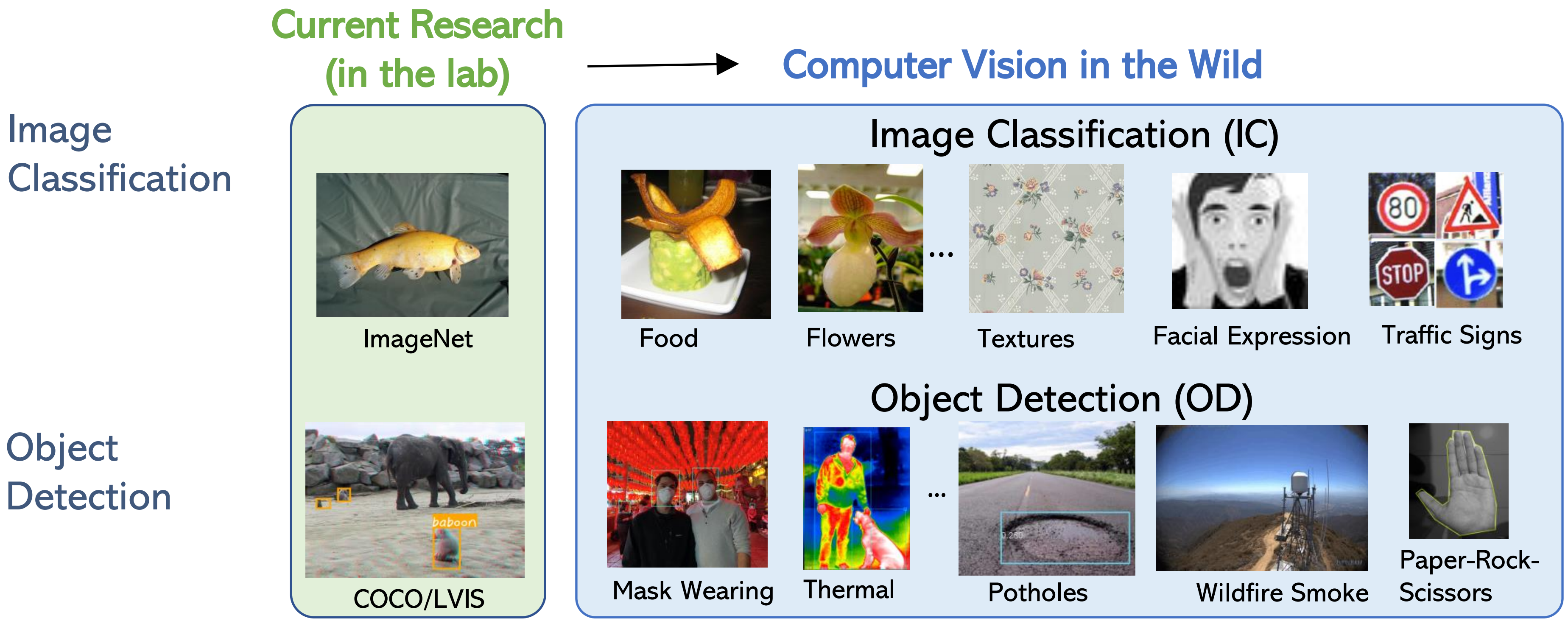
Computer Vision in the Wild:
Benchmark & Challenge Summary

October 2022

Chunyuan Li
Deep Learning Team
Microsoft Research, Redmond

Why CVinW ? Evaluation of Language-augmented Visual Task-level Transfer



Trend

- Building transferable systems (eg, foundation models) that can adapt to a wide range of CV tasks
- Inspired by the success of CLIP, many language-augmented visual models appear

Challenges

- **Fairness:** Customized task sets may favor individual pre-trained model
- **Transparency**: Detailed model adaptation process is inaccessible

What is **Computer Vision in the Wild (CVinW)** ?

