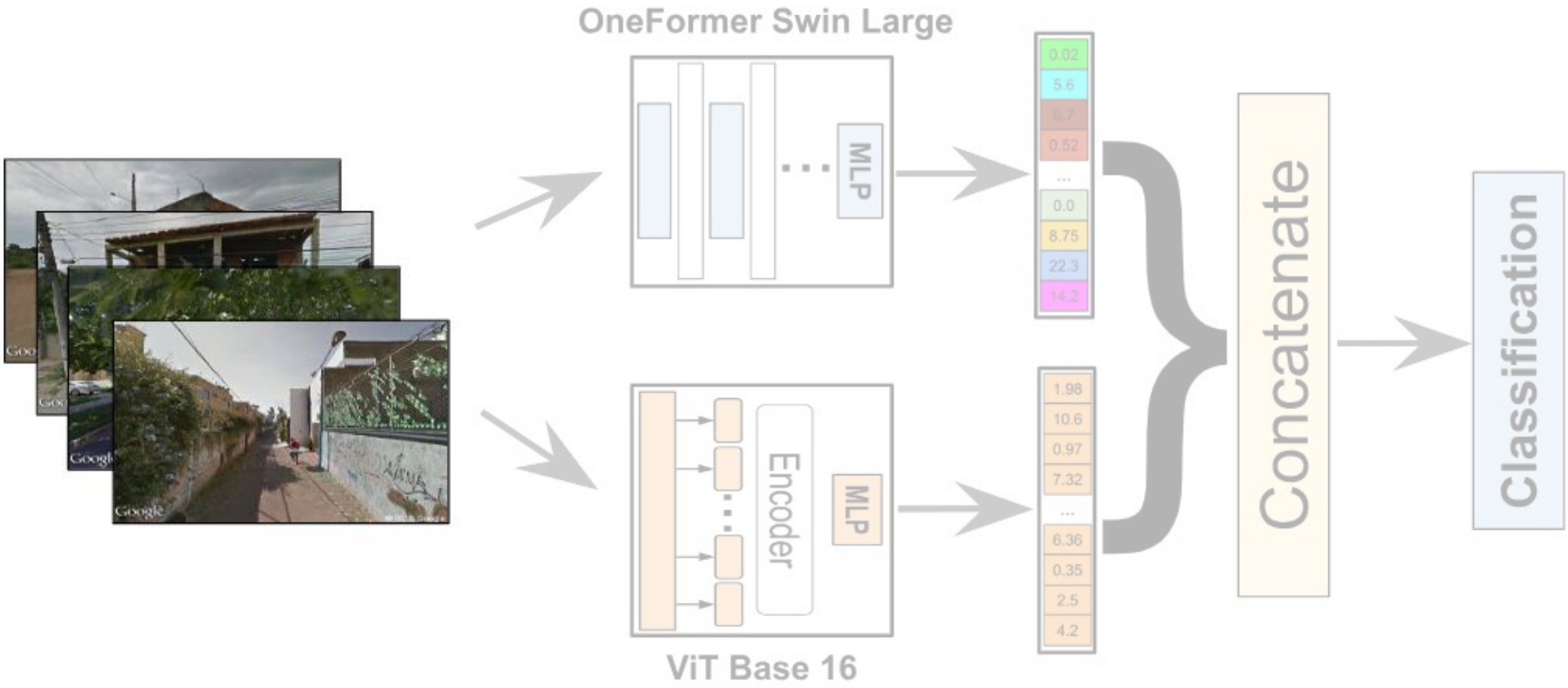
- We use the accuracy metric to compare with previous works

- We perform all experiments using a GTX 1650 Ti, 8 VRAM.

