JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Poster Tag: WED-PM-351

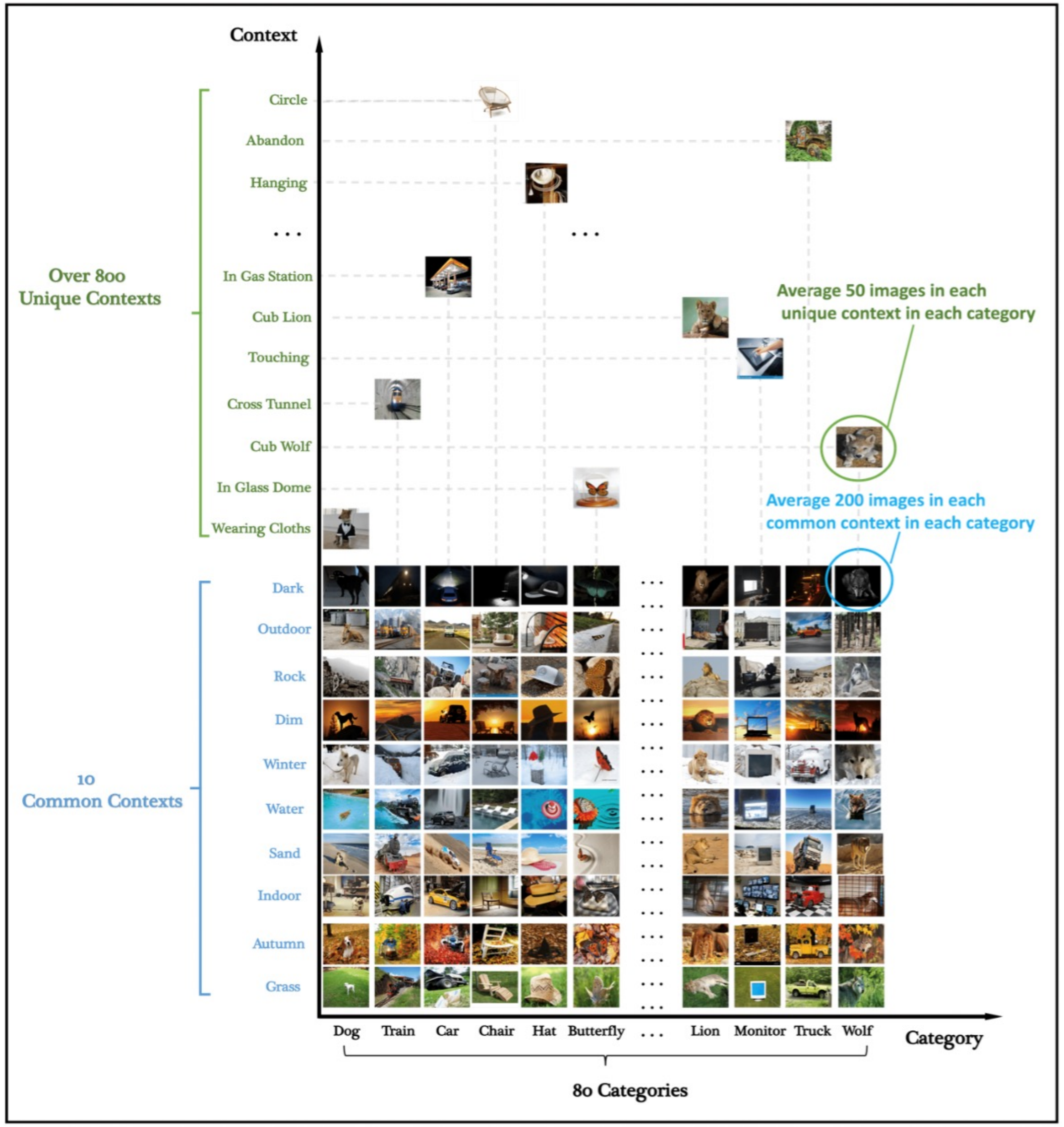
NICO++: Towards Better Benchmarking for Domain Generalization

CVPR, 2023

Xingxuan Zhang[†]*, Yue He*[†]*, Renzhe Xu, Han Yu, Zheyan Shen, Peng Cui*[*]