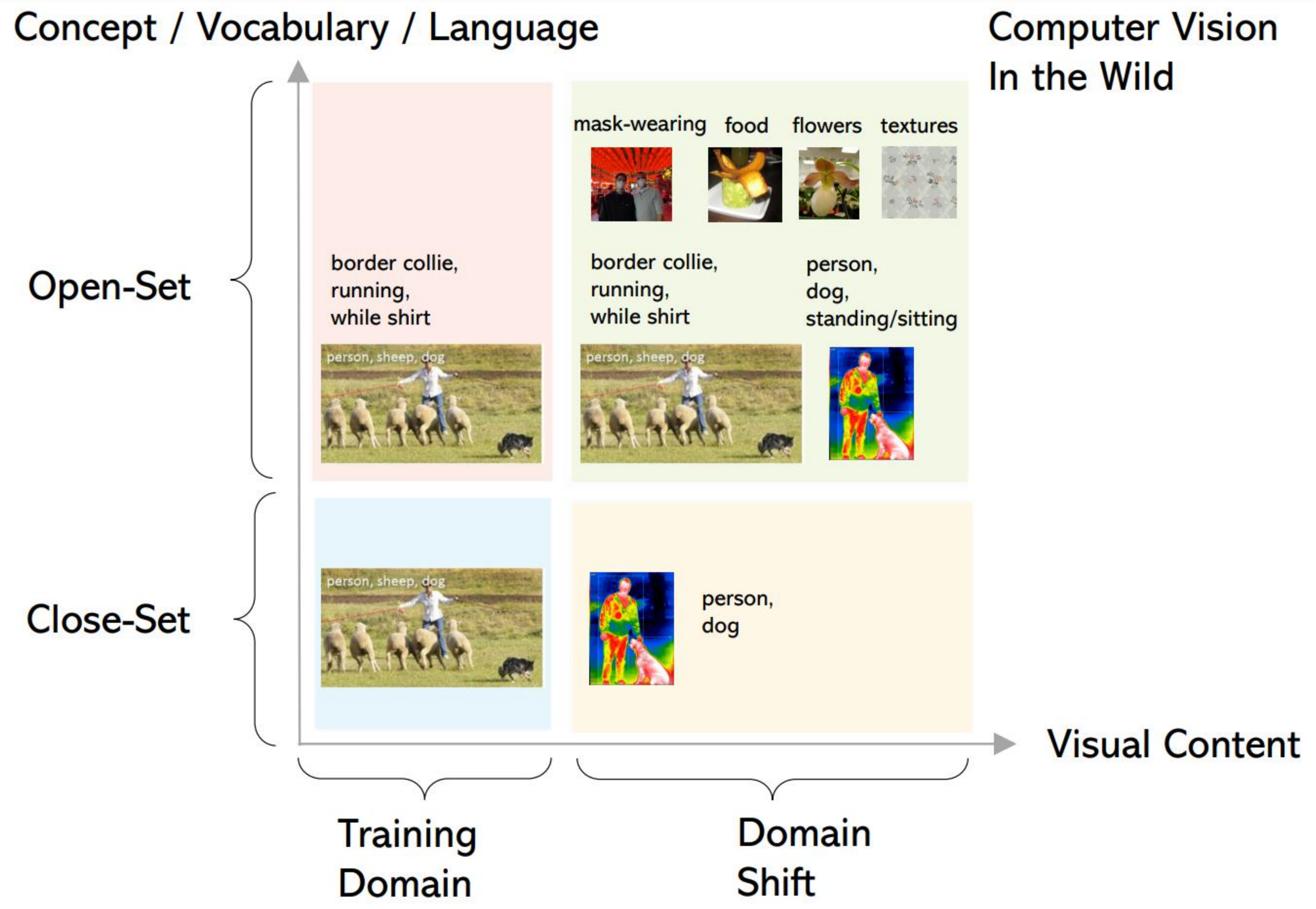
Developing a transferable foundation model/system that
can *effortlessly* adapt to *a large range of visual tasks* in the wild.

It comes with two key factors:

