COMPUTER VISION University of Freiburg

Explore our Dataset

Open-vocabulary Attribute Detection

María A. Bravo   Sudhanshu Mittal   Simon Ging   Thomas Brox

{bravoma, mittal, gings, brox}@cs.uni-freiburg.de

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Motivation

A **white** and a **spotted** horse on a field of grass.

Red traffic signal in the middle of a **wide** street.

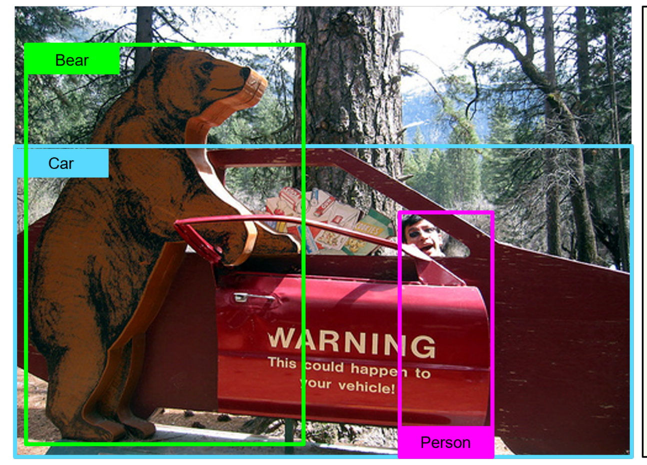
Open-vocabulary (OV) Recognition refers to the task of recognizing and understanding any visual concept in an image.

Current OV methods:

- ✓ Recognize objects beyond a closed-set of categories.
- ✓ Use available image-text pairs for supervision.
- ✓ Extend to new concepts using natural language.
- × Primarily focus on noun concepts.

⟶ **Attributes are important for an object's identity.**
They help distinguish different instances of the same class and enable better interpretation of scenes and decision-making.

OVAD: Open-vocabulary Attribute Detection Task



Object class: bear	Object class: car	Object class: person
Attributes	**Attributes**	**Attributes**
color quantity: two	color quantity: one	face exp: surprise
color: black and brown	color: red	group: single
group: single	group: single	hair color: black
material: wood	material: wood	hair length: short
maturity: adult	optical prop: opaque	hair tone: dark
position: upright	patterns: lettered	hair type: straight
size: big	state: piece / cut	maturity: adult
state: dry	texture: smooth	position: upright
texture: smooth	tone: dark	clothes color: white
tone: dark		

Objective: To evaluate the ability of visual-language models to recognize object attributes.

The OVAD task consists of two stages:

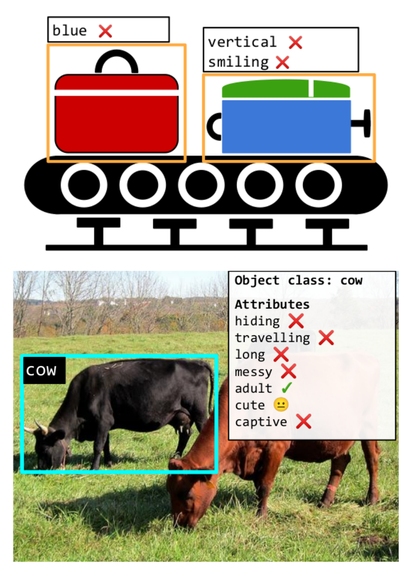
1. **Open-vocabulary object detection**: To detect an open-set of object classes.
2. **Open-vocabulary attribute recognition**: To identify an open-set of attributes for each detected object.

Attribute Benchmarks' Annotations

Test Dataset Requirements: Object and attribute annotations that are correct, dense, unambiguous, and visually consistent.

