VLP for Computer Vision in the Wild

Focused Topics: Knowledge & Benchmark

Chunyuan Li
Microsoft Research, Redmond
June 19, 2022

Zero-Shot Learning	Class-level Transfer	VL Pre-training → Task-level Transfer
Definition	Generalizing to **unseen object categories** eg, DeViSE ⚠ **Object classes** whose instances have not been observed during training	Generalizing to **unseen datasets/tasks** eg, CLIP ⚠ **Datasets** whose instances have not been observed during training
Modeling: External Knowledge	**Key**: Associate observed and non-observed classes through some form of **auxiliary information**: • **Implicit**: Pre-trained semantic embeddings • **Explicit**: Attributes, knowledge bases A rich line of research for decades	KLITE
Benchmark	• Animal with Attributes (AwA) • Caltech-UCSD Birds-200 (CUB), • SUN, • aPY, • ZS-ImageNet	ELEVATER

KLITE:

Learning Transferable Visual Models with External Knowledge

https://arxiv.org/abs/2204.08790

- Image Classification
- Object Detection

Sheng Shen[*♮], Chunyuan Li[*†♠], Xiaowei Hu[*†], Yujia Xie[†], Jianwei Yang[†]
Pengchuan Zhang[†], Anna Rohrbach[♮], Zhe Gan[†], Lijuan Wang[†], Lu Yuan[†]
Ce Liu[†], Kurt Keutzer[♮], Trevor Darrell[♮], Jianfeng Gao[†]
[†]Microsoft [♮]University of California, Berkeley

Motivating Scenarios

First time to a Japanese restaurant?

1. Language:
   Hard to understand the menu by looking at dish names
2. Knowledge:
   Waitress explains it with her knowledge
3. Image:
   Dishes served with the best fit

Takoyaki

A ***ball-shaped*** Japanese ***dumpling*** made of batter, filled with diced octopus, ***tempura scraps***, pickled ginger, and ***green onion***.