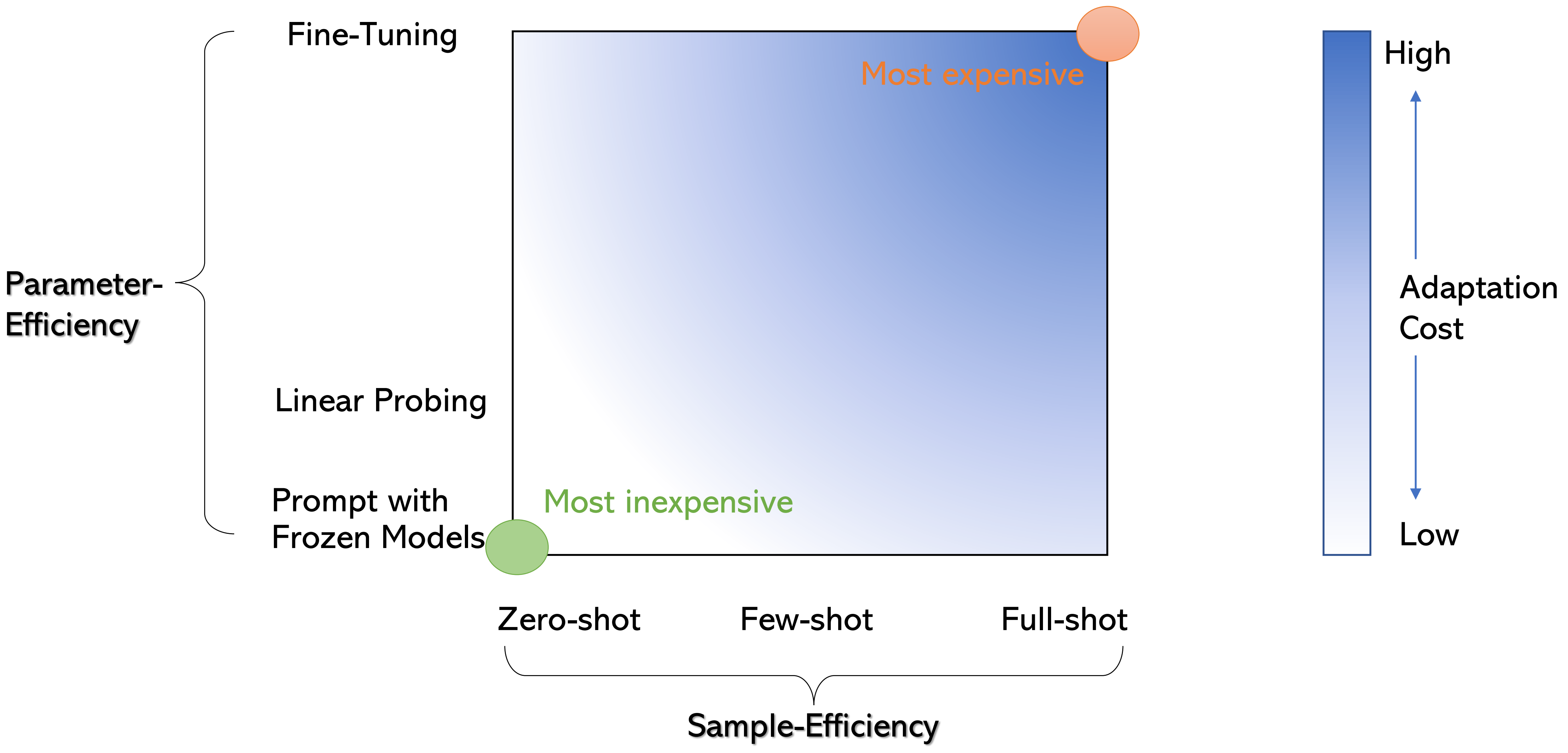
1. The task transfer scenarios are broad
2. The task transfer cost is low.

https://github.com/Computer-Vision-in-the-Wild/CVinW_Readings

1 CVinW *vs* other CV settings



2 2D space for the definition of adaptation cost



Where to start?

ICinW
Image Classification in the Wild

ODinW
Object Detection in the Wild

ELEVATER:
A Benchmark and Toolkit for Evaluating Language-Augmented Visual Models

https://arxiv.org/abs/2204.08790 NeurIPS 2022 (Benchmarks and Datasets Track)

Chunyuan Li[*1♠], Haotian Liu[*2], Liunian Harold Li[3], Pengchuan Zhang[1], Jyoti Aneja[1]
Jianwei Yang[1], Ping Jin[1], Houdong Hu[1], Zicheng Liu[1], Yong Jae Lee[2], Jianfeng Gao[1]
[1]Microsoft [2]University of Wisconsin–Madison [3]UCLA

