Benchmarking Slice Discovery Methods for Image Data

Master Thesis Presentation

Rohil Prakash Rao
Matriculation Number: 3299480

First Examiner: PD Dr. Michael Mock

Second Examiner: Prof. Dr. Stefan Wrobel

Supervisor: Sujan Sai Gannamaneni

June 2024

Table of Contents

Table of Contents

What are Slice Discovery Methods (SDMs)?

SDMs are automated techniques for identifying sytematic weaknesses of DNNs. Their goal is to identify slices of data where the DNN-under-test performs poorly such that the samples within the slice belong to a semantically coherent (human-understandable) concept.

What are Slice Discovery Methods (SDMs)?

SDMs are automated techniques for identifying sytematic weaknesses of DNNs. Their goal is to identify slices of data where the DNN-under-test performs poorly such that the samples within the slice belong to a semantically coherent (human-understandable) concept.

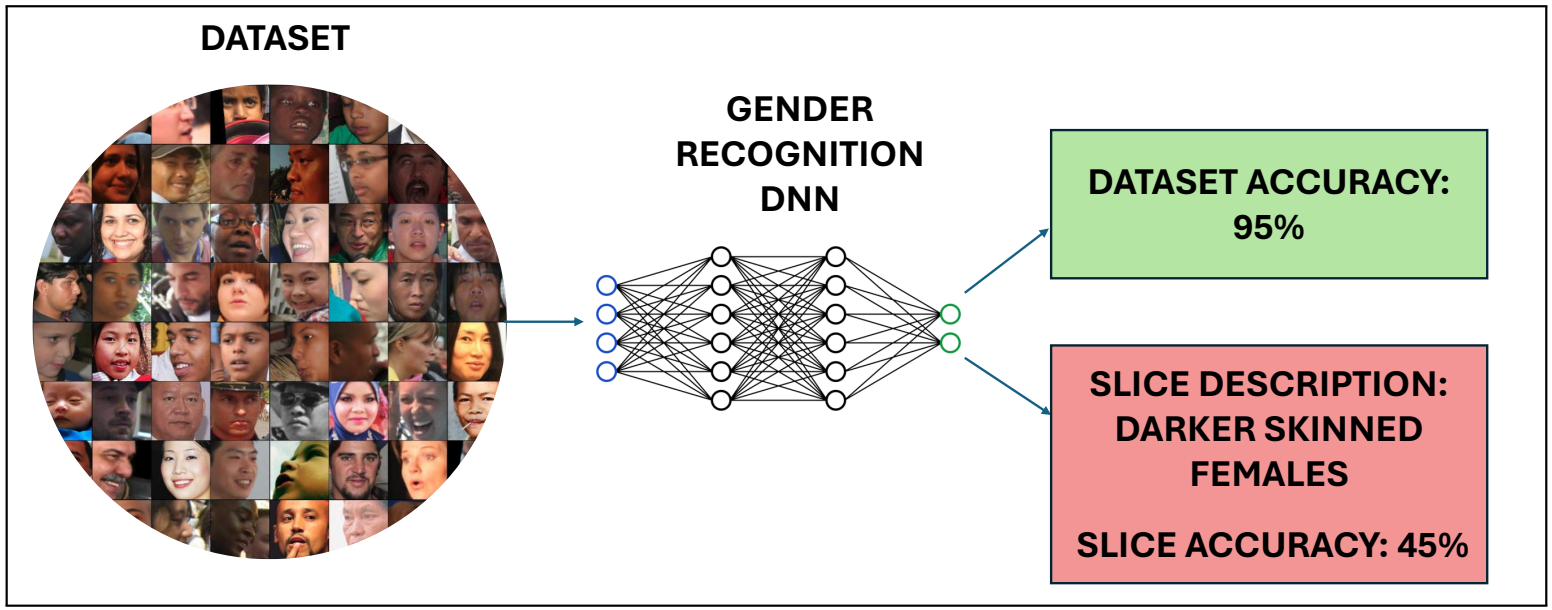


Figure: Example inspired from Boulamwini et al. [1]

What are Slice Discovery Methods (SDMs)?

SDMs are automated techniques for identifying sytematic weaknesses of DNNs. Their goal is to identify slices of data where the DNN-under-test performs poorly such that the samples within the slice belong to a semantically coherent (human-understandable) concept.

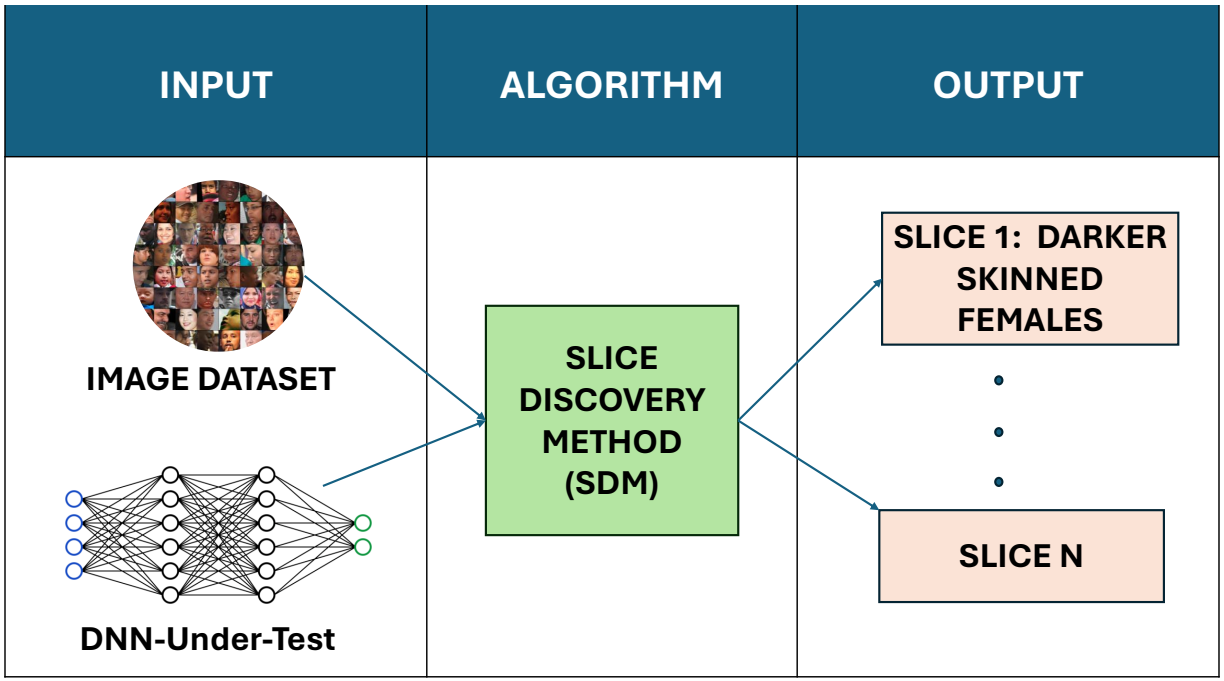
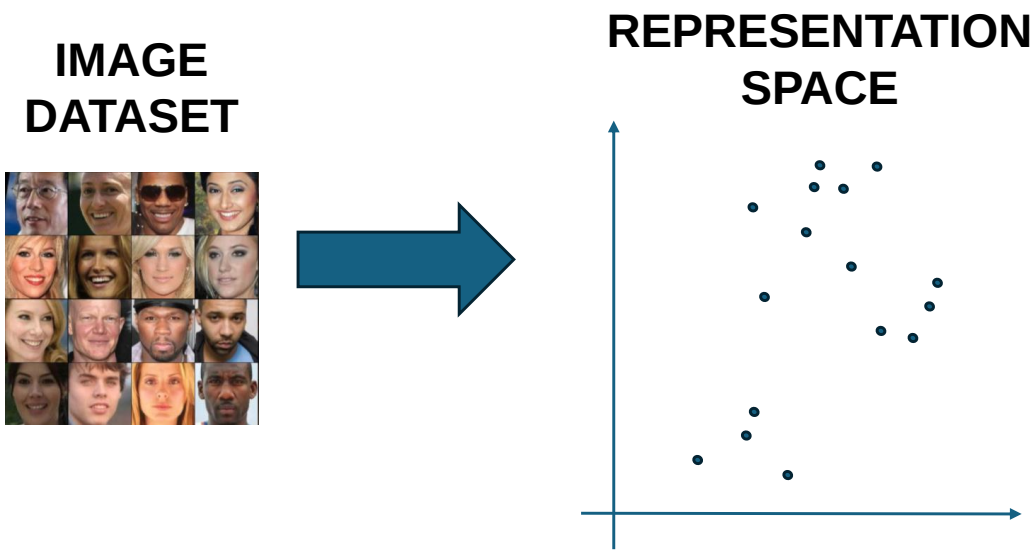