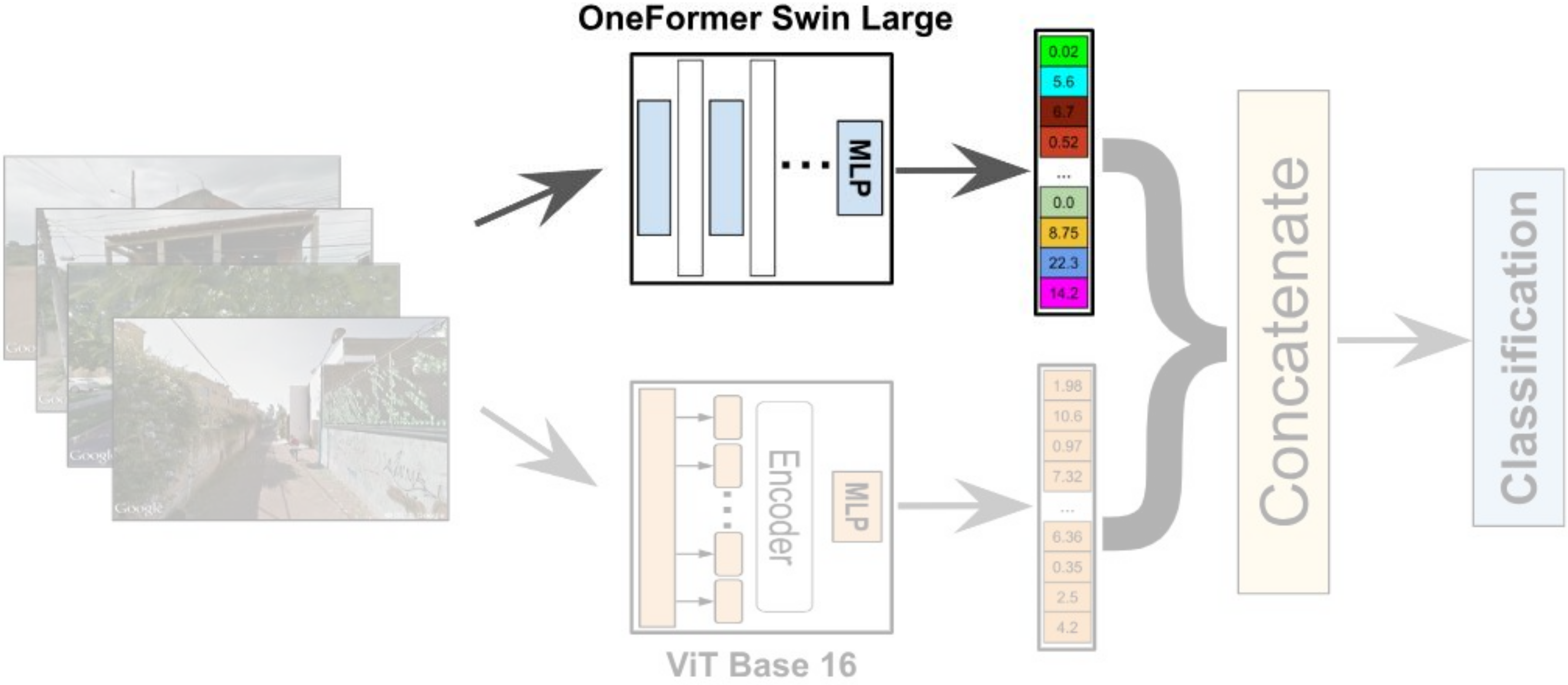
UrbanFormer



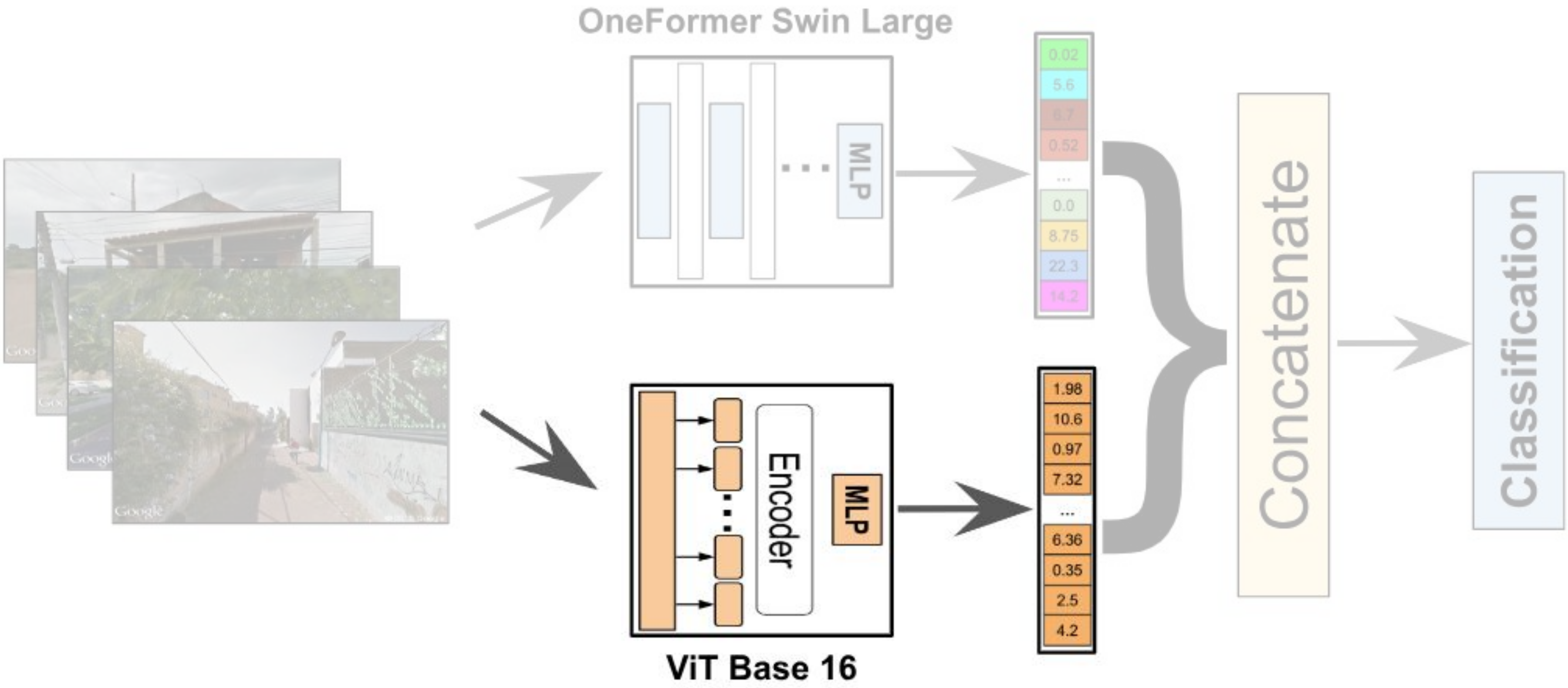
UrbanFormer



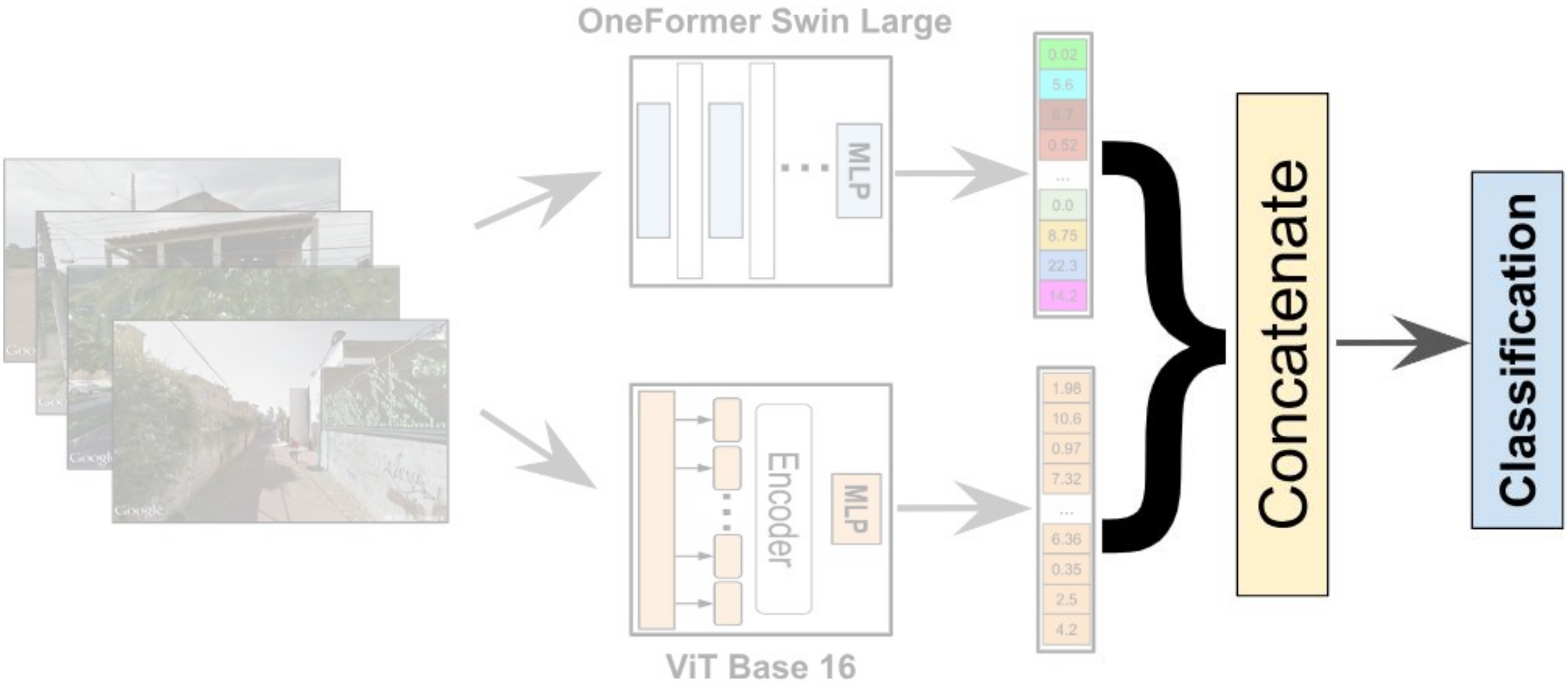
UrbanFormer



UrbanFormer



UrbanFormer



Results

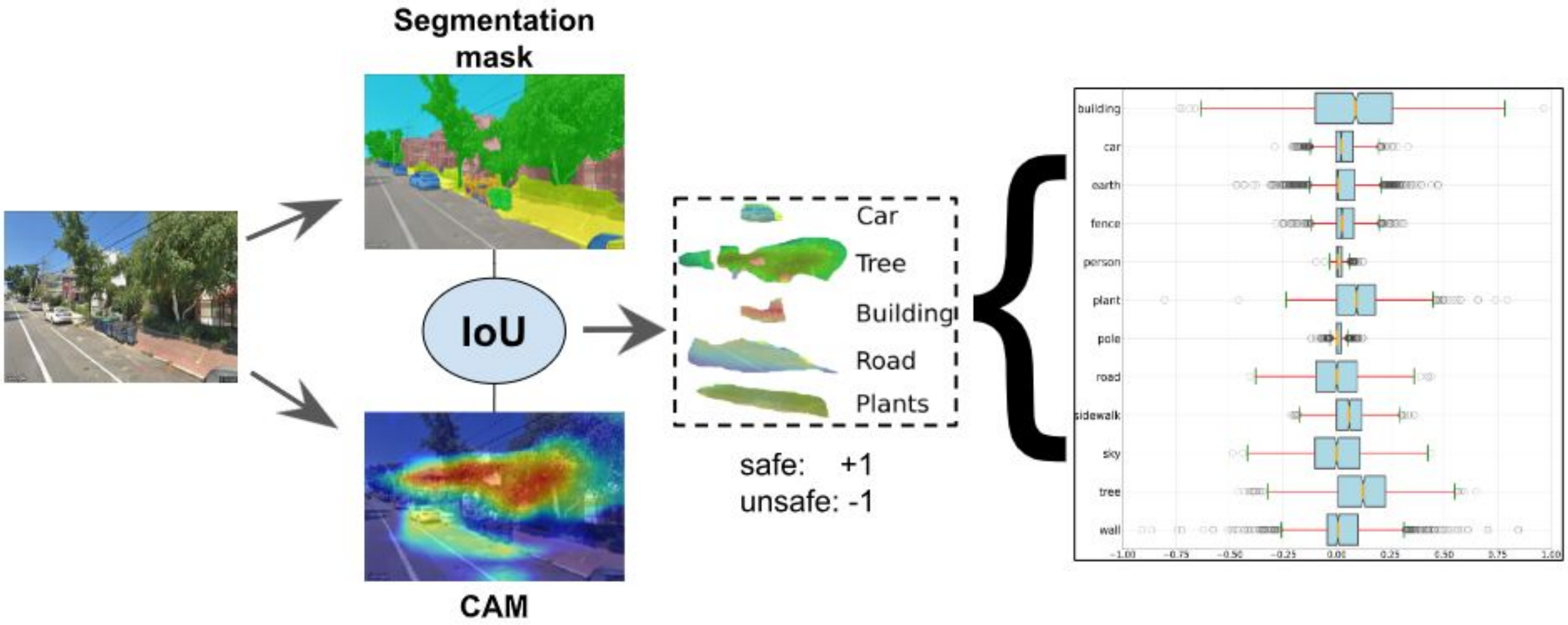
ACCURACY REPORT USING BINARY CLASSIFICATION

Model	Acc
PspNet+VGG [21]	48.38
DeepLabV3+VGG [21]	51.93
DSAPN+ResNet [43]	64.87
MTDRALN-LC [19]	65.07
MTDRALN-TC [19]	65.82
VGG+ImageNet [22]	65.72
VGG-GAP+ImageNet [22]	66.09