Benchmarks: ELEVATER

- ## Dataset Suite

- Image Classification: **20** datasets

- Object Detection: **35** datasets

- ## External Knowledge

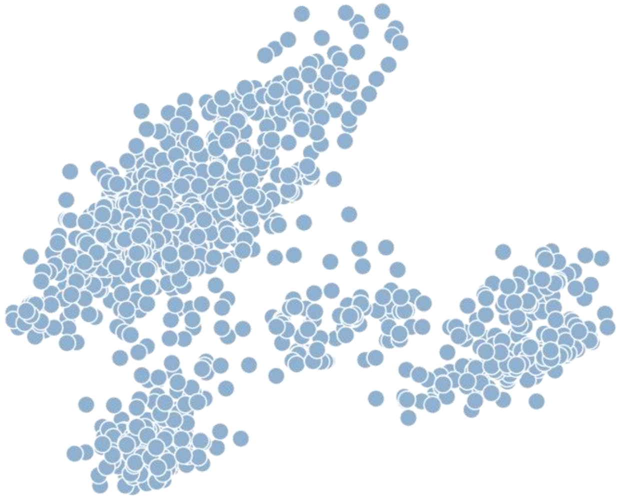
WordNet, Wiktionary, GPT-3

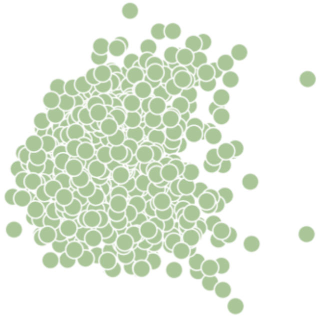
Benchmarks: A more diverse set of tasks

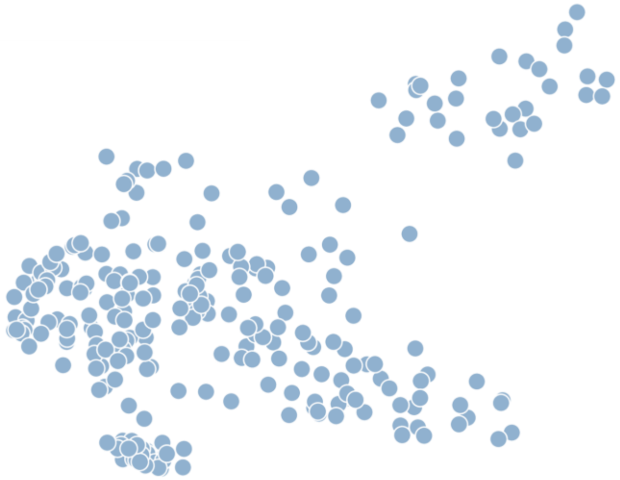
ImageNet → Image Classification in the Wild (20 datasets)

LVIS → Object Detection in the Wild (35 datasets)

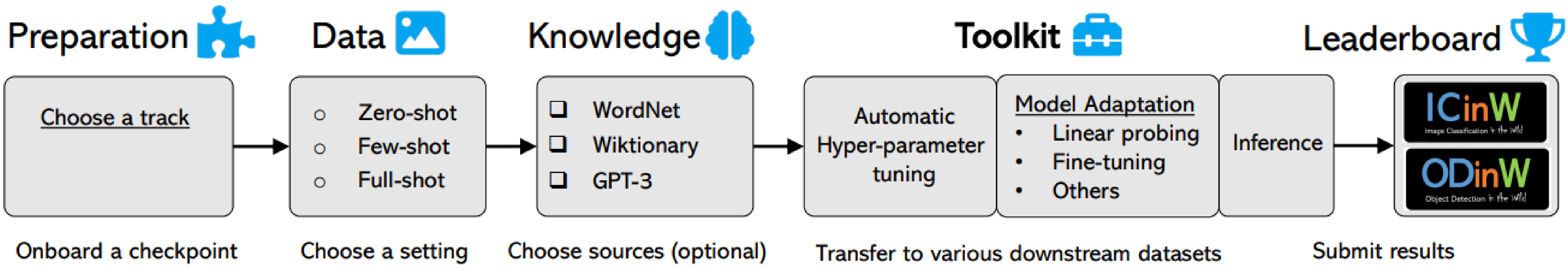
Semantic Space using PCA







	ImageNet	Image Classification in the Wild (20 datasets)	LVIS	Object Detection in the Wild (35 datasets)
Task diversity using std	0.610	0.680	0.533	0.619



Where to submit results? Challenge → Track → Phase

Challenge	Track	Definition (*IN = ImageNet)
ICinW Image Classification in the Wild	Industry	No IN-1K data; Scaling Success
	Academic	No IN-1K data; Limited pretrain data (IN21K. CC3M+12M, YFCC15M)
	ImageNet-1K in Pretraining	IN1K is allowed in pretraining, *eg,* self-supervised Learning
	Parameter-Efficiency	Efficient model adaption methods
ODinW Object Detection in the Wild	Zero-Shot	No Training examples in ODinW are used
	Full-shot	All Training examples in ODinW are used