Figure: Workflow for an SDM

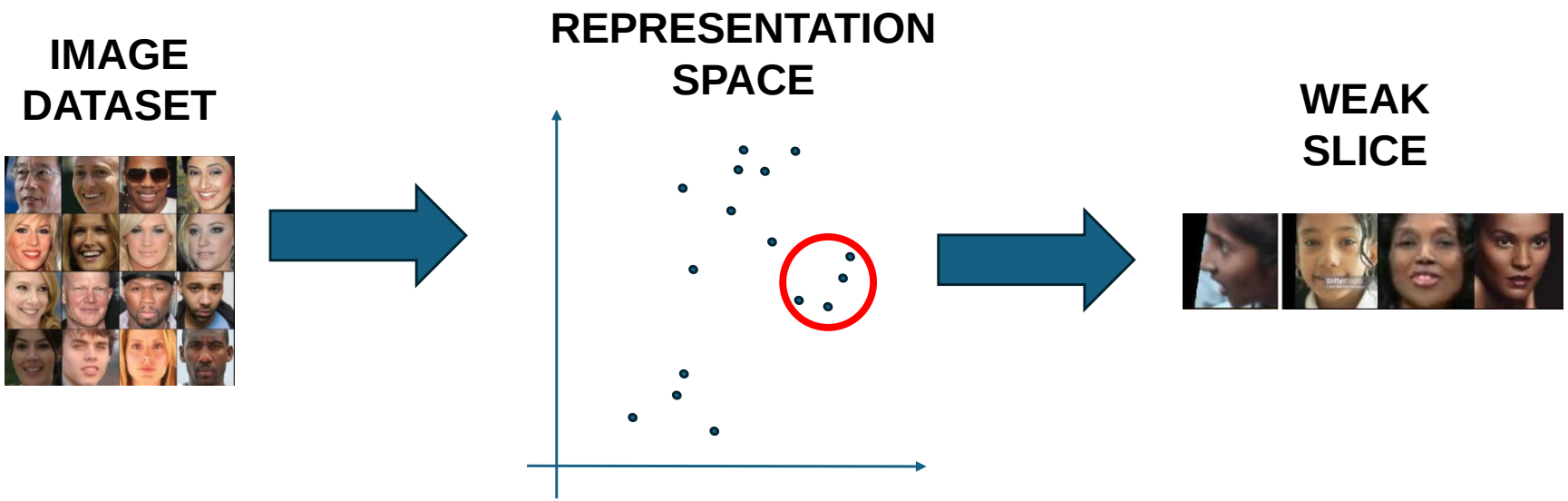
High-Level Steps of SDMs

STEP 1: EMBED



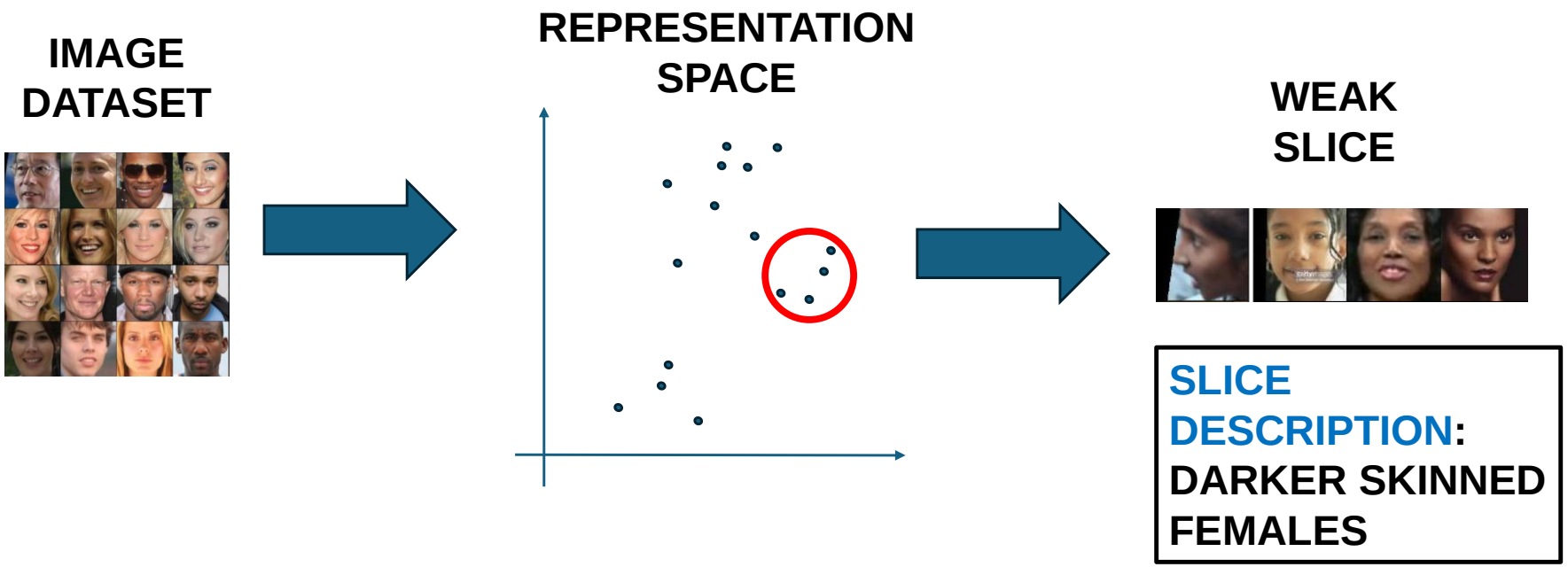
High-Level Steps of SDMs

STEP 2: SLICE



High-Level Steps of SDMs

STEP 3: DESCRIBE



Motivation:

- Recent SDMs show potential for model evaluation and auditing.
- Certifying DNNs for Autonomous Driving can be done with respect to an **Operational Design Domain (ODD)** which defines conditions and constraints under which the DNN performs safely [5].
- Essential to **Benchmark** and compare SDMs for effectiveness, actionability and alignment with a defined ODD.

Motivation:

- Recent SDMs show potential for model evaluation and auditing.
- Certifying DNNs for Autonomous Driving can be done with respect to an **Operational Design Domain (ODD)** which defines conditions and constraints under which the DNN performs safely [5].
- Essential to **Benchmark** and compare SDMs for effectiveness, actionability and alignment with a defined ODD.

Thesis Objective:

Develop a **benchmarking framework for SDMs** on image data, and evaluate SDMs on:

1. SOTA **image classification** models.
2. **Pedestrian detection** on autonomous driving datasets.

Research Questions

- **Research Question 1 (RQ1)**: Does the Slice Discovery Method (SDM) identify slices where the DNN-under-test exhibits performance degradation such that the identified slices also align with a defined Operational Design Domain (ODD) concept?

Research Questions

- **Research Question 1 (RQ1)**: Does the Slice Discovery Method (SDM) identify slices where the DNN-under-test exhibits performance degradation such that the identified slices also align with a defined Operational Design Domain (ODD) concept?

- **Research Question 2 (RQ2):** Does the Slice Discovery Method (SDM) identify slices belonging to novel concepts that go beyond the coverage of the defined Operational Design Domain (ODD) space?