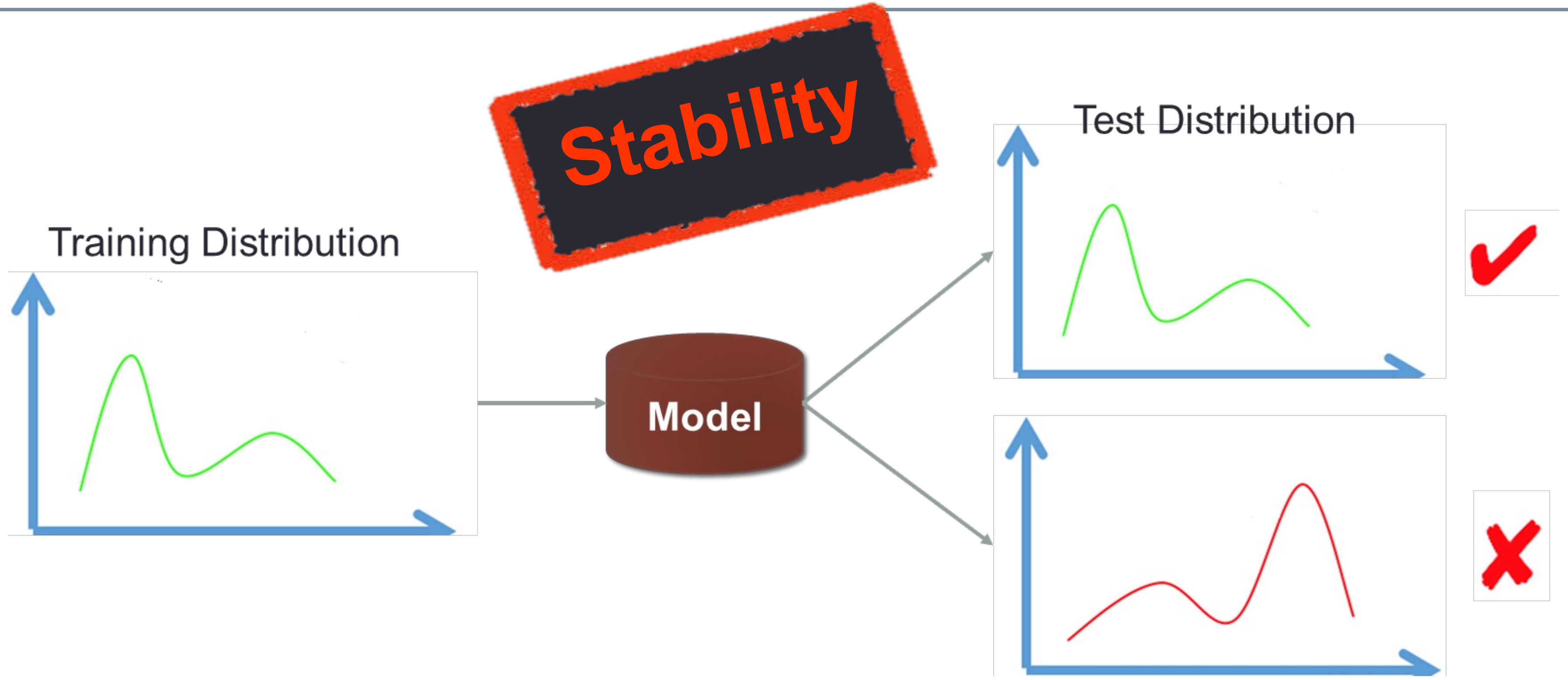
Brief view of NICO++

- A brand new large visual dataset
  - Towards better evaluation of visual OOD generalization
  - More than 200,000 images with 20 domains
  - Lower concept shift and higher covariate shift
  - Benchmarking for both standard DG and flexible DG
  - Mitigating the potential leakage of test information