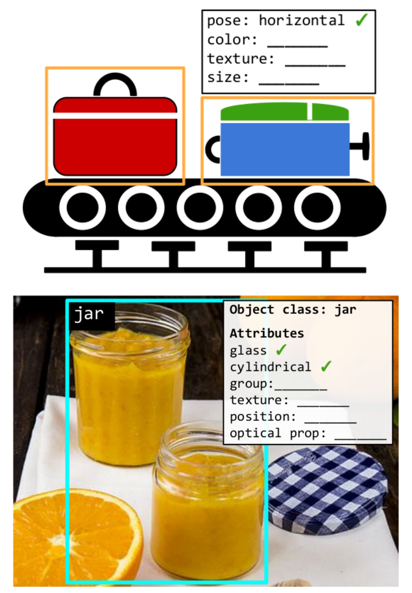
Four major types of errors in previous datasets.

(a) Incorrect ✗



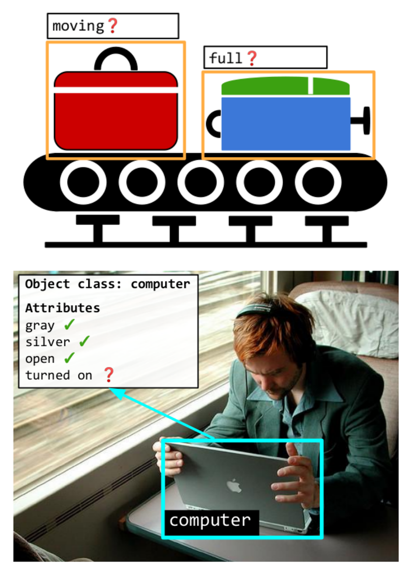
Objects with possible but incorrect attribute annotations.

(b) Missing



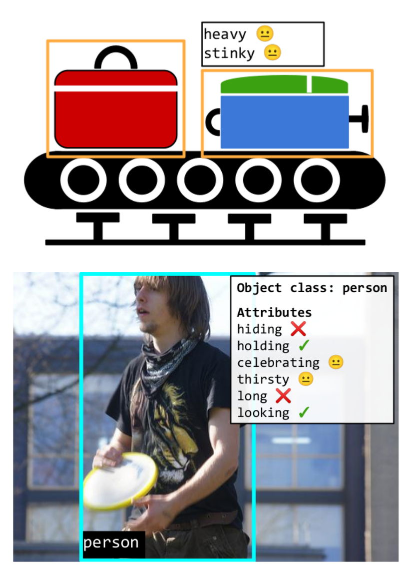
Objects with missing attribute annotations.

(c) Ambiguous ?



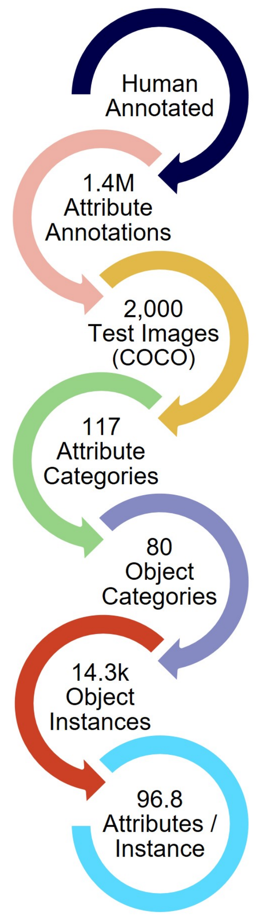
Attributes which cannot be marked using the image due to incomplete information.

(d) Non-visual



Attributes that cannot be marked using the visual information.

OVAD Dataset



Positive
clothes color: green, white
clothes pattern: lettered
face expression: neutral
group: single
maturity: young
position: standing

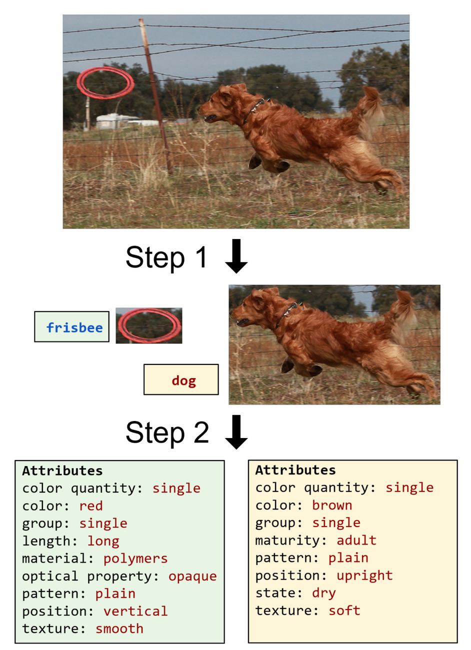
Negative
cleanliness: clean / unclean
clothes color: blue / brown / orange /…
clothes pattern: dotted / floral / plaid /…
cooked: cooked / raw
face expression: angry / happy / sleepy /…
group: group
length: long / short
material: asphalt / ceramic / glass /…
optical property: opaque / transparent /…
maturity: adult
order: messy / ordered
position: lying / sitting
state: empty / closed / dry / full /…
…