Table of Contents

SDM 1: Spotlight

- **Key Idea:** Similar input data should have similar model representations [2].
- Looks for high-loss contiguous regions in the latent space
- Uses a kernel function that assigns weights based on proximity, and optimizes to maximize the weighted average loss in these regions

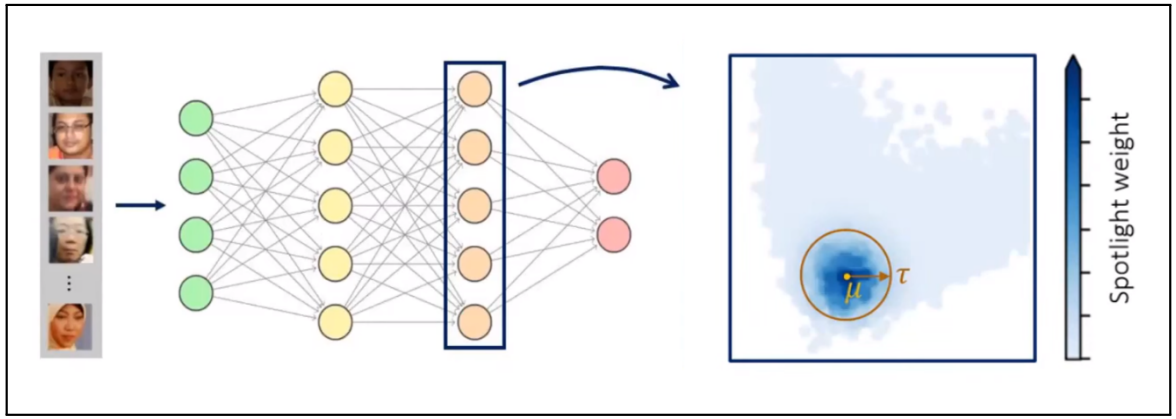


Figure: Figure taken from original paper [2]

SDM 2: Domino

1. **Embed:**
   - Images are embedded in cross-modal (image-text) space for e.g. CLIP [6]
2. **Slice:**
   - Error-Aware Gaussian Mixture Model
   - Jointly models the input embeddings, true labels and model predictions.
3. **Describe:**
   - Utilizes the multi-modal representation (for e.g. CLIP) to identify relevant phrases from a predefined corpus of descriptions

SDM 3: SVM Failure Directions

Identifies challenging sub-populations by fitting a per-class SVM model to find the hyperplane that can best separate the correct from the incorrect predictions of the DNN-under-test [4].