Risks of Today’s AI Algorithms

Most ML methods are developed under I.I.D hypothesis



Stability

Training Distribution

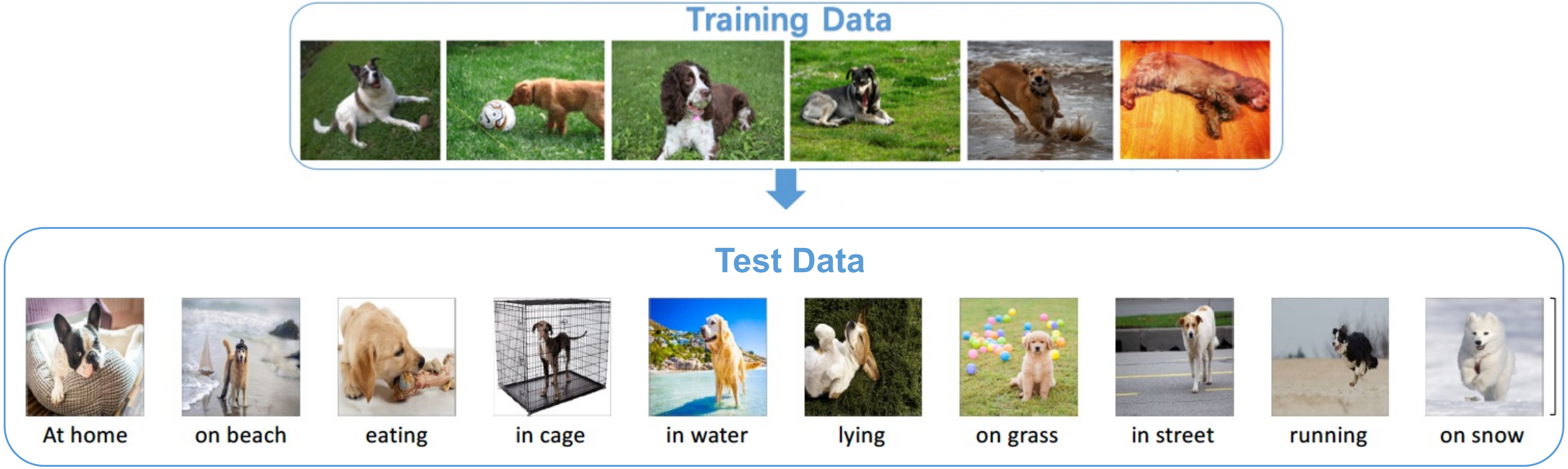
Model

Test Distribution

✔

✘

OOD Generalization in Visual Recognition

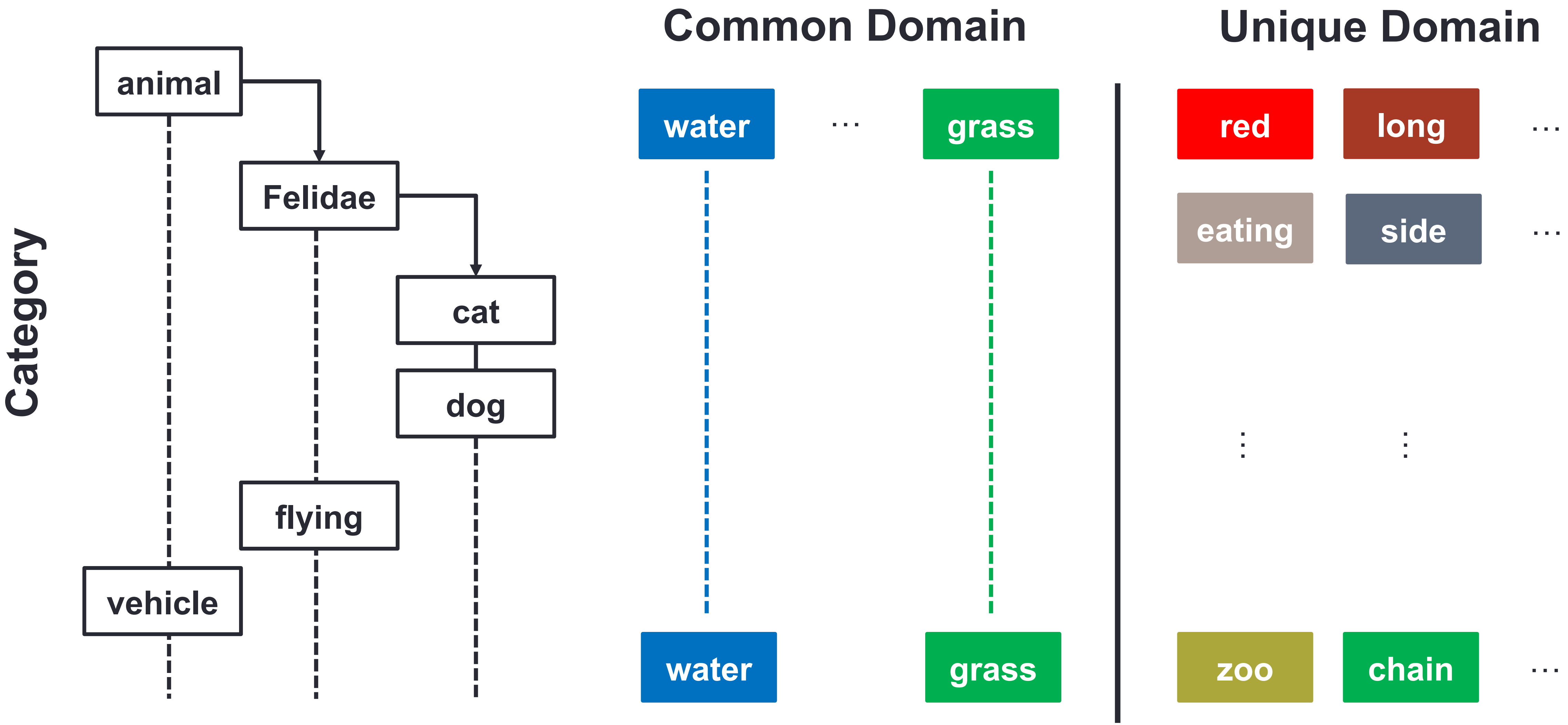


Training Data

Test Data

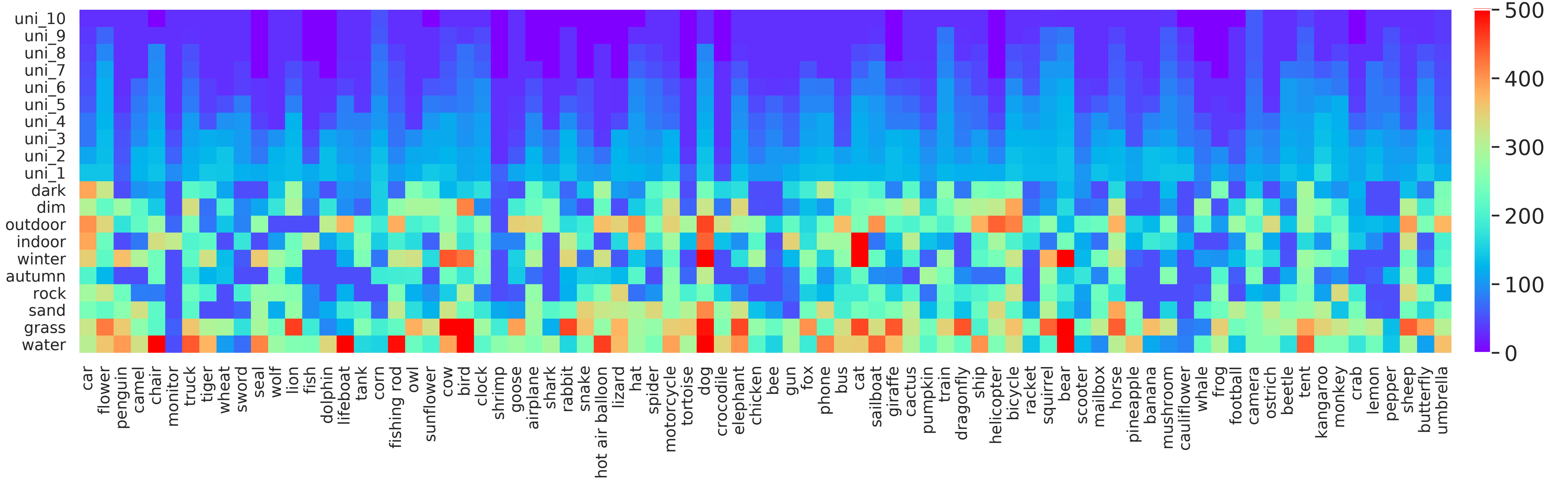
At home on beach eating in cage in water lying on grass in street running on snow