VGG+Places365 [22]	66.46
VGG-GAP+Places365 [22]	66.96
VGG19+ImageNet [4]	67.01
PSPNet+SVR [44]	70.63
DeiT+ResNet50 [32]	71.16
ViT-nn (Ours)	**71.29**
ViT-nn+OneFormer (Ours)	**75.68**

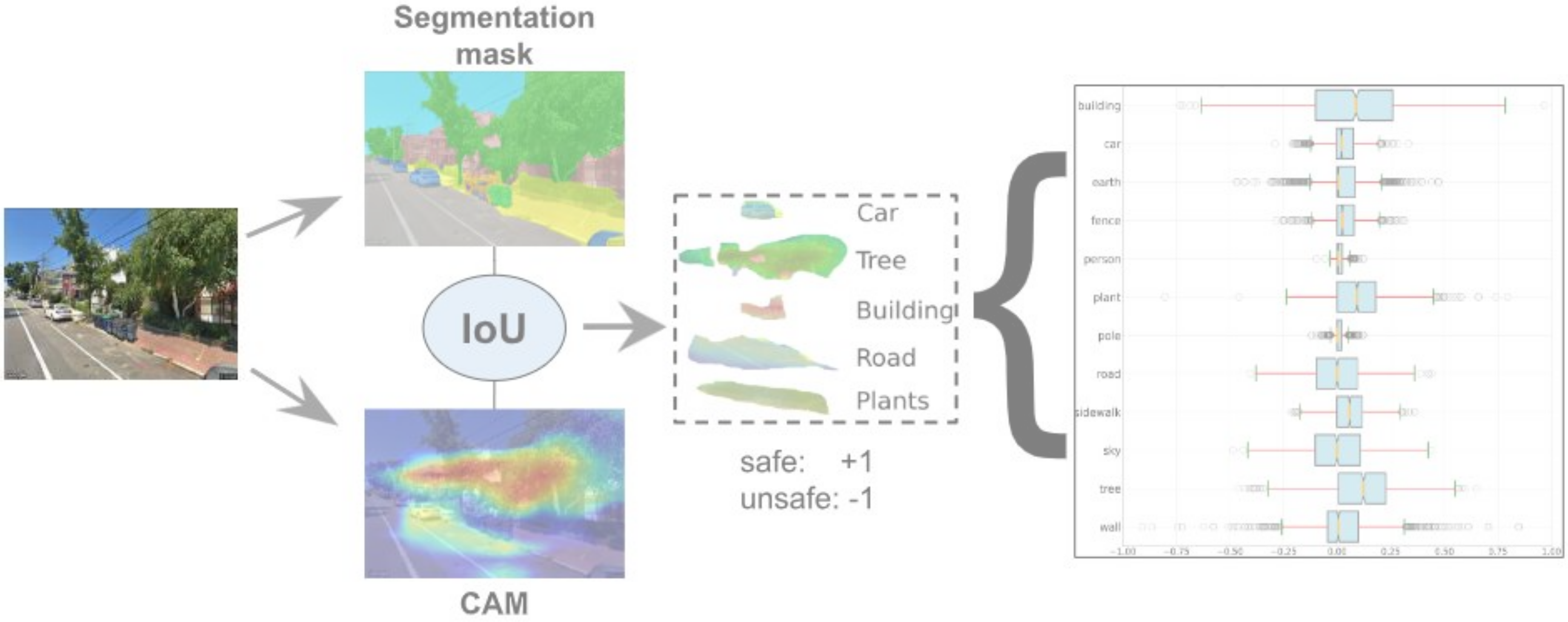
ACCURACY REPORT USING 10-LABEL CLASSIFICATION

Model	Acc
ResNet50 [18]	71.33
SegFormerB5+RF [46]	42.8
VGG19 [46]	75.2
ConvNeXt-B [46]	76.4
SFB5+ConvNeXt-B+RF [46]	78.1
ViT-nn (Ours)	74.97
ViT-nn+OneFormer (Ours)	**78.68**