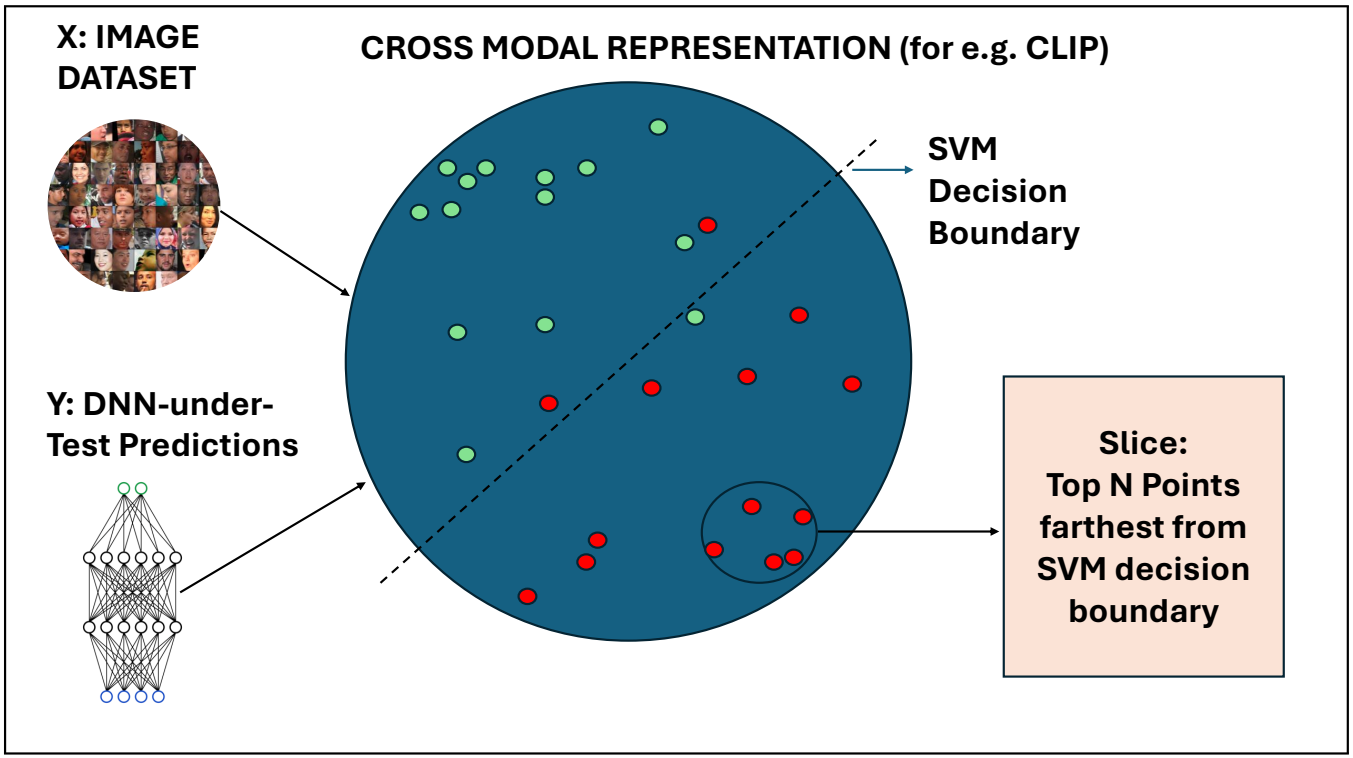


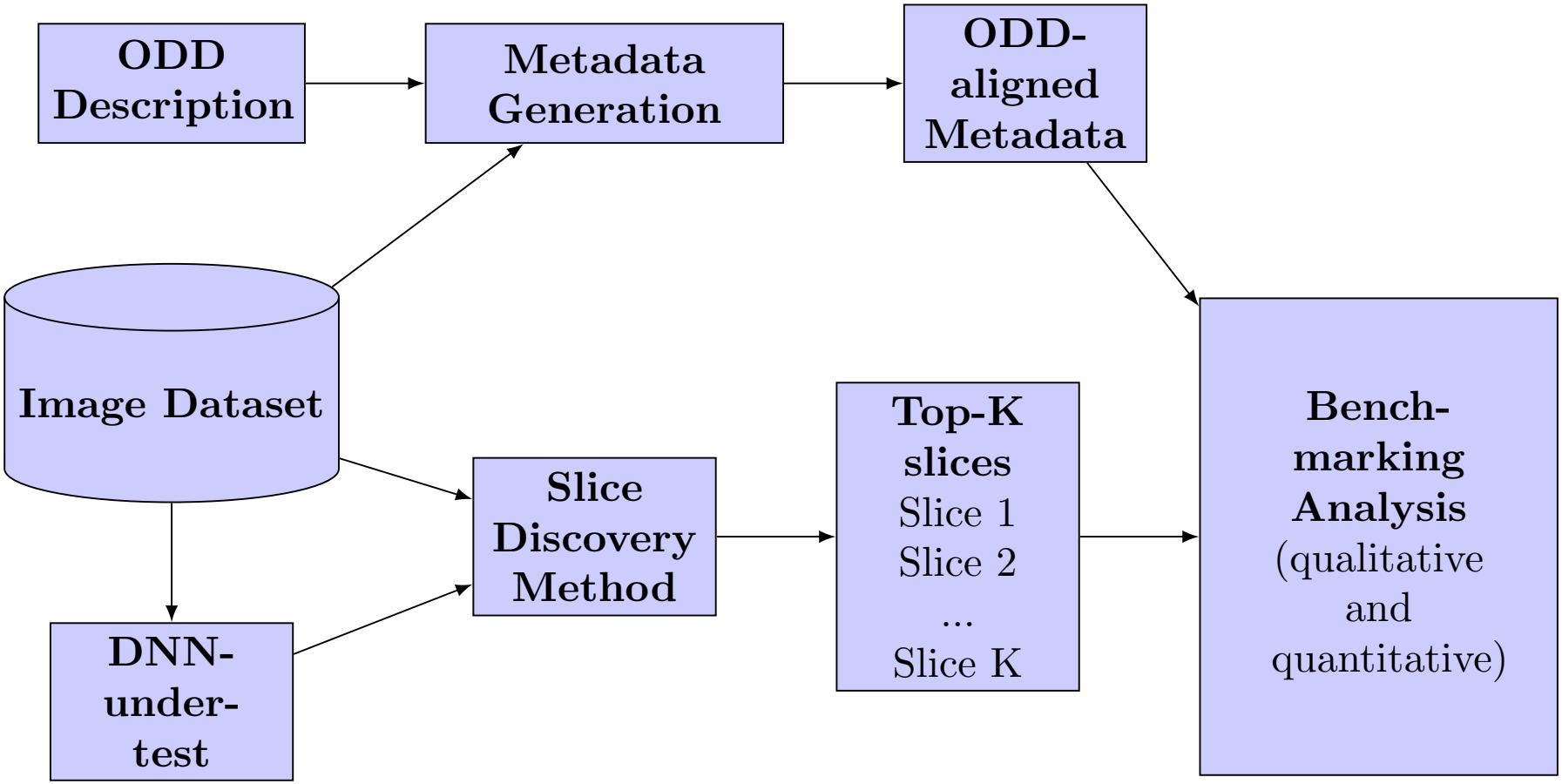
Figure: SVM Failure Directions Overview

SDM	Image Embedding Space	Slicing Algorithm	Slice Descriptions
Spotlight [2]	DNN-under-test	Kernel Function	
Domino [3]	Cross-Modal (e.g. CLIP)	Gaussian Mixture Model	✓
SVM Failure Directions [4]	Cross-Modal (e.g. CLIP)	Support Vector Machine	✓

Table: SDMs chosen to be benchmarked for the thesis

Table of Contents

Benchmarking overview

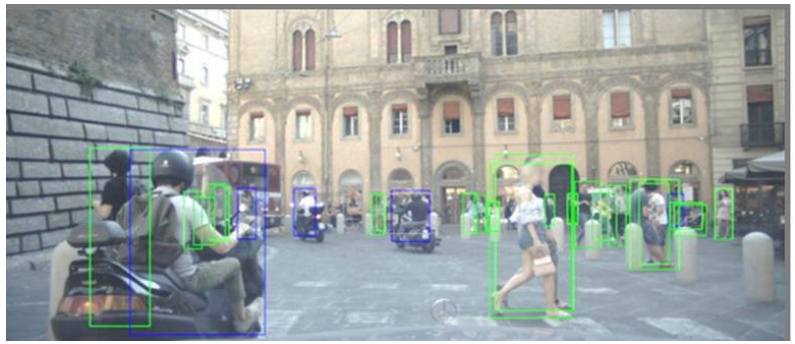


Datasets

CelebA

RailSem19 (RS19)

ImageNet21K-P



EuroCity Persons (ECP)

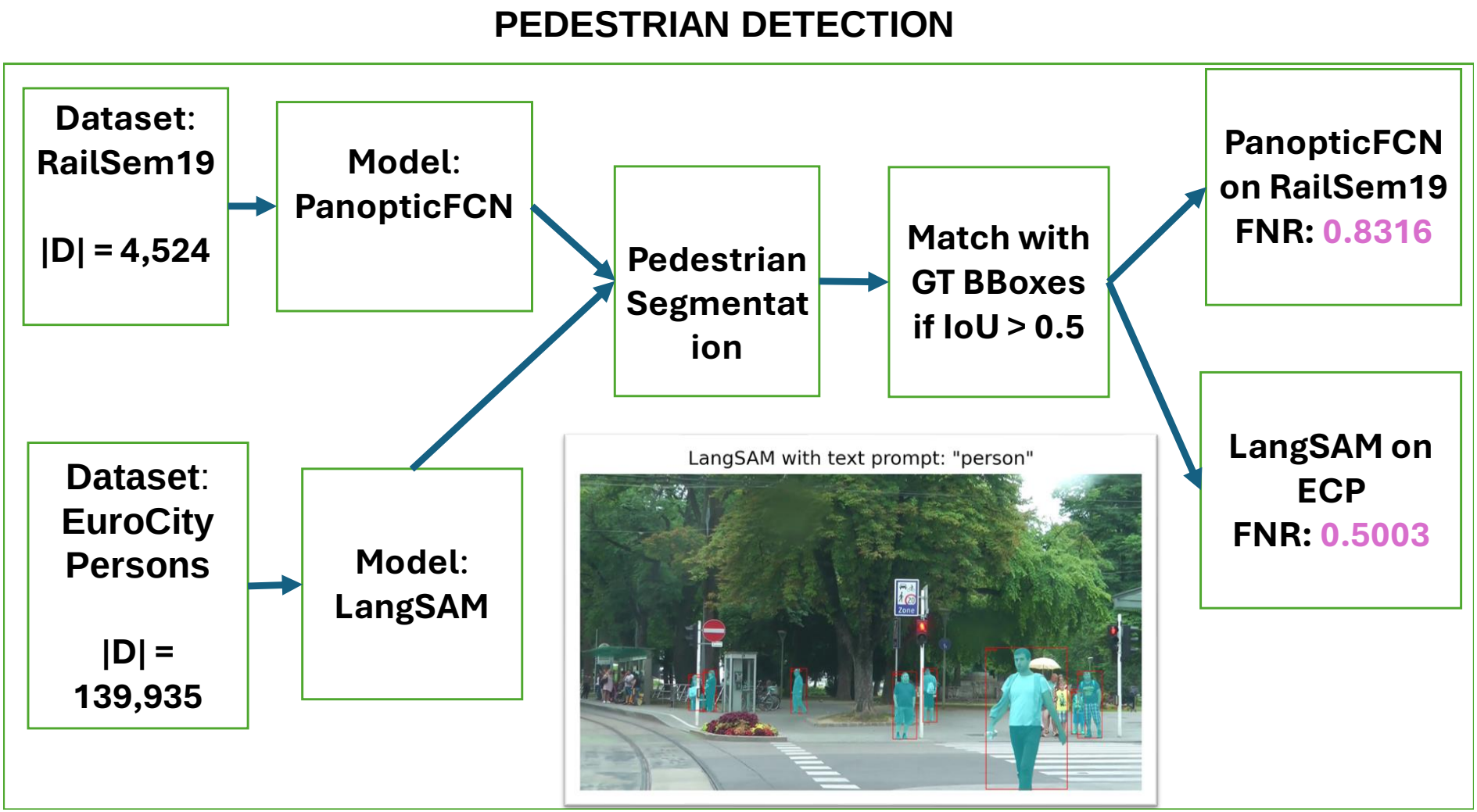
Evaluated Tasks

Image Classification

- **Task:** Classify image as ‘Person’ or ‘Not Person’
- **Dataset:** CelebA (‘Person’ images), ImageNet21K-S (’Not Person’ images)
- **Model:** ViT-B/16 [7] pre-trained on ImageNet21K-P
- **Accuracy:** 0.97 on both ‘Person’ and ‘Not Person’ classes at level-0

Figure: Inference example [7]

Evaluated Tasks



Slice Coherence Metrics

For a dataset D and a slice $S \subseteq D$, each image has attributes $a_1, a_2, \ldots, a_n$ (true or false). Let S_{a_i} be the subset of S where a_i is true.