Structure of NICO++

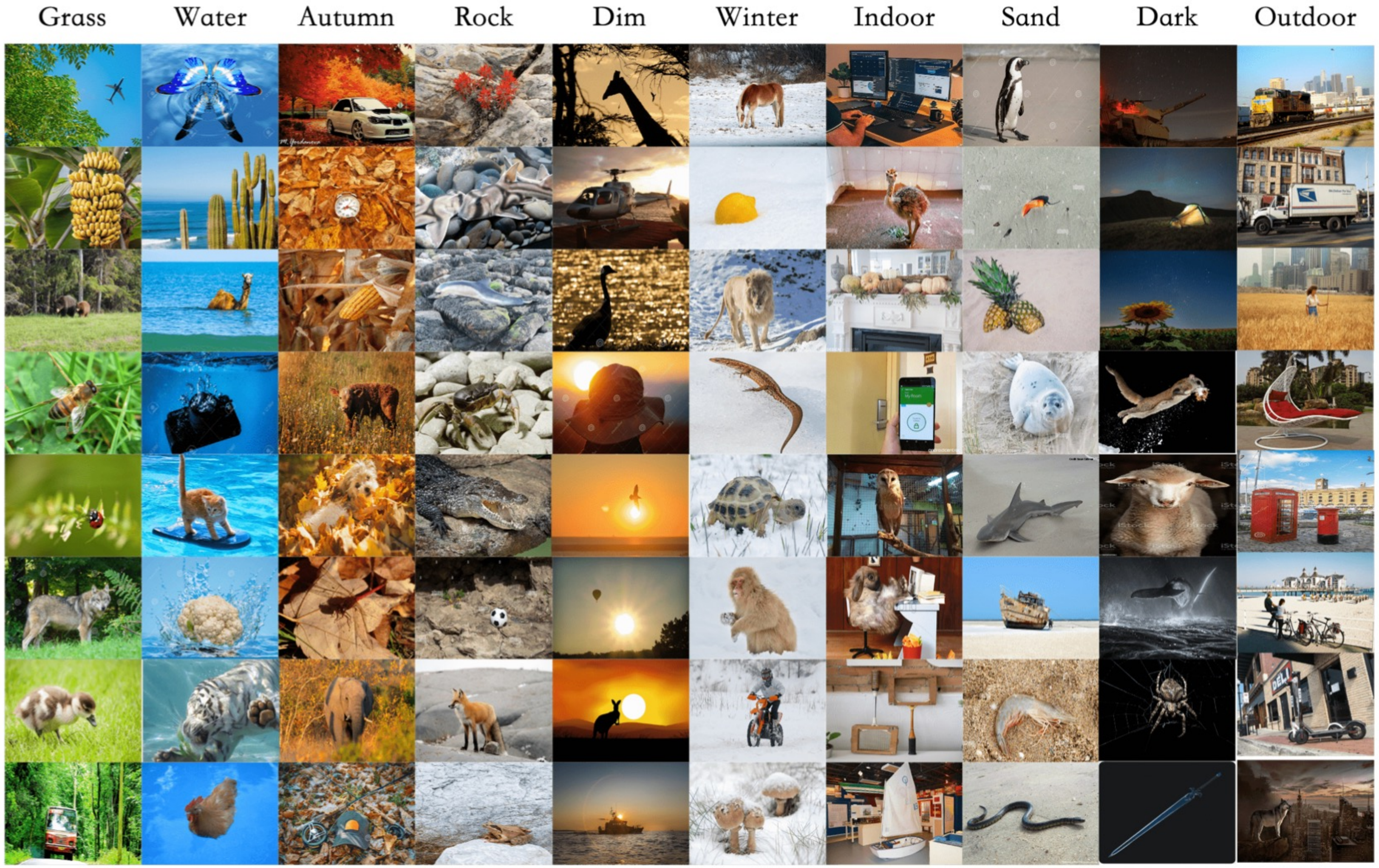


Statistics of NICO++

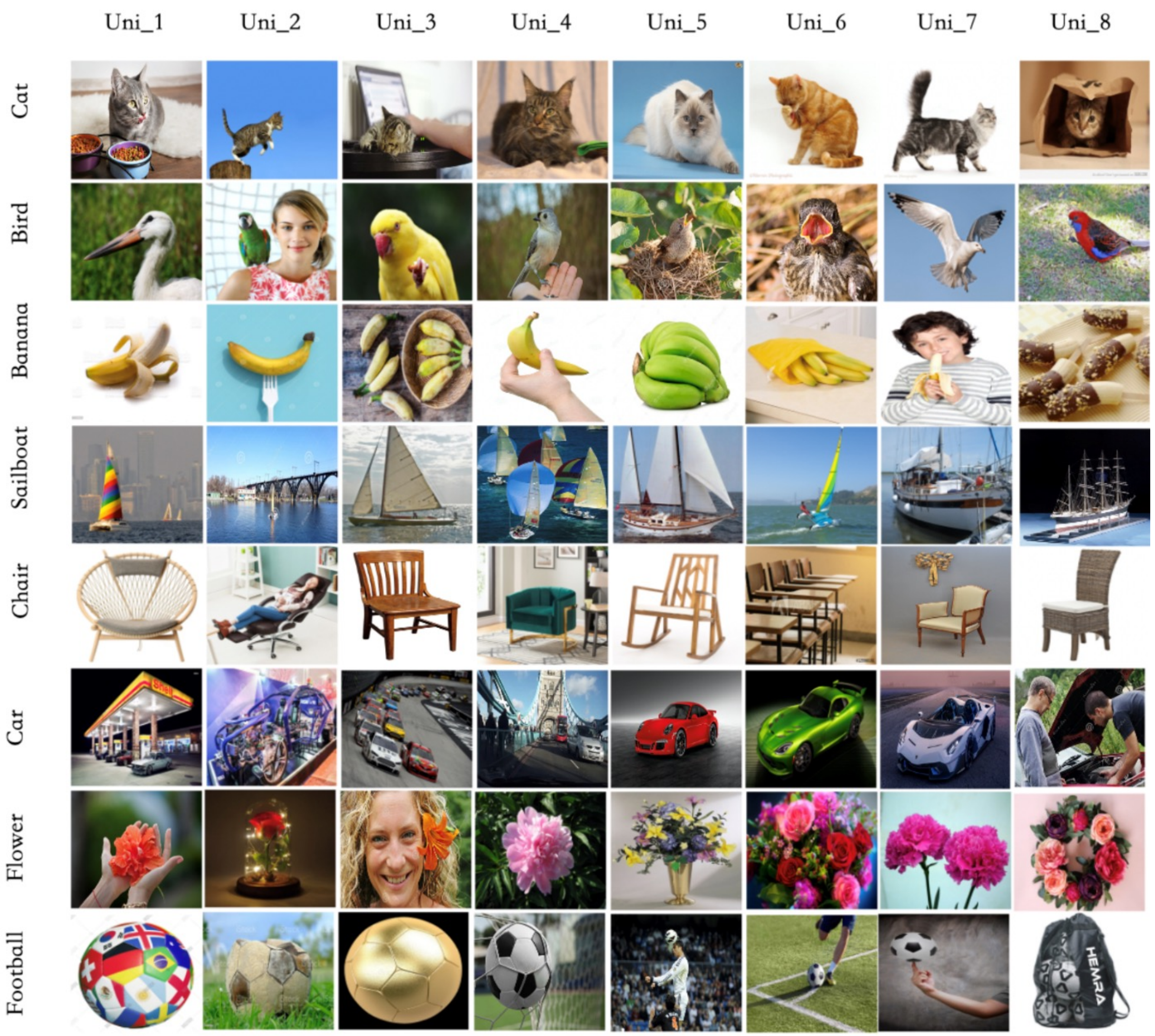
- 80 categories
- 10 common domains for each category
- 10 unique domains for each category
- Still growing…



Common Domains



Unique Domains



Covariate shift and concept shift

$$\varepsilon_{te}(h) \leq \varepsilon_{tr}(h) + \mathcal{M}_{\text{cov}}\left(\mathcal{D}_{tr}, \mathcal{D}_{te}; \mathcal{H}, \ell\right) + \mathcal{M}^{\text{min}}_{\text{cpt}}\left(\mathcal{D}_{tr}, \mathcal{D}_{te}, f_{tr}, f_{te}; \ell\right).$$