A collaborative community-effort

-- to benchmark the SOTA foundation vision models

Berkeley UNIVERSITY OF CALIFORNIA

UCLA

WISCONSIN UNIVERSITY OF WISCONSIN-MADISON

COLUMBIA UNIVERSITY

NYU

UNIVERSITY of WASHINGTON

UNIVERSITY OF CALIFORNIA SANTA CRUZ

sensetime

HUAWEI

TEXAS The University of Texas at Austin

University at Buffalo The State University of New York

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Tencent

Alibaba Group

ByteDance

ZHEJIANG UNIVERSITY 1897

University of Science and Technology of China 1958

TSINGHUA UNIVERSITY 1911

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

idea INTERNATIONAL DIGITAL ECONOMY ACADEMY 粤港澳大湾区数字经济研究院

Om Research

Challenge Talks

-- one talk for each track

- ICinW Industry Track | Chinese CLIP | Junyang Lin (Alibaba)
- ICinW Academic Track | K-LITE | Sheng Shen (University of California, Berkeley)
- ICinW ImageNet-1K in Pre-training | Bamboo | Yuanhan Zhang (Nanyang Technological University)
- ICinW Parameter-Efficiency | ProDA | Yuning Lu (University of Science and Technology of China)
- ODinW Zero-Shot Track | DetCLIP | Jianhua Han (Huawei)
- ODinW Full-Shot Track | DINO | Shilong Liu (IDEA & Tsinghua)

Criterion

- **Ranking:** Top ranked methods by October 20, 2022 (3 days ago)
- **Availability**: Make a video presentation before Challenge
- **Deduplication**: No duplicated presentation between workshop paper and challenge
- **Note**: The ranking has been changing in the past 3 days

1

ICinW

Image Classification in the Wild

- Industry Track

- Larger models are better (4B > 1.6B > 1.0B > 0.4B)
- Foreign image-text pre-training is effective to IC in English
- Generative models such as GIT has a large space to improve for IC
- ALIGN is comparable with CLIP

		20 datasets		19 datasets			
Team	Method	Average Score	Rank	Average Score**	Rank **	# Model Params [M]	# Vision Backbone Params [M]
Microsoft	Turing Bletchley v2	73.5	1	73.6	1	4,240.3	3,668.84
TinyCLIP	X-CLIP	71.2	2	71.8	2	1,602.1	992.71
clip	clip	69.1	3	69.4	3	986.1	632.08
ELEVATER	CLIP [ViT-L/14 336x336]	66.8	4	67.2	4	427.9	304.29
OFA-Team	Chinese CLIP [ViT-H/14-	62.3	5	62.7	6	957.6	632.08
ELEVATER	CLIP [ViT-B/16]	60.0	6	60.1	8	149.7	86.19
eceipeno	eceipeno	57.2	7	57.5	9	182.3	86.65
ELEVATER	CLIP [ViT-B/32]	56.8	8	56.9	10	151.3	87.85
GIT - Single Generative Model		55.3	9	55.6	11	0.0	638.52
YT	pclip	54.9	10	55.0	12	151.6	88.14
DeCLIP	DeCLIP [ViT-B/32]	51.0	11	50.8	13	154.6	89.82
cathy4k	Chinese CLIP					406.8	303.97
ALIGN_LARGE				67.0	5	304.0	304.00
ALIGN				62.1	7	86.0	86.00

Results on October 22, 2022 PT

2

ICinW

Image Classification in the Wild

\- Academic Track

- MaskCLIP ranks 1st on ICinW
- External knowledge is useful

- The conclusions are inconsistent between ICinW and ImageNet-1K; Be more careful when designing architectures and objectives