Slice Attribute Proportion (SAP): The proportion of the slice S characterized by attribute a_i.

$$SAP = |S_{a_i}|/|S|$$

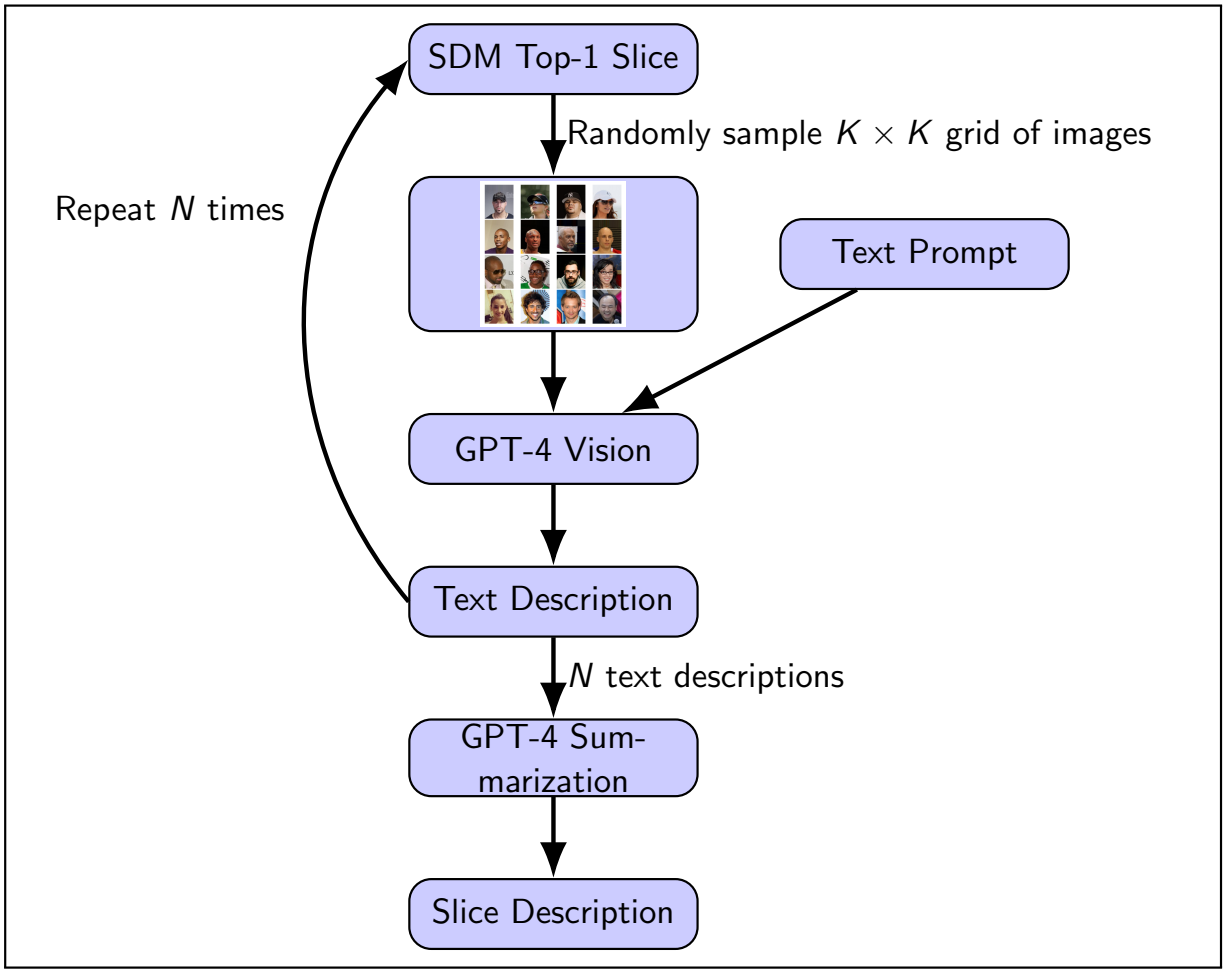
Slice Attribute Coverage (SAC): The proportion of the dataset attribute a_i contained within the slice S.

$$SAC = |S_{a_i}|/|D_{a_i}|$$

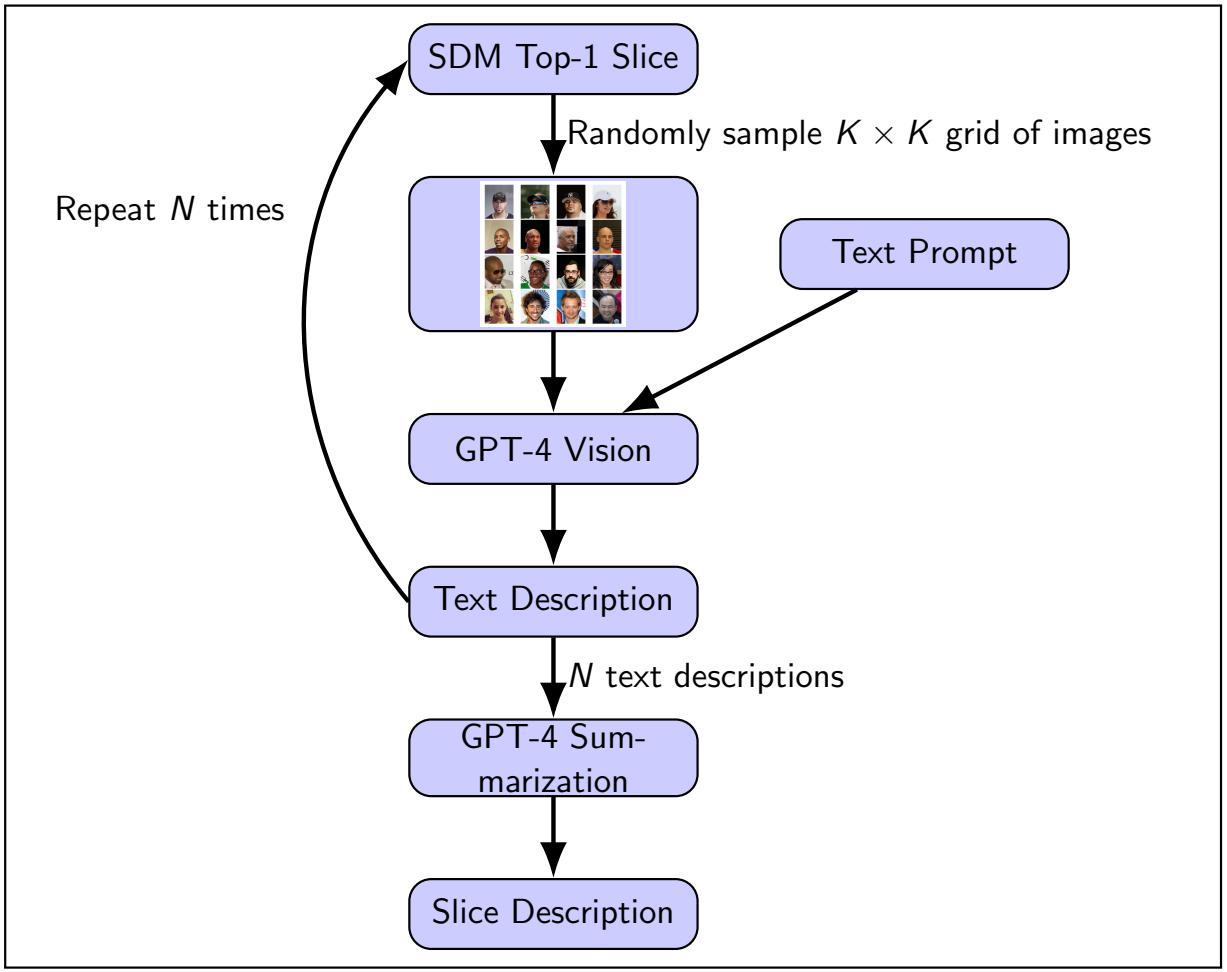
Slice to Dataset Attribute Ratio of Proportions (SDARP): The ratio of the attribute proportion in the slice to the attribute proportion in the dataset.