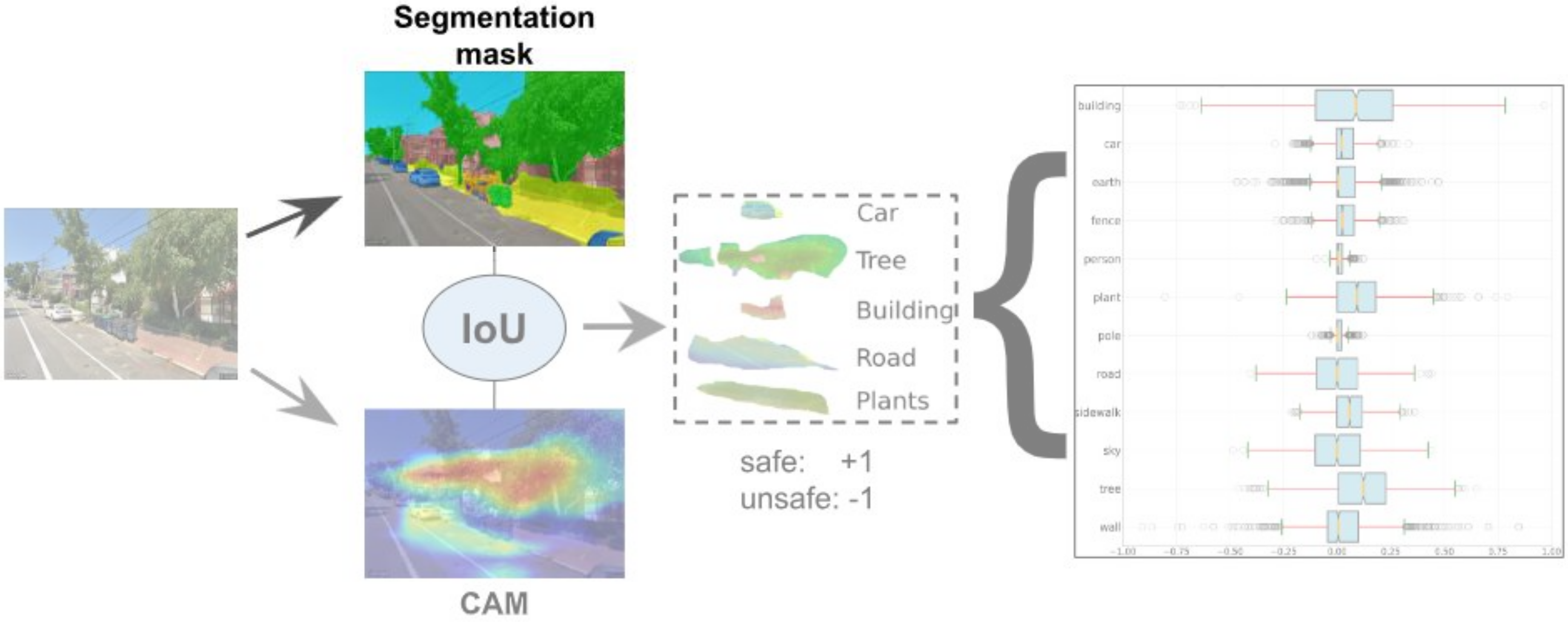
Explanation



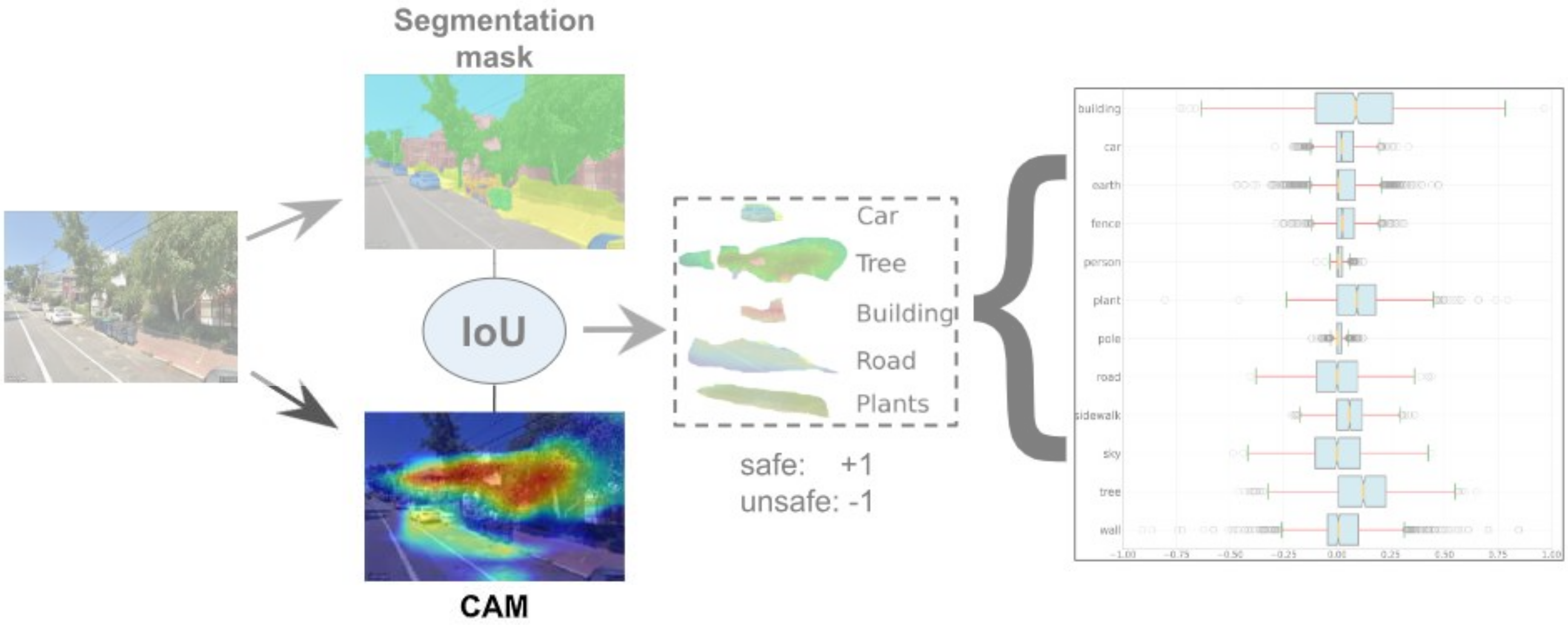
Explanation



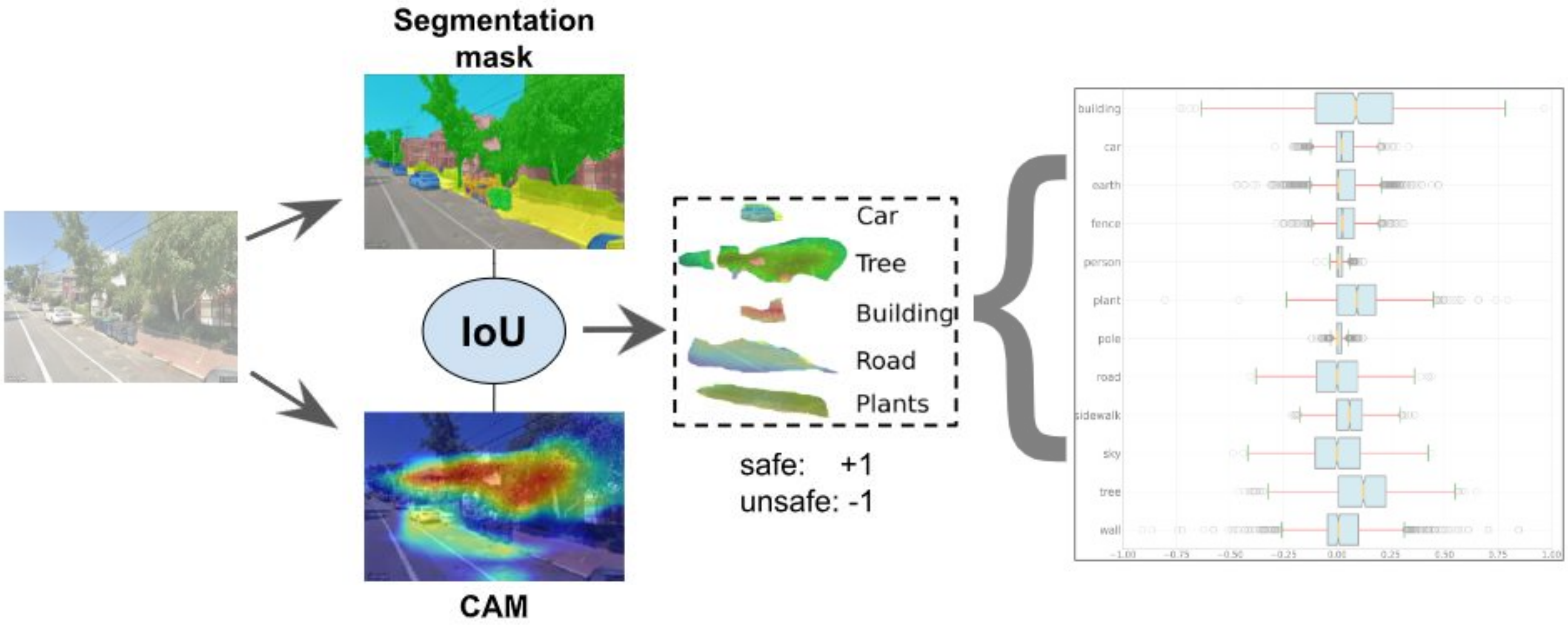
Explanation