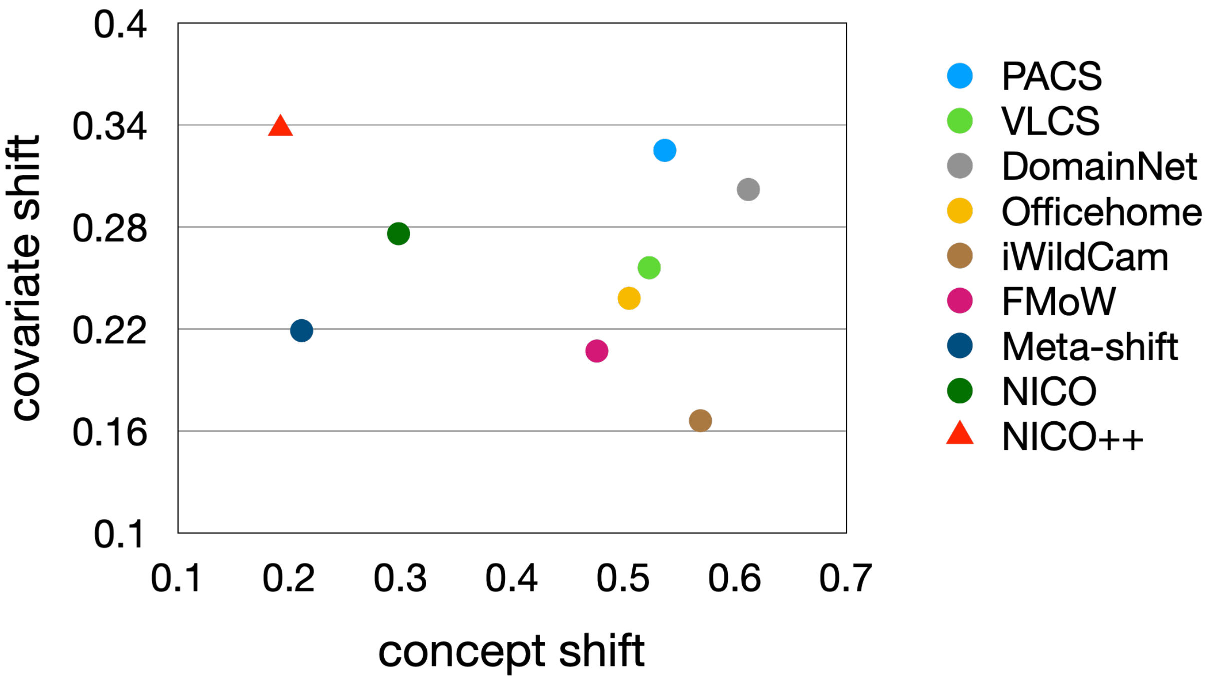
prediction error on target data

prediction error on source data

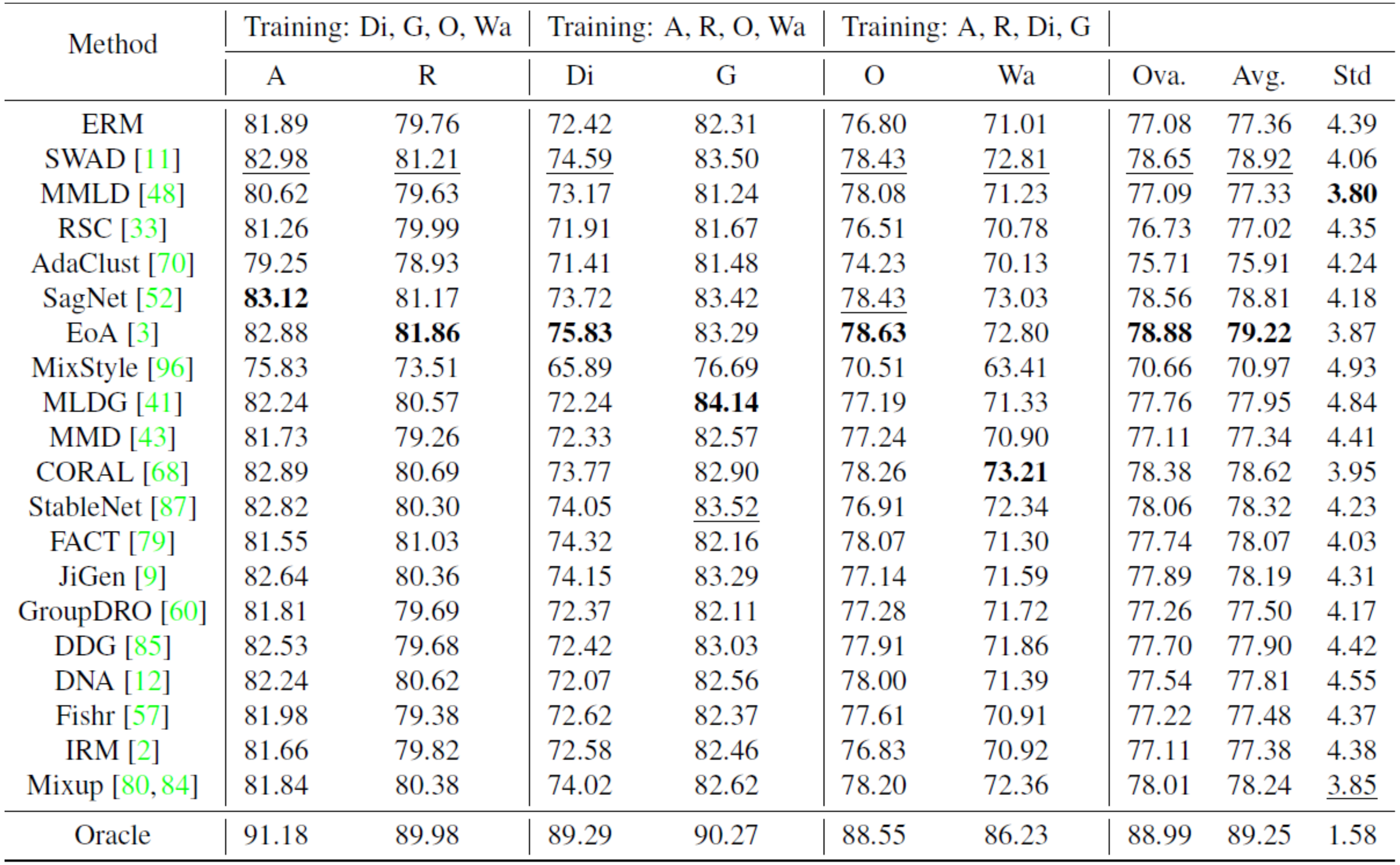
covariate shift between source and target data

concept shift between source and target data

- The covariate shift $p(x)$ reflects the difficulty of solving OOD problems.
- The concept shift indicates the labeling function shift between target and source data, which is not solvable using algorithms.
- NICO++ shows stronger covariate shift and lower concept shift compared with other OOD datasets.