Unknown
gender: female / male
hair color: black / blue / brown / …
hair length: bald / long / short
hair tone: dark / light
hair type: curly / straight

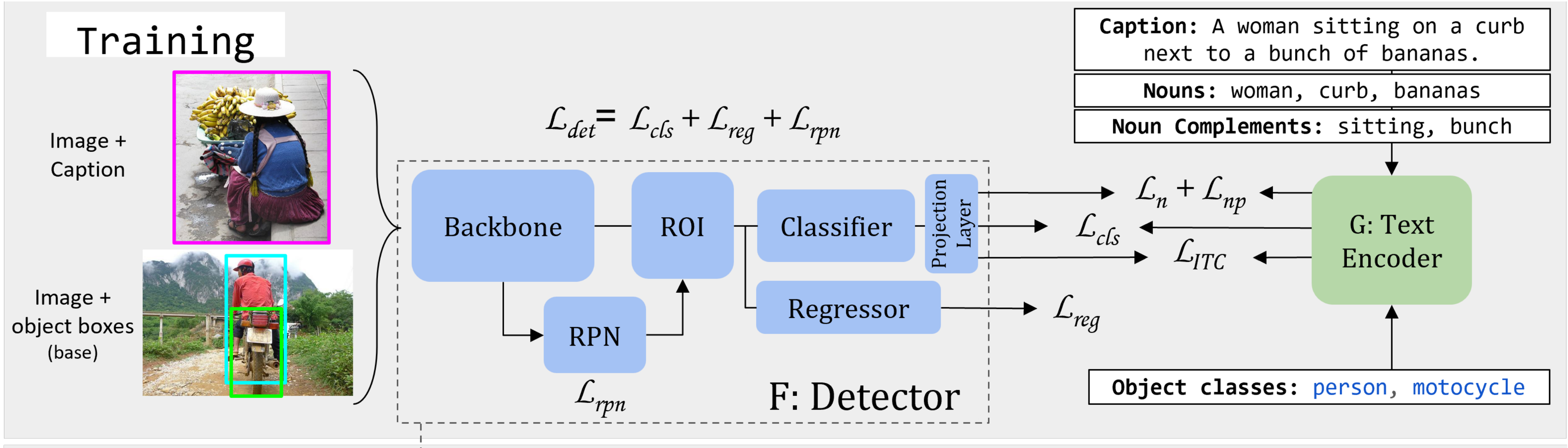
Attributes are positive, negative or unknown.