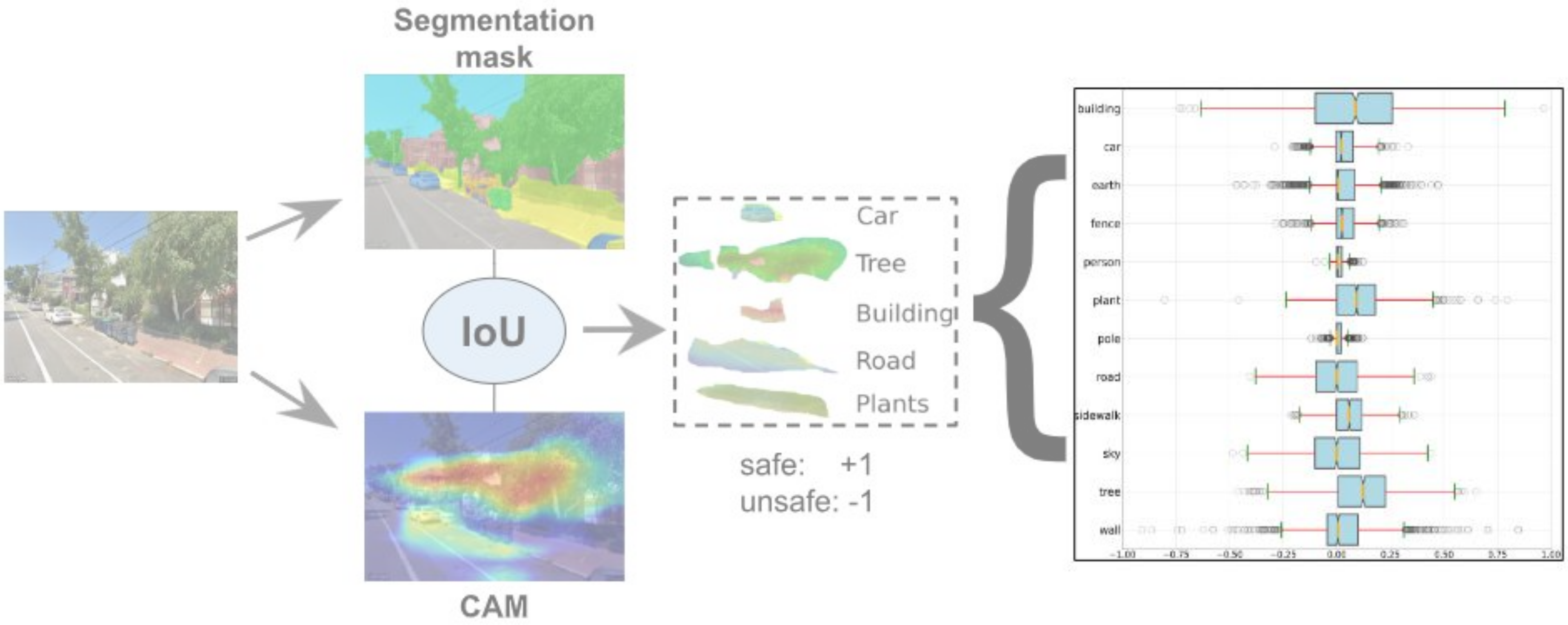
Explanation



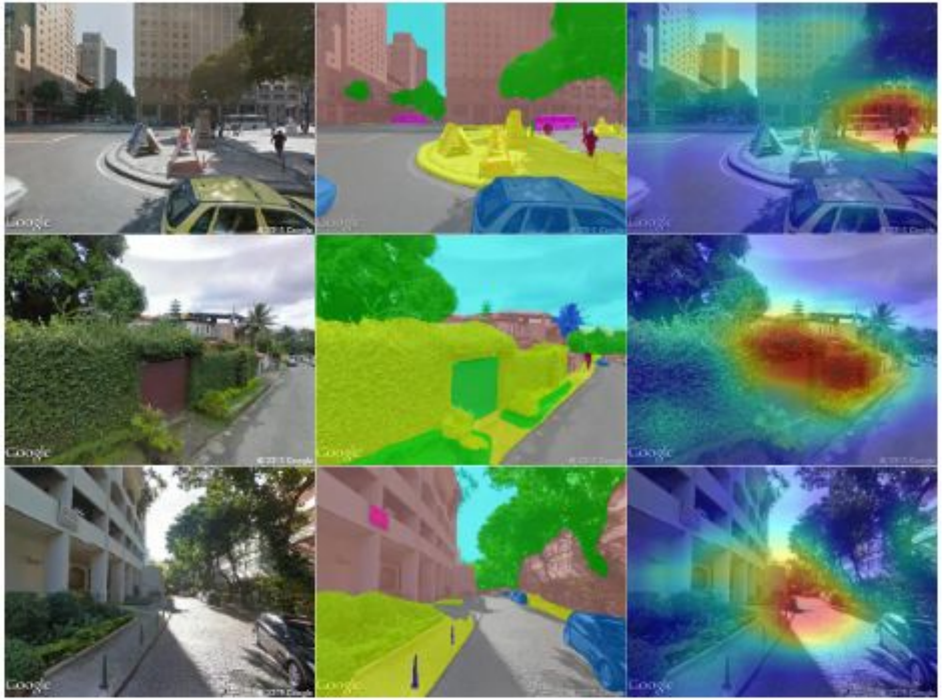
Explanation



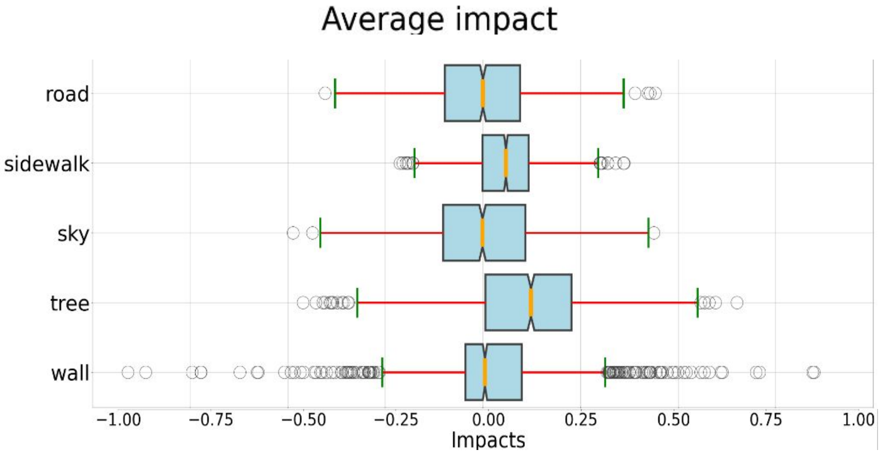
Explanation