Benchmark: Standard DG



Method	Training: Di, G, O, Wa		Training: A, R, O, Wa		Training: A, R, Di, G				
	A	R	Di	G	O	Wa	Ova.	Avg.	Std
ERM	81.89	79.76	72.42	82.31	76.80	71.01	77.08	77.36	4.39
SWAD [11]	82.98	81.21	74.59	83.50	78.43	72.81	78.65	78.92	4.06
MMLD [48]	80.62	79.63	73.17	81.24	78.08	71.23	77.09	77.33	**3.80**
RSC [33]	81.26	79.99	71.91	81.67	76.51	70.78	76.73	77.02	4.35
AdaClust [70]	79.25	78.93	71.41	81.48	74.23	70.13	75.71	75.91	4.24
SagNet [52]	**83.12**	81.17	73.72	83.42	78.43	73.03	78.56	78.81	4.18
EoA [3]	82.88	**81.86**	**75.83**	83.29	**78.63**	72.80	**78.88**	**79.22**	3.87
MixStyle [96]	75.83	73.51	65.89	76.69	70.51	63.41	70.66	70.97	4.93
MLDG [41]	82.24	80.57	72.24	**84.14**	77.19	71.33	77.76	77.95	4.84
MMD [43]	81.73	79.26	72.33	82.57	77.24	70.90	77.11	77.34	4.41
CORAL [68]	82.89	80.69	73.77	82.90	78.26	**73.21**	78.38	78.62	3.95
StableNet [87]	82.82	80.30	74.05	83.52	76.91	72.34	78.06	78.32	4.23
FACT [79]	81.55	81.03	74.32	82.16	78.07	71.30	77.74	78.07	4.03
JiGen [9]	82.64	80.36	74.15	83.29	77.14	71.59	77.89	78.19	4.31
GroupDRO [60]	81.81	79.69	72.37	82.11	77.28	71.72	77.26	77.50	4.17
DDG [85]	82.53	79.68	72.42	83.03	77.91	71.86	77.70	77.90	4.42
DNA [12]	82.24	80.62	72.07	82.56	78.00	71.39	77.54	77.81	4.55
Fishr [57]	81.98	79.38	72.62	82.37	77.61	70.91	77.22	77.48	4.37
IRM [2]	81.66	79.82	72.58	82.46	76.83	70.92	77.11	77.38	4.38
Mixup [80, 84]	81.84	80.38	74.02	82.62	78.20	72.36	78.01	78.24	3.85
Oracle	91.18	89.98	89.29	90.27	88.55	86.23	88.99	89.25	1.58

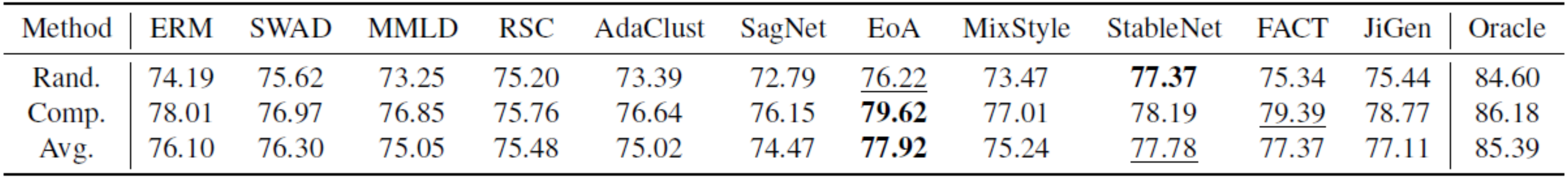
6 public common domains

- Split into 3 groups
  - Autumn, Rock
  - Dim, Grass
  - Outdoor, Water
- For each standard DG setting
  - 4 domains for training
  - 2 domains for test
- 4 private domains
  - Used for NICO Challenge
  - Will be released after more common domains are added this year

Results

- Current SOTA show their effectiveness
  - EoA, CORAL, StableNet…
- A gap between SOTA and oracle
  - Spacious room for improvement

Benchmark: Flexible DG



Method	ERM	SWAD	MMLD	RSC	AdaClust	SagNet	EoA	MixStyle	StableNet	FACT	JiGen	Oracle
Rand.	74.19	75.62	73.25	75.20	73.39	72.79	76.22	73.47	**77.37**	75.34	75.44	84.60
Comp.	78.01	76.97	76.85	75.76	76.64	76.15	**79.62**	77.01	78.19	79.39	78.77	86.18
Avg.	76.10	76.30	75.05	75.48	75.02	74.47	**77.92**	75.24	77.78	77.37	77.11	85.39