$$SDARP = \frac{|S_{a_i}|/|S|}{|D_{a_i}|/|D|}$$

Improving Slice Descriptions using GPT-4 Vision



Improving Slice Descriptions using GPT-4 Vision



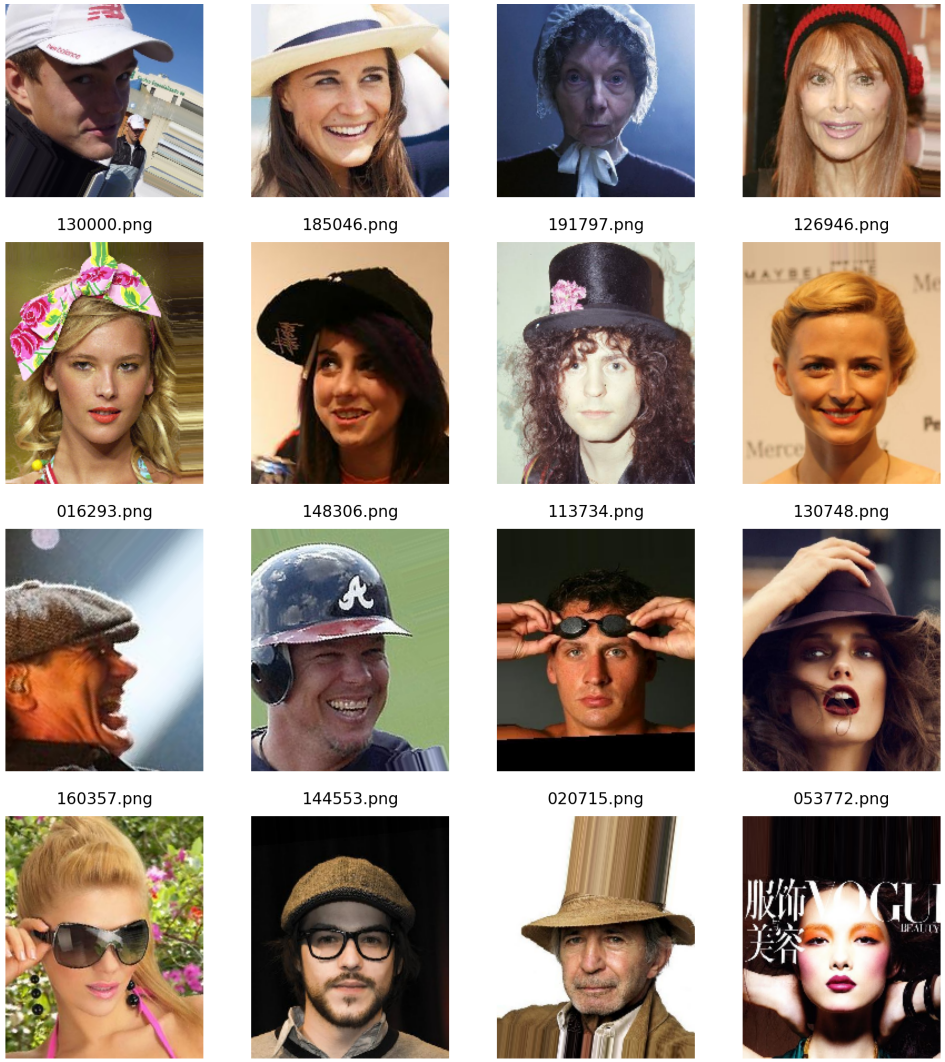
- We use: N, K = 10
- Two types of text prompts:
  1. Obtain Natural Language Descriptions for Slice
  2. Obtain Attribute-Value Descriptions for Slice

Table of Contents

Domino's top-1 Slice for CelebA

Domino Slice Descriptions

- a photo of an undercover person
- tony blair photo of a person
- a photo of a disguised person



Domino's top-1 Slice for CelebA

Attribute	**Accuracy**
Wearing_Hat: True	0.5566
Pale_Skin: True	0.9479
Eyeglasses: True	0.9613
High_Cheekbones: False	0.9627
Smiling: False	0.9630

Table: Lowest performing attributes based on the accuracy of ViT-B/16 on CelebA

Coherence Metrics for CelebA

Attribute	$\lvert S_a \rvert$	$\lvert D_a \rvert$	SAP $\frac{\lvert S_a \rvert}{\lvert S \rvert}$	DAP $\frac{\lvert D_a \rvert}{\lvert D \rvert}$	SAC $\frac{\lvert S_a \rvert}{\lvert D_a \rvert}$	SDARP $\frac{SAP}{DAP}$
Domino						
Wearing_Hat: True	2628	5801	0.5996	0.0334	0.453	17.9471
Pale_Skin: True	302	6281	0.0689	0.0362	0.0481	1.9048
Eyeglasses: True	452	11378	0.1031	0.0655	0.0397	1.5738
SVM Failure Directions (SVM FD)						
Wearing_Hat: True	2330	5801	0.6709	0.0334	0.4017	20.0812
Pale_Skin: True	245	6281	0.0705	0.0362	0.039	1.9502
Eyeglasses: True	409	11378	0.1178	0.0655	0.0359	1.7972
Spotlight						
Wearing_Hat: True	2829	5801	0.8083	0.0334	0.4877	24.1938
Pale_Skin: True	206	6281	0.0589	0.0362	0.0328	1.6271
High_Cheekbones: False	2534	90485	0.724	0.5211	0.028	1.3893

Table: Coherence Metrics for the top-3 attributes (sorted by SDARP) of the top-1 slice identified by different methods on CelebA.

Coherence Metrics for CelebA

Attribute	$\lvert S_a \rvert$	$\lvert D_a \rvert$	SAP $\frac{\lvert S_a \rvert}{\lvert S \rvert}$	DAP $\frac{\lvert D_a \rvert}{\lvert D \rvert}$	SAC $\frac{\lvert S_a \rvert}{\lvert D_a \rvert}$	SDARP $\frac{SAP}{DAP}$
Domino						
Wearing_Hat: True	2628	5801	0.5996	0.0334	0.453	17.9471
Pale_Skin: True	302	6281	0.0689	0.0362	0.0481	1.9048
Eyeglasses: True	452	11378	0.1031	0.0655	0.0397	1.5738
SVM Failure Directions (SVM FD)						
Wearing_Hat: True	2330	5801	0.6709	0.0334	0.4017	20.0812
Pale_Skin: True	245	6281	0.0705	0.0362	0.039	1.9502
Eyeglasses: True	409	11378	0.1178	0.0655	0.0359	1.7972
Spotlight						
Wearing_Hat: True	2829	5801	0.8083	0.0334	0.4877	24.1938
Pale_Skin: True	206	6281	0.0589	0.0362	0.0328	1.6271
High_Cheekbones: False	2534	90485	0.724	0.5211	0.028	1.3893

Table: Coherence Metrics for the top-3 attributes (sorted by SDARP) of the top-1 slice identified by different methods on CelebA.

Domino's top-1 Slice for ImageNet21K-S

Domino Slice Descriptions

- a photo of a church person
- a photo of an altar person
- a photo of a monastic person

Coherence Metrics for ImageNet21K-S

Class name	Accuracy
bullfighting	0.031
boxing	0.2392
pontifical	0.3707
little_theater	0.8324
madras	0.8508

Table: Lowest performing classes based on the accuracy of ViT-B/16 on ImageNet21K-S

Coherence Metrics for ImageNet21K-S

Attribute	$\lvert S_a \rvert$	$\lvert D_a \rvert$	SAP $\frac{\lvert S_a \rvert}{\lvert S \rvert}$	DAP $\frac{\lvert D_a \rvert}{\lvert D \rvert}$	SAC $\frac{\lvert S_a \rvert}{\lvert D_a \rvert}$	SDARP $\frac{SAP}{DAP}$
Domino						
pontifical: True	317	464	0.1809	0.0025	0.6832	71.0092
bullfighting: True	131	870	0.0748	0.0048	0.1506	15.6504
madras: True	124	838	0.0708	0.0046	0.1480	15.3798
SVM Failure Directions (SVM FD)						
boxing: True	701	1292	0.3850	0.0071	0.5426	54.2567
bullfighting: True	432	870	0.2372	0.0048	0.4966	49.6549
pontifical: True	123	464	0.0675	0.0025	0.2651	26.5085
Spotlight						
bullfighting: True	862	870	0.2362	0.0048	0.9908	49.4314
boxing: True	1263	1292	0.3460	0.0071	0.9776	48.7703
pontifical: True	327	464	0.0896	0.0025	0.7047	35.1596

Table: Coherence Metrics for the top-3 attributes of the top-1 slice identified by different methods on ImageNet21K-S.