Safe samples

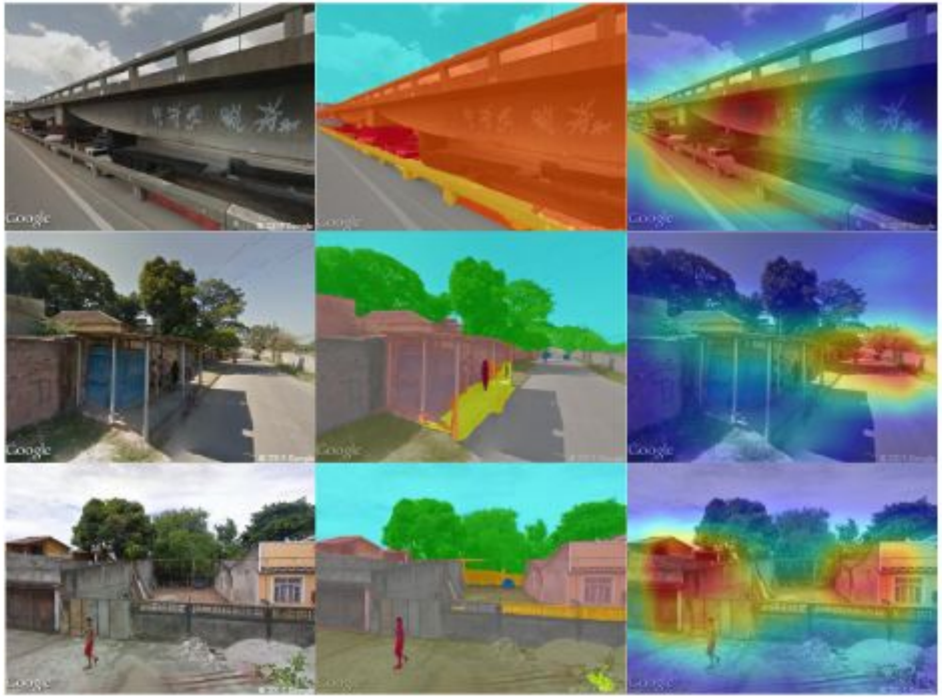


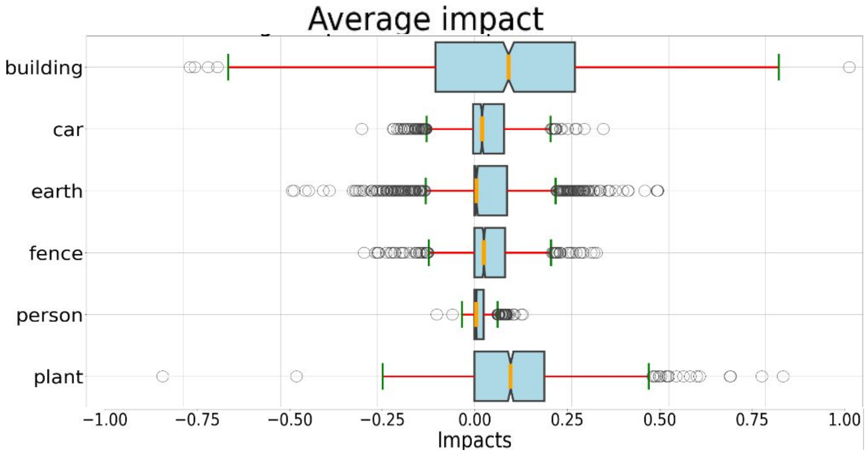
. . .



Unsafe samples

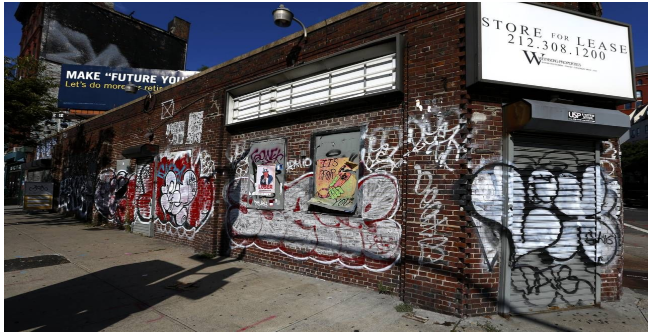
. . .