- Not all models outperform CLIP trained on YFCC

			Zero-shot on 20 datasets in ICinW			Zero-shot
Rank	Team	Method	Average Score	# Model Params [M]	# Vision Backbone Params [M]	[Ref.] ImageNet-1K
1	DLight	MaskCLIP	48.9	196.0	94.22	56.5
2	KLITE	Swin-B; 3 datasets	45.5	150.7	86.74	57.8
3	YT	YT-CLIP	44.5	151.2	87.79	52.9
4	ELEVATER	UniCL + FocalB [3 datasets]	44.0	155.4	91.44	54.2
5	Gramer	UniCL [SwinB, 3 datasets]	43.2	150.7	86.74	52.2
6	ELEVATER	UniCL+ViT-B [IN21K + GCC]	42.4	149.6	85.8	45.1
7	ELEVATER	UniCL+FlatFocal-B [IN21K + GCC]	41.8	150.5	86.65	47.4
8	ELEVATER	UniCL+DaViT-B [IN21K + GCC]	40.4	150.9	86.93	47.3
9	ELEVATER	UniCL+Focal-B [IN21K+GCC]	39.5	155.4	91.44	47.1
10	DeCLIP	DeCLIP [ViT-B/32, YFCC-15M]	37.9	154.6	89.82	
11	FILIP	FILIP(ViT-B32,YFCC15M)	34.5	177.3	88.05	
12	ELEVATER	CLIP [ViT-B/32, YFCC-15M]	32.0	151.3	87.85	
13	SLIP	SLIP (ViT-B/YFCC-15M)	31.2	172.3	87.85	
14	ELEVATER	SLIP (YFCC15M)	31.2	172.3	87.85	
15	MS-CLIP	MS-CLIP (ViT/B32|YFCC)	30.3	132.4	86.89	36.5
16	ELEVATER	UniCL [Swin-T, ImageNet-21K]	27.2	91.4	27.52	
17	CyCLIP	CyCLIP + [ResNet-50, GCC-3M]	25.8	102.0	38.32	22.0
18	ELEVATER	ResNet-50, GCC-3M	25.3	102.0	38.32	19.8

Results on October 22, 2022 PT

3

ICinW

Image Classification in the Wild

- ImageNet1K in Pretraining

- Bamboo ranks 1st on ICinW; Data-centric AI is effective
- Image self-supervised learning methods are popular
- Zero-shot FLAVA outperforms FT and LP many models

Rank	Team	Method	Average Score	# Vision Backbone Params [M]	# Trainable Params [K]
1	Bamboo	Bamboo-ViTB/16 LP	63.7	85.8	34.6
2	ELEVATER	ViT [ViT-B/16, LP]	57.6	85.8	44.3
3	ELEVATER	ViT [ViT-B/16, FT]	57.2	85.8	85842.96
4	Amazon-m5	DeCL [ViT-B/16, LP]	54.7	85.8	85842.96
5	CACR	CACR [ViT-B/16, LP]	54.5	86.57	7.69
6	ELEVATER	DeiT [ViT-B/16, LP]	54.1	85.8	44.3
7	ELEVATER	DeiT [ViT-B/16, FT]	54.1	85.8	85842.96
8	CACR	CACR [ViT-B/16, LP]	53.6	86.57	7.69
9	ELEVATER	MoCo-v3 [ViT-B/16, LP]	50.2	85.8	44.3
10	CACR	CACR [ViT-B/16, FT]	48.9	86.57	86424.05
11	Facebook AI Research	FLAVA (PMD-ZeroShot)	48.7	241.36	0.0
12	Bamboo	Bamboo-ViTB/16 FT	48.4	85.8	85833.26
13	SupMAE	SupMAE [ViT-B/16, FT]	46.8	85.8	85842.96
14	CAE v1	Baidu (CAE v1 [ViT-B/16, LP])	44.1	85.81	44.26
15	ELEVATER	MoCo-v3 [ViT-B/16, FT]	39.3	85.8	85842.96
16	CAE v1	Baidu (CAE v1 [ViT-B/16, FT])	37.9	85.81	85807.87
17	ELEVATER	MAE [ViT-B/16, FT]	36.1	85.8	85842.96
18	ELEVATER	MAE [ViT-B/16, LP]	33.4	85.8	44.3
19	ELEVATER	From-Scratch [ViT-B/16, FT]	20.8	85.8	85842.96
20	CV-Team	ViT-Contrastive	20.7	86.57	86570.73
21	ELEVATER	From-Scratch [ViT-B/16, LP]	19.6	85.8	44.3

Results on October 22, 2022 PT

4

ICinW

Image Classification in the Wild

- Parameter-Efficiency

A single number that measures both **prediction accuracy** and **parameter-efficiency**

A Constant *eg*, 10^8

$$\mathrm{PE} = \texttt{score} * \exp(\log_{10}(\texttt{\# trainable-parameters}/M_0 + 1))$$

- ProDA ranks the 1st
- Advanced adaptation methods from NLP are not really better than linear probing?