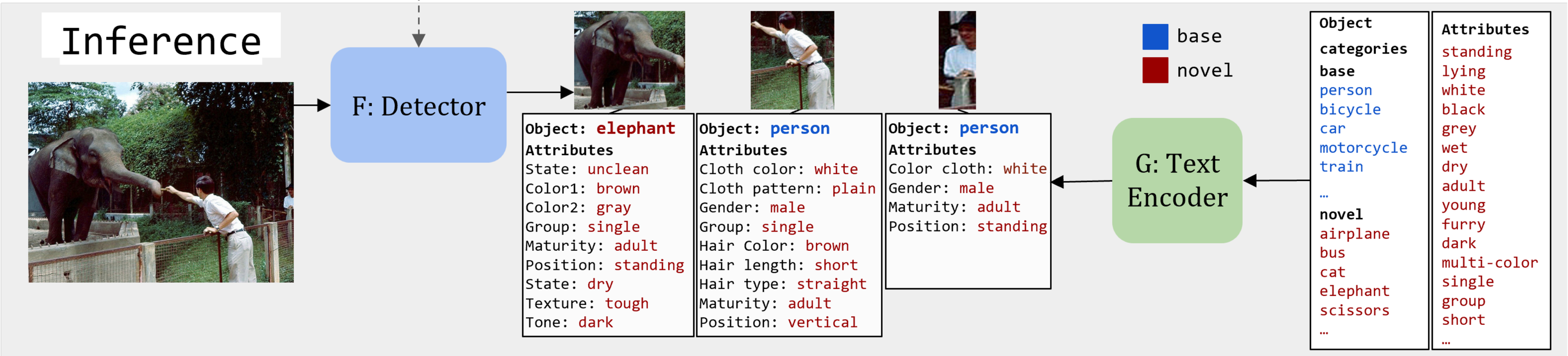
Evaluation Modes:

1. Full evaluation: (Steps 1+2)
   Detect objects and their attributes
2. Box-oracle: (Step 2)
   Detect attributes given the object box



OVAD Baseline





Training: A two-stage detector that matches image regions with text embeddings.

1. Use image-caption pairs and object detection annotations from base object classes.
2. Extract parts-of-captions: nouns and noun complements, as signals for learning visual-text alignment.

Inference:

1. Generates visual embeddings for objects.
2. Detects objects and attributes via cosine similarity with the class text embeddings.

Parts-of-caption Ablation

box+cls $\mathcal{O}^B$	captions	nouns	noun phrases	noun comp.	OVAD mAP	AP_{50} - OVD-80 Novel (32)
✓					11.7	0.3
✓	✓				15.0	19.2
✓	✓	✓			16.2	23.2
✓	✓	✓	✓		15.9	23.7
✓	✓	✓		✓	18.8	24.7

- The parts-of-caption help the model segregate the information in the caption, improving the object and attribute performance.
- Noun complements makes the attribute supervision more explicit and improves the performance.

Results

OVD Models on OVAD, Full Evaluation Setting

Method	OVAD All	OVAD Head	OVAD Medium	OVAD Tail	Generalized OVD-80 Novel (32)	Base (48)	All (80)
Chance	8.6	36.0	7.3	0.6	-	-	-
OV-Faster-RCNN	11.7	34.4	13.1	1.9	0.3	53.3	32.1
VL-PLM [1]	13.2	32.6	16.3	2.6	19.7	58.8	43.2
Detic [2]	13.3	44.4	13.4	2.3	20.0	49.2	37.5
Rasheed et al. [3]	14.6	33.5	18.7	2.8	32.5	56.6	46.9
LocOv [4]	14.9	42.8	17.2	2.2	22.5	52.5	40.5
OVR [5]	15.1	46.3	16.7	2.1	17.9	51.8	38.2
OVAD Baseline	**18.8**	**47.7**	**22.0**	**4.6**	24.7	49.1	39.3

- OVAD Baseline outperforms the latest OVD models on the OVAD task.
- OVD methods achieve results above the chance level on attribute detection even when trained only for object detection.
- Methods that incorporate image region with text-parts alignment (LocOv, OVR, OVAD Baseline) achieve better performance.