Limitations

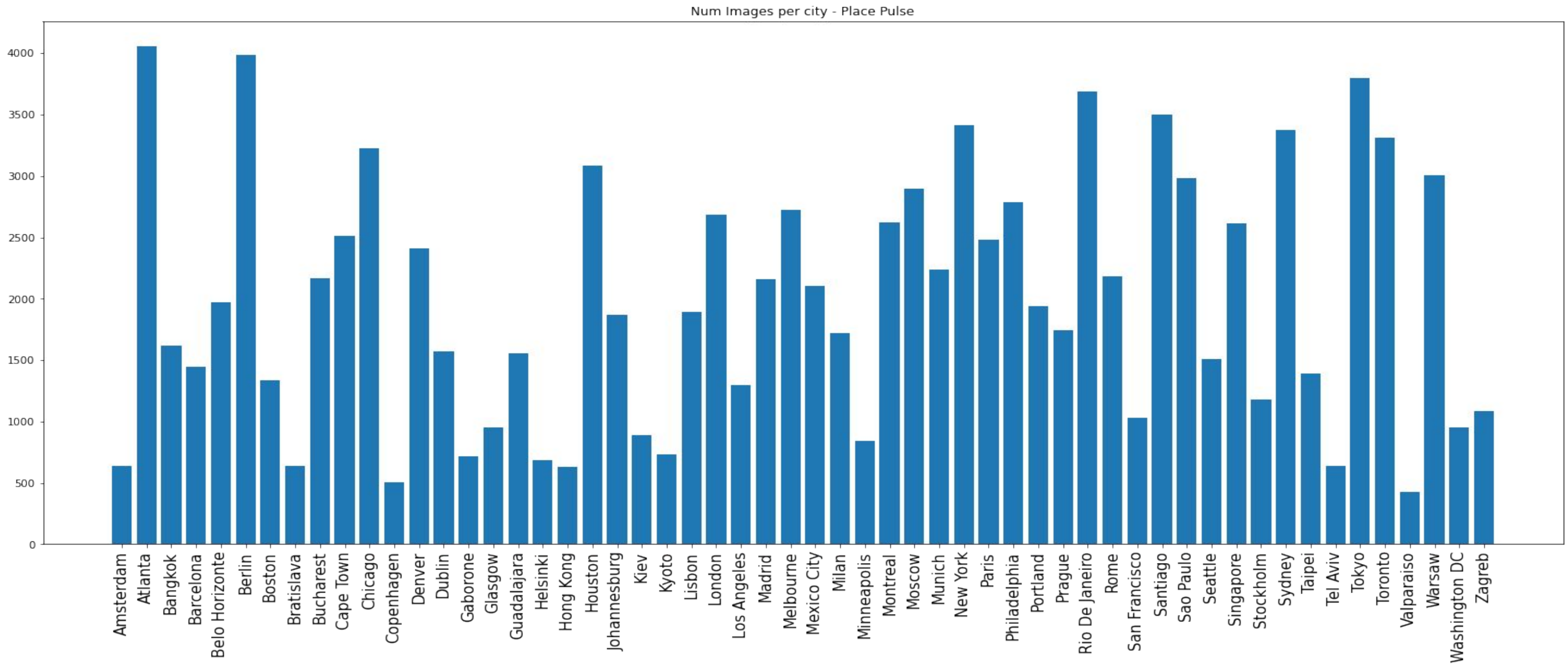
Individual perception

Safe perception	Unsafe perception	
		**New York***
		**Tokyo****

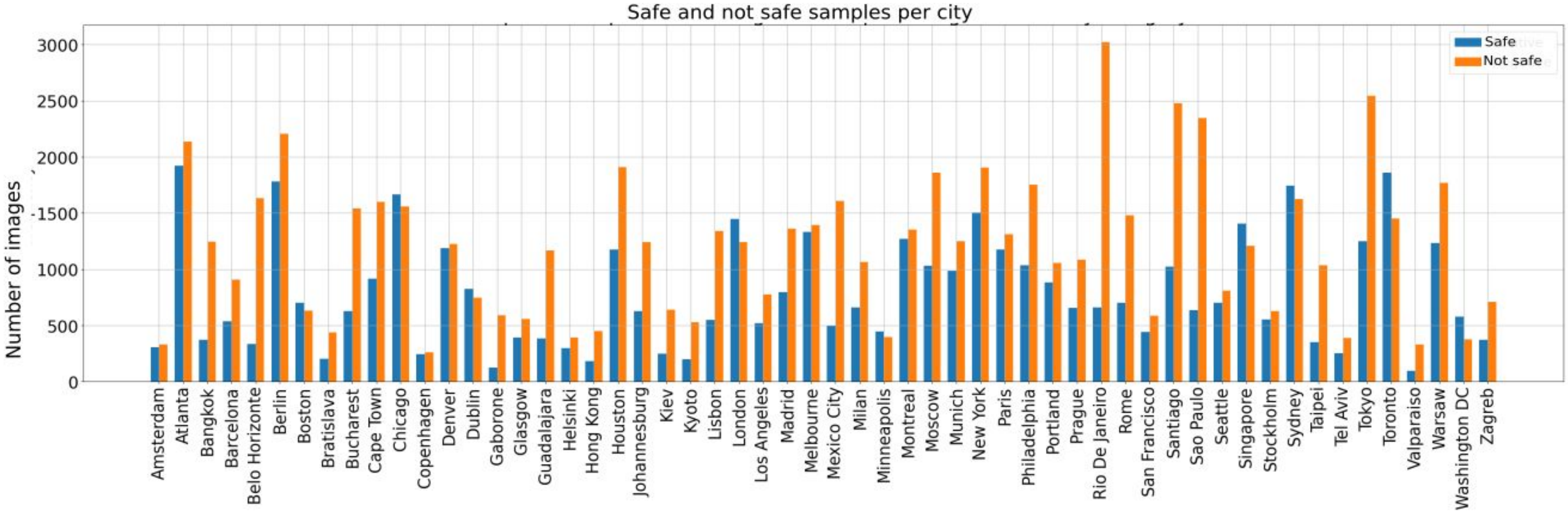
*https://www.nytimes.com/2019/08/08/nyregion/newyorktoday/times-square-panic-safety.html#:~:text=Actually%2C%20Times%20Square%20is%20one,23%2C000%20major%20crimes%20were%20recorded.