Coherence Metrics for ImageNet21K-S

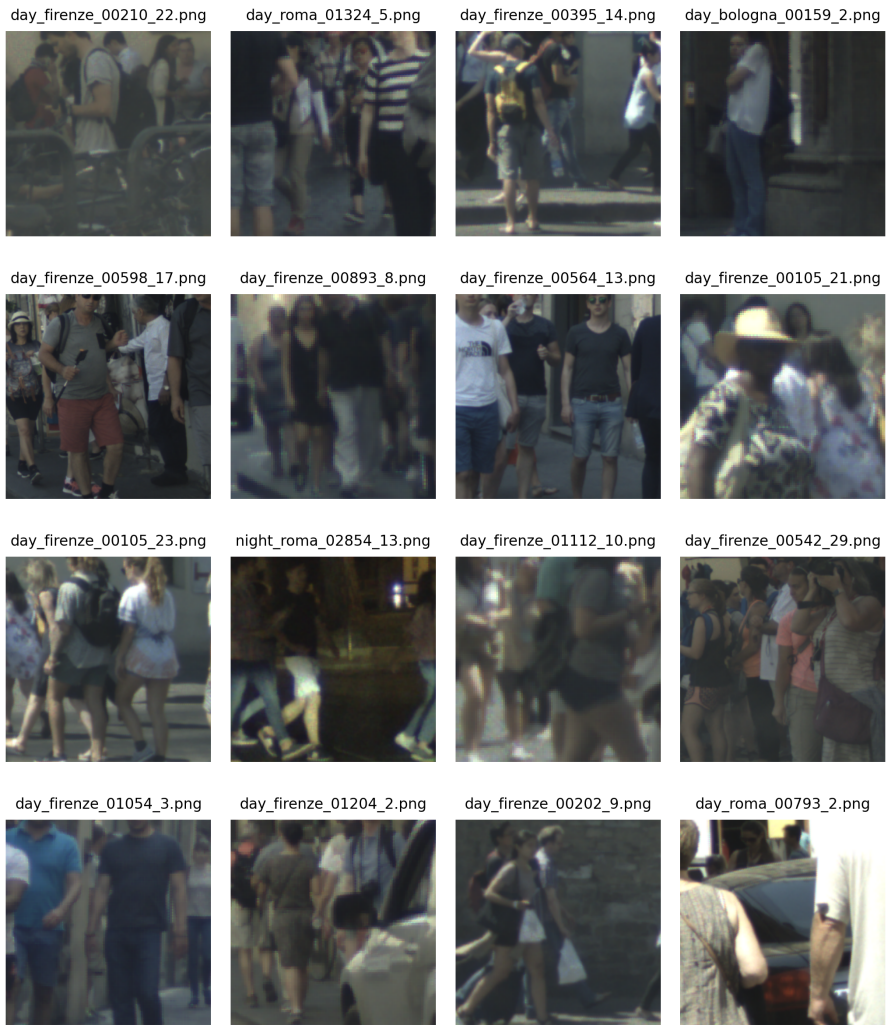
Attribute	$\lvert S_a \rvert$	$\lvert D_a \rvert$	SAP $\frac{\lvert S_a \rvert}{\lvert S \rvert}$	DAP $\frac{\lvert D_a \rvert}{\lvert D \rvert}$	SAC $\frac{\lvert S_a \rvert}{\lvert D_a \rvert}$	SDARP $\frac{SAP}{DAP}$
Domino						
pontifical: True	317	464	0.1809	0.0025	0.6832	71.0092
bullfighting: True	131	870	0.0748	0.0048	0.1506	15.6504
madras: True	124	838	0.0708	0.0046	0.1480	15.3798
SVM Failure Directions (SVM FD)						
boxing: True	701	1292	0.3850	0.0071	0.5426	54.2567
bullfighting: True	432	870	0.2372	0.0048	0.4966	49.6549
pontifical: True	123	464	0.0675	0.0025	0.2651	26.5085
Spotlight						
bullfighting: True	862	870	0.2362	0.0048	0.9908	49.4314
boxing: True	1263	1292	0.3460	0.0071	0.9776	48.7703
pontifical: True	327	464	0.0896	0.0025	0.7047	35.1596

Table: Coherence Metrics for the top-3 attributes of the top-1 slice identified by different methods on ImageNet21K-S.

Domino's top-1 Slice for ECP

Domino Slice Descriptions

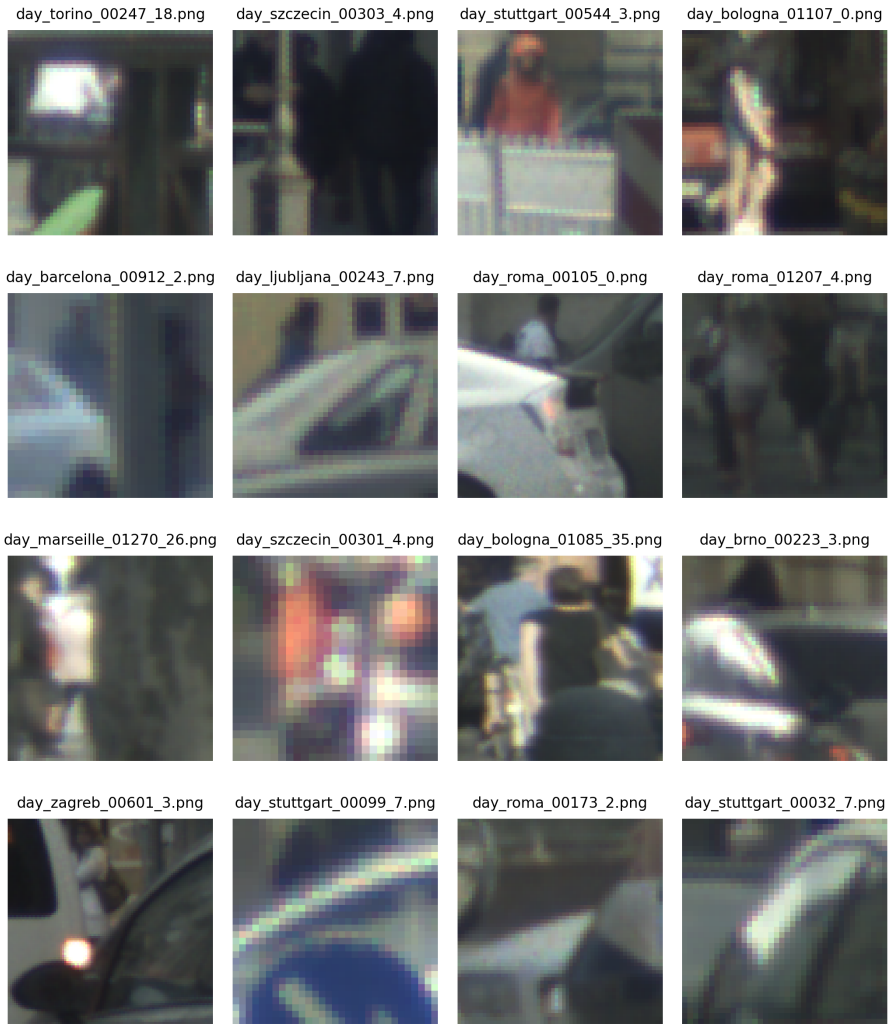
- a occluded photo of a person with a group of people in the day.
- a photo of a person in the day.



SVM FD's top-1 Slice for ECP

SVM FD Slice Descriptions

- a photo of a middle-aged person in sunny weather in the day.
- a photo of a middle-aged person in sunny weather.