Rank	Team	Method	Accuracy-Efficiency Metric	Average Score	# Vision Backbone Params [M]	# Trainable Params [K]
1	ProDA	ProDA [CLIP, ViT-B/16]	0.71	70.7	86.19	262.14
2	ELEVATER	Linear Probe [CLIP, ViT-B/16]	0.68	68.3	86.19	29.57
3	PEViT-Adapter	Adapter [CLIP, ViT-B/32]	0.65	65.1	89.06	1211.65
4	ELEVATER	Linear Probe [CLIP, ViT-B/32]	0.65	65.3	87.85	29.57
5	PEViT-LoRA	LoRA [CLIP, ViT-B/32]	0.61	61.5	88.0	151.05
6	ELEVATER	Ref: Zero-Shot [CLIP, ViT-B/16]	0.6	60.0	86.19	0.0
7	ELEVATER	Ref: Zero-Shot [CLIP, ViT-B/32]	0.57	56.8	87.85	0.0
8	ELEVATER	Fine-tuning [CLIP, ViT-B/16]	0.53	69.1	86.19	86222.21
9	ELEVATER	Fine-tuning [CLIP, ViT-B/32]	0.48	63.3	87.85	87878.78

Results on October 22, 2022 PT

5

ODinW
Object Detection in the Wild
- Zero-Shot Track

Thanks for the suggestion from Matthias Minderer

- Florence ranks 1st with average;
- DetCLIP ranks 1st with median

- Larger pre-training dataset leads to better performance; Though MDETR trained on a small dataset, the results are not bad

Rank	Team	Method	Average Score (Common Metric)	Median Score (Robust Metric)
1	Florence	FL-1.5-D5	25.8	14.3
2	DetCLIP-team	DetCLIP	24.9	18.3
3	GLIPv2_team	GLIPv2-T	22.3	8.9
4	OmLab	OmDet	19.7	8.9
5	ODinW_Team	GLIP-T	19.6	5.1
6	FIBER	FIBER	19.5	10.4
7	OmLab	OmDet	19.0	8.9
8	Google Research	OWL-ViT L/14 @ 672	18.8	9.8
9	ODinW_Team	GLIP-T (B)	12.8	2.2
10	ODinW_Team	GLIP-T (A)	11.4	1.6
11	MDETR-NYU	MDETR - ENB5	10.7	3.0
12	MDETR-NYU	MDETR - ENB3	10.1	2.7
13	MDETR-NYU	MDETR - R101	9.9	3.1

Results on October 22, 2022 PT

6

ODinW

Object Detection in the Wild

- Full-Shot Track

- OmLab ranks 1st with average and median;
- DINO, the best performing OD head on COCO, performs well on ODinW

Full-Shot:

Rank	Team	Method	Average Score	Median Score
1	OmLab	OmDet	67.1	71.2
2	IDEA-CVR-DINO	DINO-SwinT	66.7	68.5
3	OmLab	OmDet_Base	65.7	65.7
4	IDEA-CVR-DINO	DINO-SwinT(Merged	65.3	65.1
6	ODinW_Team	DyHead-T	63.2	64.9
7	ODinW_Team	GLIP-T	62.6	62.1

Few-Shot:

Rank	Team	Method	Average Score	Median Score
1	OmLab	OmDet	42.4	41.7
2	IDEA-CVR-DINO	DINO-SwinT	41.2	41.1
3	ODinW_Team	GLIP-T	38.9	33.7
4	ODinW_Team	DyHead-T	37.5	36.7

Results on October 22, 2022 PT

📢 Call for Collaboration

-- benchmarking the transfer ability of SoTA vision models

- **Criterion**
  - **Goals**:
    - Summary of Challenge results on ICinW and ODinW
    - A comprehensive technical report to benchmark the best vision checkpoints and adaptation methods
    - A shared view to push CVinW
  - **Authorship:** Contributors with valid submissions are encouraged to co-author the report
  - **Timeline:** **The** 1st version by the end of 2022; Continual updating arXiv when necessary

- **Inspiring Examples** BIGBench and BigScience in NLP

- **Future Update**

  https://computer-vision-in-the-wild.github.io/eccv-2022/

Challenge Presentation Agenda

CVinW

Computer Vision in the Wild

https://computer-vision-in-the-wild.github.io/eccv-2022/

Talk	Challenge	Track
1	ICinW	Industry
2		Academic
3		ImageNet-1K in Pretraining
4		Parameter-Efficiency
5	ODinW	Zero-Shot
6		Full-shot