**https://www.japantimes.co.jp/news/2019/10/04/national/media-national/rip-off-bars-japan-tourist-boom/

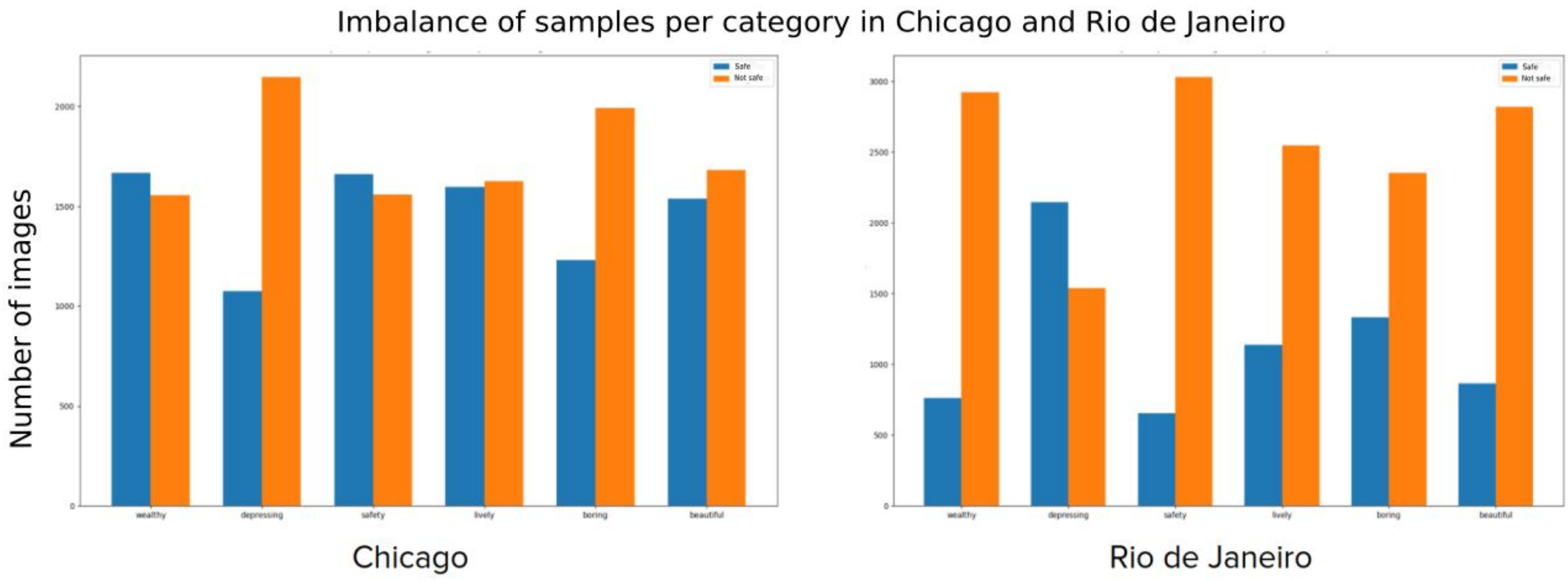
Lack of samples



Imbalance of samples



Imbalance of samples

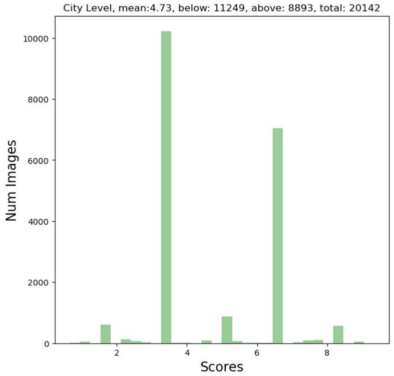


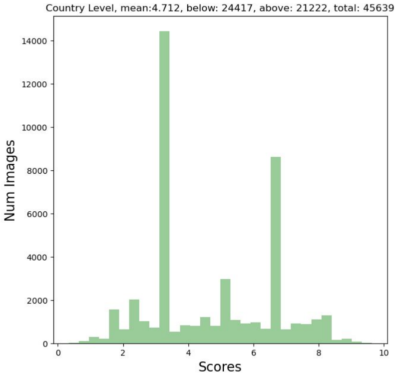
***Positive Samples:** safe, beautiful, wealthy, lively, not depressing, not boring.
***Negative Samples:** not safe, not beautiful, not wealthy, not lively, depressing, boring.

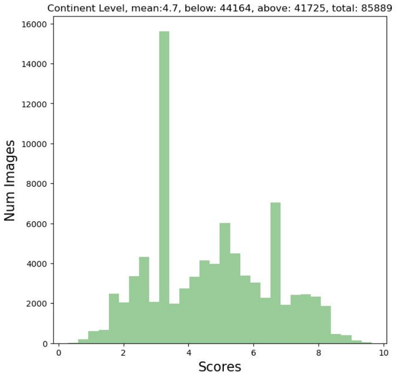
Non-Reliable Score Distribution

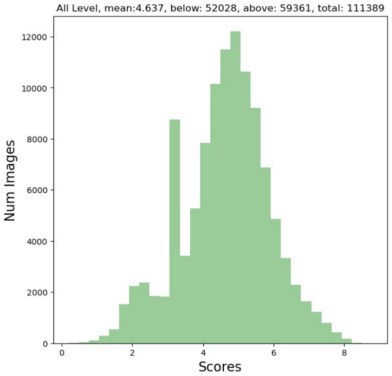
City | Country | Continent | Global

World

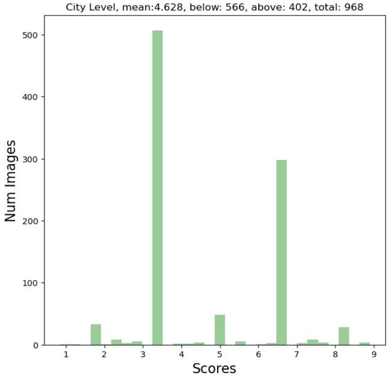


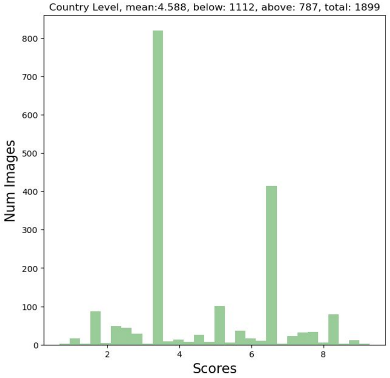


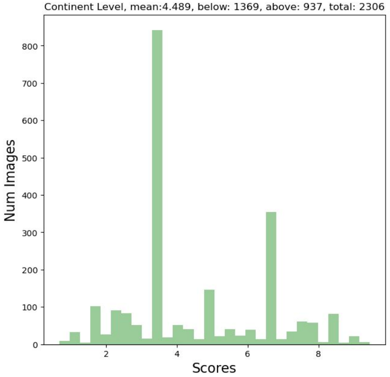


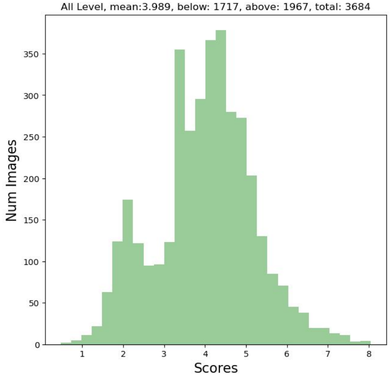


Rio de Janeiro

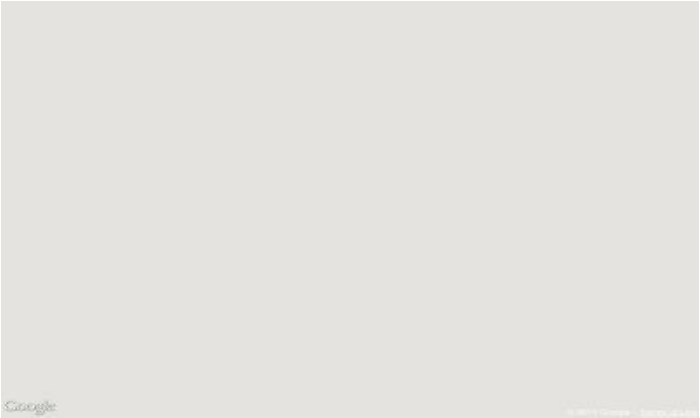


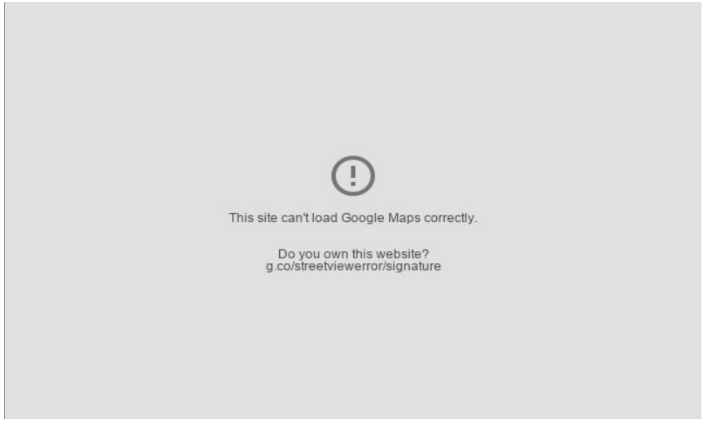






Faulty/Blank/None samples





Conclusions

Conclusions

- We **propose a methodology** to analyze the Place Pulse 2.0 dataset since we thought that is better to focus on data first instead of model complexity.

- We **develop** a new transformer-based model called **UrbanFormer**, aiming to improve street view imagery classification applied to urban safety perception

- We **evaluate** the importance of **visual elements** within images by measuring the intersection over union (IoU) between segmentation masks and model-generated explanations, providing deeper insights into model interpretability and feature relevance.

- We expose the Place Pulse limitations founded through this analysis.

THANKS!

Any Questions?