Coherence Metrics for EuroCity Persons

Attribute	**Accuracy**
occlusion_gt: 80	0.0342
pose: facing sideways	0.1417
occlusion_gt: 40	0.1632
truncated: 80	0.1786
pose: facing forwards	0.2260
in-group: True	0.2811
objects: Car	0.3097
hair-color: gray	0.3192
construction-worker: True	0.3395
city: lyon	0.3476

Table: Lowest performing attributes based on the accuracy of LangSAM on ECP

Coherence Metrics for EuroCity Persons

Attribute	$\lvert S_a \rvert$	$\lvert D_a \rvert$	SAP $\frac{\lvert S_a \rvert}{\lvert S \rvert}$	DAP $\frac{\lvert D_a \rvert}{\lvert D \rvert}$	SAC $\frac{\lvert S_a \rvert}{\lvert D_a \rvert}$	SDARP $\frac{SAP}{DAP}$
Domino						
in-group: True	241	1594	0.1664	0.0114	0.1512	14.6112
city: Firenze	684	8218	0.4724	0.0587	0.0832	8.0435
city: Roma	577	13036	0.3985	0.0932	0.0443	4.2775
occlusion_gt: 80	267	7185	0.1844	0.0513	0.0372	3.5912
shirt-color: bright	114	3321	0.0787	0.0237	0.0343	3.3174
SVM Failure Directions (SVM FD)						
occlusion_gt: 80	239	7185	0.1707	0.0513	0.0333	3.3248
hair-color: gray	242	7540	0.1729	0.0539	0.0321	3.2081
brightness: medium	419	20295	0.2993	0.1450	0.0206	2.0636
occlusion_gt: 40	654	33112	0.4671	0.2366	0.0198	1.9742
pose: facing-forwards	937	47827	0.6693	0.3418	0.0196	1.9582

Table: Coherence Metrics for the top-3 attributes of the top-1 slice identified by different methods on ECP.

Coherence Metrics for EuroCity Persons

Attribute	$\lvert S_a \rvert$	$\lvert D_a \rvert$	SAP $\frac{\lvert S_a \rvert}{\lvert S \rvert}$	DAP $\frac{\lvert D_a \rvert}{\lvert D \rvert}$	SAC $\frac{\lvert S_a \rvert}{\lvert D_a \rvert}$	SDARP $\frac{SAP}{DAP}$
Domino						
in-group: True	241	1594	0.1664	0.0114	0.1512	14.6112
city: Firenze	684	8218	0.4724	0.0587	0.0832	8.0435
city: Roma	577	13036	0.3985	0.0932	0.0443	4.2775
occlusion_gt: 80	267	7185	0.1844	0.0513	0.0372	3.5912
shirt-color: bright	114	3321	0.0787	0.0237	0.0343	3.3174
SVM Failure Directions (SVM FD)						
occlusion_gt: 80	239	7185	0.1707	0.0513	0.0333	3.3248
hair-color: gray	242	7540	0.1729	0.0539	0.0321	3.2081
brightness: medium	419	20295	0.2993	0.1450	0.0206	2.0636
occlusion_gt: 40	654	33112	0.4671	0.2366	0.0198	1.9742
pose: facing-forwards	937	47827	0.6693	0.3418	0.0196	1.9582

Table: Coherence Metrics for the top-3 attributes of the top-1 slice identified by different methods on ECP.

GPT Descriptions: Natural Language

SDM	Top 3 ODD	SDM Top-1 Description	GPT-4V Summary Description
Domino	pontifical, bullfighting, madras	A photo of a church person	The images collectively showcase diverse human activities and professions, reflecting traditions, cultural practices, and unexpected role portrayals from around the world, with a focus on Spain.
SVM FD	boxing, bullfighting, pontifical	A photo of a black instrumentality with a person	The images depict diverse scenarios of physical confrontation and competition, ranging from bullfighting to boxing, illustrating human interactions and confrontational activities across cultural events and sports settings.
Spotlight	bullfighting, boxing, pontifical	N/A	These images capture the theme of sports, competitions, and cultural activities across various regions, emphasizing physical and performative events.

Table: Natural Language Descriptions for the top-1 slice for ImageNet-21K-S

GPT Descriptions: Attribute-Value

SDM	Summary Description (Within ODD Attributes)	Summary Description (Outside ODD Attributes)
Domino	Class Bullfighting: 8/10 Class Formalwear: 8/10 Class Classroom: 7/10	Class Human: Present Class Social Event: Present Class Specialized Clothing: Varied
SVM FD	Class Bull: 4 Class Bullfighting: 4 Class Boxing: 3	Attr: Sports, Value: Boxing, Bullfighting, Billiards Attr: Human Activities, Value: True Attr: Class Sport, Value: Sport-related Activities
Spotlight	Class Bullfighting: Yes Class Boxing: Yes Class Formalwear: Yes	Attribute: Class Sport, Value: Yes Attribute: Class Performance, Value: High Attribute: Class Human, Value: High

Table: Attribute-Values Descriptions for the top-1 slice for **ImageNet21K-S**

Table of Contents

Insights on Slice Error metrics

- **Slice size:** minimum 1% and maximum 2.5% of the dataset
- **Slice Failure Rate**: SDMs consistently identify high-error slices. Minimum Slice Failure Rate: 0.78

	CelebA	IN21K-S	ECP	RS19
Domino	-0.9362	-0.9639	-0.4991	-0.1684
SVM FD	-0.7610	-0.8305	-0.3418	-0.1467
Spotlight	-0.8121	-0.7946		

Table: Comparison of performance degradation of SDMs across different datasets

Insights on Slice Coherence

	CelebA	IN21K-S	ECP	RS19
Domino	Wearing Hat: True 0.5996	pontifical: True 0.1809	In-group: True 0.1664	pose: forwards **0.6071**
SVM FD	Wearing Hat: True 0.6709	boxing: True **0.3850**	occlusion gt: 80 **0.1707**	brightness: high 0.1304
Spotlight	Wearing Hat: True **0.8083**	bullfighting: True 0.2362		

Table: Slice Attribute Proportion of Top-1 Attributes identified across experiments

- Non-trivial to draw conclusion for Pedestrian Detection datasets due to noisy metadata

Answers to Research Questions

RQ1: Does the SDM identify top-1 slices that are underperforming, and align with a concept in the defined ODD?

- SDMs consistently identify top-1 slices with DNN performance degradation.
- **Visual inspection** and **Slice Descriptions** suggest coherence in some cases but it is subjective and may not suffice for safety-critical applications.
- **ODD-based analysis** showed that in some cases the top-1 slices did not have a single dominant concept.

Answers to Research Questions

RQ2: Does the SDM identify slices that detect novel concepts beyond the coverage of the ODD?

- SDMs identify top-1 slices with multiple dominant ODD attributes.
- Analysis did not reveal a conclusive higher-level encompassing concept.
- GPT-4 Vision shows promise in identifying novel concepts beyond defined ODD. Requires further evaluations.

Conclusion

Contributions:

- Proposed a benchmarking framework
- Evaluated three SDMs on four real-world datasets
- Two tasks: Image Classification and Object Detection
- Proposed metrics to measure slice coherence
- Evaluation of GPT-4 Vision for enhancing slice descriptions

Future Work:

- Novel slice description and visualization techniques
- Coherence metrics for reliable slice discovery

References I

[1] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[2] Greg d'Eon, Jason d'Eon, James R. Wright, and Kevin Leyton-Brown. The spotlight: A general method for discovering systematic errors in deep learning models. *CoRR*, abs/2107.00758, 2021. URL `https://arxiv.org/abs/2107.00758`.

[3] Sabri Eyuboglu, Bojan Karlaš, Christopher Ré, Ce Zhang, and James Zou. dcbench: A benchmark for data-centric ai systems. In *Proceedings of the Sixth Workshop on Data Management for End-To-End Machine Learning*, pages 1–4, 2022.

[4] Saachi Jain, Hannah Lawrence, Ankur Moitra, and Aleksander Madry. Distilling model failures as directions in latent space, 2022.

References II

[5] Philip Koopman and Frank Fratrik. How many operational design domains, objects, and events? In *SafeAI@AAAI*, 2019.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[7] Tal Ridnik, Emanuel Ben Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *CoRR*, abs/2104.10972, 2021. URL `https://arxiv.org/abs/2104.10972`.