Large Vision-Language Models on OVAD, Box-oracle Setting

Method	Training Data	OVAD-Box All	Head	Medium	Tail
Chance	-	8.6	36.0	7.3	0.6
CLIP RN50 [6]	400M (9)	15.8	42.5	17.5	4.2
CLIP VIT-B16 [6]	400M (9)	16.6	43.9	18.6	4.4
Open CLIP RN50 [7]	12M (7b)	11.8	41.0	11.7	1.4
Open CLIP ViT-B16 [7]	400M (8b)	16.0	45.4	17.4	3.8
Open CLIP ViT-B32 [7]	2B (8c)	17.0	44.3	18.4	5.5
ALBEF [8]	4M (1a,3,4,7a)	15.6	43.1	17.3	3.7
ALBEF [8]	14M (1a,3,4,7)	15.3	43.7	17.1	3.0
ALBEF [8]	14M (1a,3,4,7) + ft(2)	21.0	44.2	23.9	9.4
BLIP [9]	14M (1a,3,4,7)	17.0	46.6	18.3	5.0
BLIP [9]	129M (1a,3,4,7,8a)	18.2	44.4	20.7	5.7
BLIP [9]	129M (1a,3,4,7,8a) + ft(1a)	24.3	51.0	28.5	9.7
BLIP-2_{Large} [10]	129M (1a,3,4,7,8a)	20.1	49.3	23.2	5.9
BLIP-2 [10]	129M (1a,3,4,7,8a)	21.6	44.7	24.0	10.3
BLIP-2 [10]	129M (1a,3,4,7,8a) + ft(1a)	25.5	49.8	30.5	10.9
X-VLM [11]	4M (1*,3*,4,7a)	25.9	**50.3**	32.0	9.8
X-VLM [11]	16M (1*,3*,4,5*,6*,7) + ft(2)	26.2	48.7	31.2	12.1
X-VLM [11]	16M (1*,4*,4,5*,6*,7)	**28.1**	49.7	**34.2**	**12.9**
OVAD Baseline-Box	0.11M (1a,1b$^{*\text{base}}$)	21.4	48.0	26.9	5.2

* use of localization information from the annotations.

+ ft: final fine-tuning using the captions of this dataset.

(#) Dataset	#Images	#Captions	#Objects	#Regions
(1a) COCO Captions	0.12M	0.57M	-	-
(1b) COCO Objects	0.12M	-	0.86M	-
(2) RefCOCO+	0.019M	-	-	0.14M
(3) VG	0.10M	-	2.5M	5.4M
(4) SBU Captions	1M	1M	-	-
(5) OpenImages	1.7M	0.67M	4.4M	3.3M
(6) Objects365	1.8M	-	29M	-
(7a) CC-3M	2.95M	2.95M	-	-
(7b) CC-12M	11.1M	11.1M	-	-
(8a) LAION	115M	115M	-	-
(8b) LAION	400M	400M	-	-
(8c) LAION	2B	2B	-	-
(9) CLIP 400M	400M	400M	-	-

- VLMs tend to focus on object classes and struggle with fine-grained aspects like attributes.
- The quality of the training data has a greater impact than its quantity or model size.
- Fine-grained alignment between image regions and text (X-VLM) significantly improves the understanding of visual attributes.

Conclusions / Contributions

- We propose the open-vocabulary attribute detection (OVAD) task to study vision-language models' ability to recognize attributes.
- We introduce the OVAD benchmark, a clean and densely annotated object-level attribute dataset for evaluating the OVAD task.
- We provided a baseline method that exploits fine-grained information contained in captions.
- We found that the performance of foundation models on attributes stays clearly behind their performance on objects.

References

[1] S. Zhao et al., "Exploiting unlabeled data with vision and language models for object detection," in *ECCV*, 2022.
[2] X. Zhou et al., "Detecting twenty-thousand classes using image-level supervision," in *ECCV*, 2022.
[3] H. A. Rasheed et al., "Bridging the gap between object and image-level representations for open-vocabulary detection," in *NeurIPS*, 2022.
[4] M. Bravo et al., "Localized vision-language matching for open-vocabulary object detection," in *GCPR*, 2022.
[5] A. Zareian et al., "Open-vocabulary object detection using captions," in *CVPR*, 2021.
[6] A. Radford et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021.
[7] G. Ilharco et al., "Openclip," 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5143773
[8] J. Li et al., "Align before fuse: Vision and language representation learning with momentum distillation," in *NeurIPS*, 2021.
[9] ——, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.
[10] ——, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023.
[11] Y. Zeng et al., "Multi-grained vision language pre-training: Aligning texts with visual concepts," in *ICML*, 2022.

Project page

Acknowledgements

This work was supported by the German Academic Exchange Service (DAAD) - 57440921 Research Grants - Doctoral Programmes in Germany, 2019/20. This work was partially funded by the German Research Foundation (DFG) - 401269959 and 417962828.