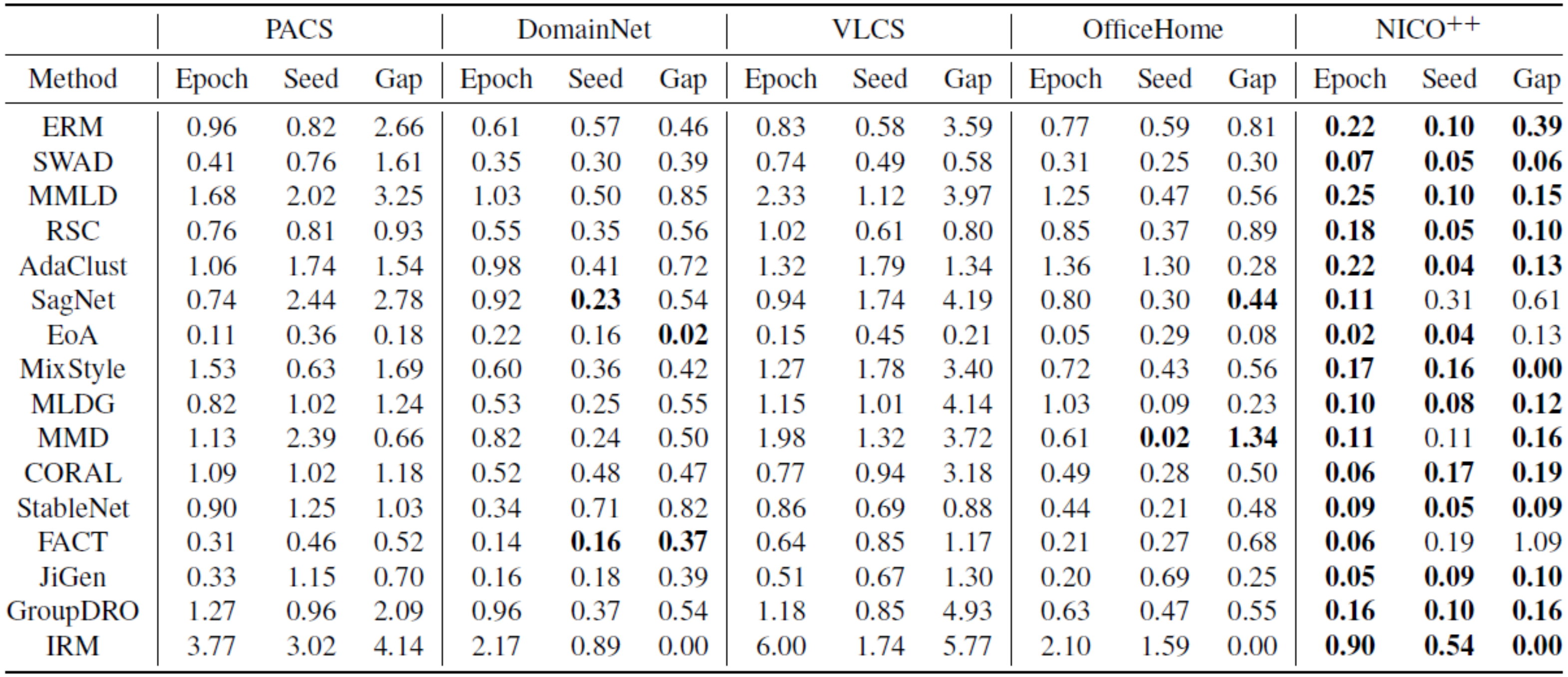
20 domains (10 common + 10 unique) for each category

- Compositional
  - 14 domains for training
    - 2 major domains: all images each domain
    - 12 minor domains: 50 images each domain
  - 6 domains for test
  - Major domains for one category can be the test domains for other categories
- Random
  - 4 fixed training domains across all categories
    - 2 major domains: all images each domain
    - 2 minor domains: 50 images each domain
  - Other 16 domains
    - 12 minor domains for training: 50 images each domain
    - 4 test domains for test
  - The two major domains for every category cannot be the test domains for any category

Results

- Current SOTA show their effectiveness
  - EoA, FACT, StableNet…
- A gap between SOTA and oracle
  - Spacious room for improvement

Test Variance and Model Selection



	PACS			DomainNet			VLCS			OfficeHome			NICO^{++}		
Method	Epoch	Seed	Gap	Epoch	Seed	Gap	Epoch	Seed	Gap	Epoch	Seed	Gap	Epoch	Seed	Gap
ERM	0.96	0.82	2.66	0.61	0.57	0.46	0.83	0.58	3.59	0.77	0.59	0.81	**0.22**	**0.10**	**0.39**
SWAD	0.41	0.76	1.61	0.35	0.30	0.39	0.74	0.49	0.58	0.31	0.25	0.30	**0.07**	**0.05**	**0.06**
MMLD	1.68	2.02	3.25	1.03	0.50	0.85	2.33	1.12	3.97	1.25	0.47	0.56	**0.25**	**0.10**	**0.15**
RSC	0.76	0.81	0.93	0.55	0.35	0.56	1.02	0.61	0.80	0.85	0.37	0.89	**0.18**	**0.05**	**0.10**
AdaClust	1.06	1.74	1.54	0.98	0.41	0.72	1.32	1.79	1.34	1.36	1.30	0.28	**0.22**	**0.04**	**0.13**
SagNet	0.74	2.44	2.78	0.92	**0.23**	0.54	0.94	1.74	4.19	0.80	0.30	**0.44**	**0.11**	0.31	0.61
EoA	0.11	0.36	0.18	0.22	0.16	**0.02**	0.15	0.45	0.21	0.05	0.29	0.08	**0.02**	**0.04**	0.13
MixStyle	1.53	0.63	1.69	0.60	0.36	0.42	1.27	1.78	3.40	0.72	0.43	0.56	**0.17**	**0.16**	**0.00**
MLDG	0.82	1.02	1.24	0.53	0.25	0.55	1.15	1.01	4.14	1.03	0.09	0.23	**0.10**	**0.08**	**0.12**
MMD	1.13	2.39	0.66	0.82	0.24	0.50	1.98	1.32	3.72	0.61	**0.02**	**1.34**	**0.11**	0.11	**0.16**
CORAL	1.09	1.02	1.18	0.52	0.48	0.47	0.77	0.94	3.18	0.49	0.28	0.50	**0.06**	**0.17**	**0.19**
StableNet	0.90	1.25	1.03	0.34	0.71	0.82	0.86	0.69	0.88	0.44	0.21	0.48	**0.09**	**0.05**	**0.09**
FACT	0.31	0.46	0.52	0.14	**0.16**	**0.37**	0.64	0.85	1.17	0.21	0.27	0.68	**0.06**	0.19	1.09
JiGen	0.33	1.15	0.70	0.16	0.18	0.39	0.51	0.67	1.30	0.20	0.69	0.25	**0.05**	**0.09**	**0.10**
GroupDRO	1.27	0.96	2.09	0.96	0.37	0.54	1.18	0.85	4.93	0.63	0.47	0.55	**0.16**	**0.10**	**0.16**
IRM	3.77	3.02	4.14	2.17	0.89	0.00	6.00	1.74	5.77	2.10	1.59	0.00	**0.90**	**0.54**	**0.00**

Potential shortcuts for model selection in DG

- Select the best hyperparameters via test performance
- Select the best epoch model checkpoint
- Select the best seed model checkpoint

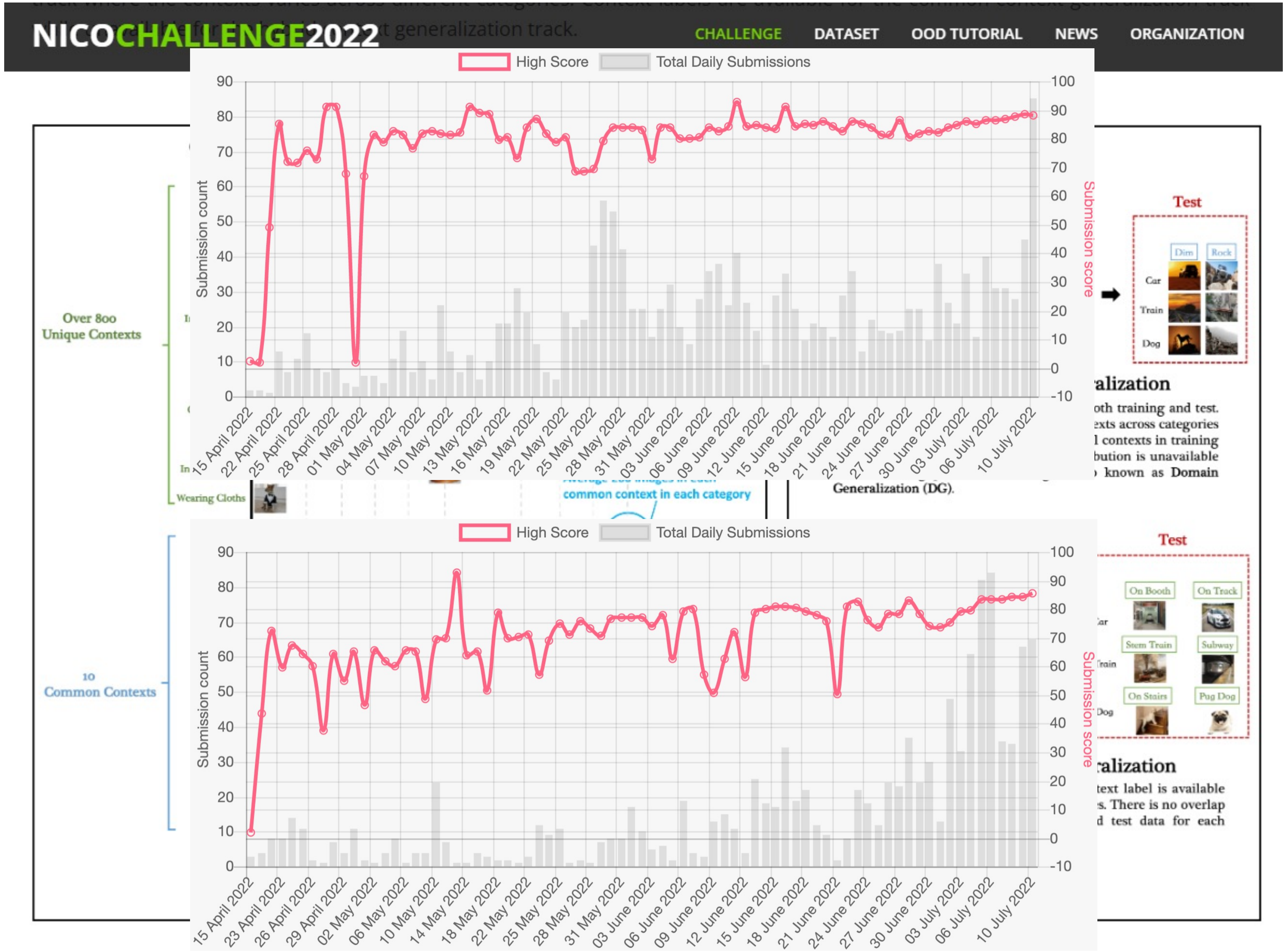
Metrics measuring the potential leakage

- Test variance across epochs
- Test variance across seeds
- Gap between standard model selection and oracle model selection

NICO++ squeezes all of them!   =>   A fairer comparison for DG

NICO Challenge 2022

https://nicochallenge.com/



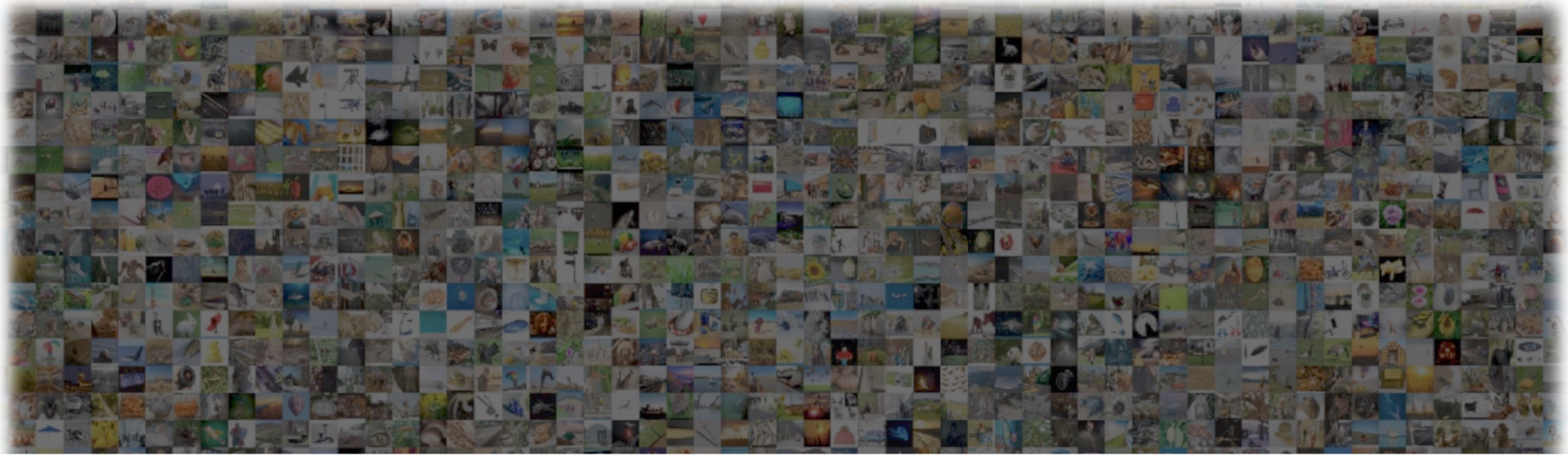
Statistics

- A total of 178 teams participated.
- Over 4,000 results submitted for public test.

NICO Challenge 2023 is coming!!!

More Categories *More Domains*

More Challenging Tracks



Join to solve visual problems in open applications!

Thanks!

Github link:
https://github.com/xxgege/NICO-plus

Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, Peng Cui. NICO++: Towards better bechmarks for Domain Generalization. CVPR, 2023.