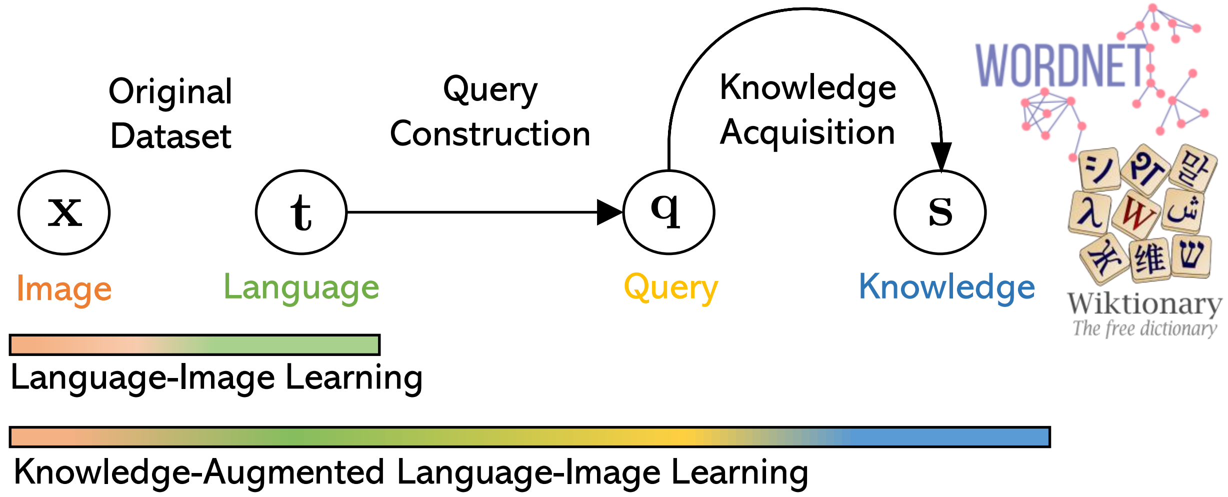
Sashimi

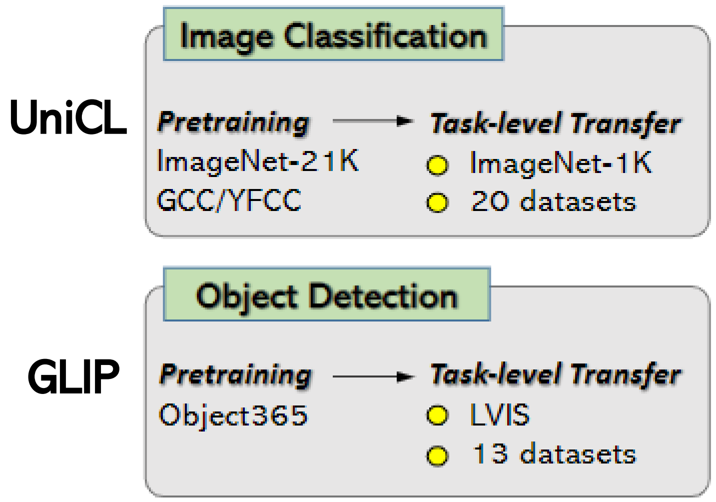
A dish consisting of ***thin slices*** or pieces of ***raw fish or meat***.

Q: How humans generalize to novel concepts?
A: Instead of trying to memorize all concepts, humans leverage the structured knowledge

Idea: External knowledge is generally available for a variety of domains (eg, textbooks, databases), Can we leverage them to build a systematic and generic approach for task-level visual transfer?

K-Lite: Knowledge-augmented Language Image Training and Evaluation





Order sashimi from Oishi Japanese Restaurant for delivery or take-out!

sashimi

1. WordNet Hierarchy:
   [sashimi, dish, nutriment, food, substance, matter, physical_entity, entity]
2. WordNet Definition:
   very thinly sliced raw fish
3. Wiktionary Definition:
   A dish consisting of thin slices or pieces of raw fish or meat.

The knowledge-augmentation process in executed in two phases:

- **Training**: the model is endowed with an ability to read and understand a specific knowledge source
- **Evaluation**: knowledge provides an additional information source to enhance model inference

Image Classification (Zero & Few-shot Task Transfer to ImageNet-1K and 20 public datasets)

Baseline: UniCL is the academic version of **Microsoft Florence**, trained on large public datasets

Training Data		Method	ImageNet-1K	20 datasets	
Dataset	# Samples		Zero-shot	Zero-shot	Linear Probing
ImageNet-21K	13M (full)	UniCL	28.16	27.15	53.07 $\pm$ 4.15
	13M (full)	K-LITE	**30.23**	**33.44**	**53.92** $\pm$ **1.05**
GCC-15M + ImageNet-21K	15M (half)	UniCL	41.64	36.31	53.86 $\pm$ 2.73
	15M (half)	K-LITE	44.26	39.53	55.91 $\pm$ 2.53
	15M (half)	K-LITE$^\diamond$	47.30	40.32	57.38 $\pm$ 2.70
	28M (full)	UniCL	46.83	38.90	57.92 $\pm$ 3.31
	28M (full)	K-LITE	**48.76**	**41.34**	**58.56** $\pm$ **3.12**

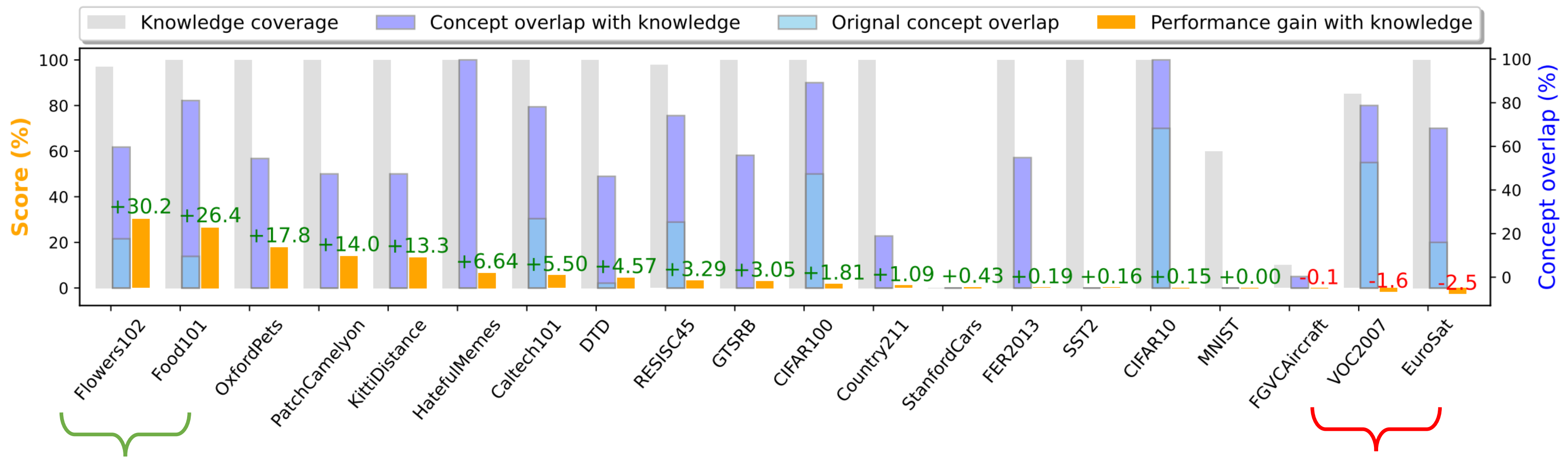
Sample-efficiency in Pre-training:

When scaled up to the largest academic datasets, K-LITE achieves the prior best performance with only a half number of pre-training image-text pairs

UniCL: Unified contrastive learning in image-label-text space, CVPR 2022

Florence: A new foundation model for computer vision

Image Classification (Why does external knowledge help zero-shot transfer?)



✓ **English marigold**: Any of the Old World plants, of the genus Calendula, with orange, yellow or reddish flowers.

✗ **Wallflower**: Any of several short-lived herbs or shrubs of the Erysimum genus with bright yellow to red flowers.

✓ **Lobster bisque**: A thick creamy soup made from fish, shellfish, meat or vegetables.

✗ **Hot and sour soup**: Any one of several soups, served in various Asian cuisines, which are both spicy and sour

✗ **bus**: A motor vehicle for transporting large numbers of people along roads.

✓ **car**: A wheeled vehicle that moves independently, with at least three wheels

annual crop land: arable land

permanent crop land: arable land

Knowledge **benefits** the most, when

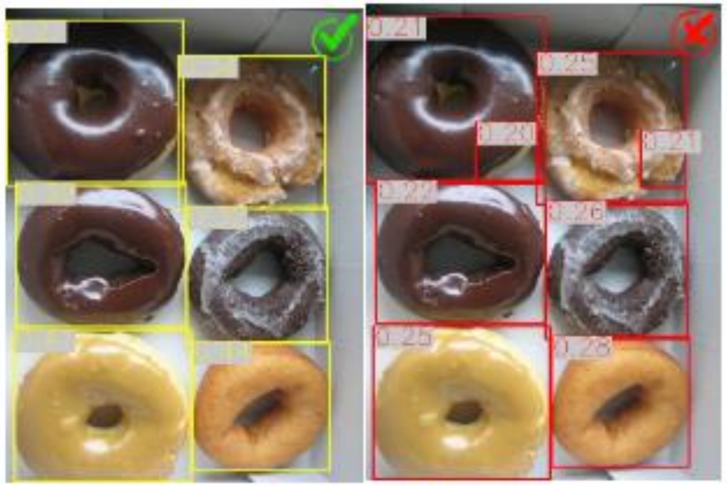
- Fine-grained datasets, which typically present many rare concepts

Knowledge **hurts** performance, when

- Knowledge coverage/quality is low
- Spurious words are contained

Object Detection (Zero-shot Task Transfer to LVIS and 13 datasets)

Method	LVIS								13 datasets			
	APr	APc	APf	-	$\mathcal{S}_{\text{LVIS}}$	$\mathcal{S}_{\text{wn\_path}}$	$\mathcal{S}_{\text{wn\_def}}$	$\mathcal{S}_{\text{wiki\_def}}$	-	$\mathcal{S}_{\text{wn\_path}}$	$\mathcal{S}_{\text{wn\_def}}$	$\mathcal{S}_{\text{wiki\_def}}$
GLIP-A [49]	14.2	13.9	23.4	18.5	-	-	-	-	28.8	-	-	-
Baseline GLIP$^{\heartsuit}$	8.6	14.0	23.1	17.9	17.6	17.1	17.2	15.0	27.5	26.8	21.0	18.5
K-LITE	14.8	18.6	24.8	16.9	**21.3**	18.7	21.4	20.5	25.0	30.3	28.4	**31.7**



doughnut: a small ring-shaped friedcake
bun: small rounded bread either plain or sweet

GLIP: Grounded Language-Image Pre-training, CVPR 2022

ELEVATER:

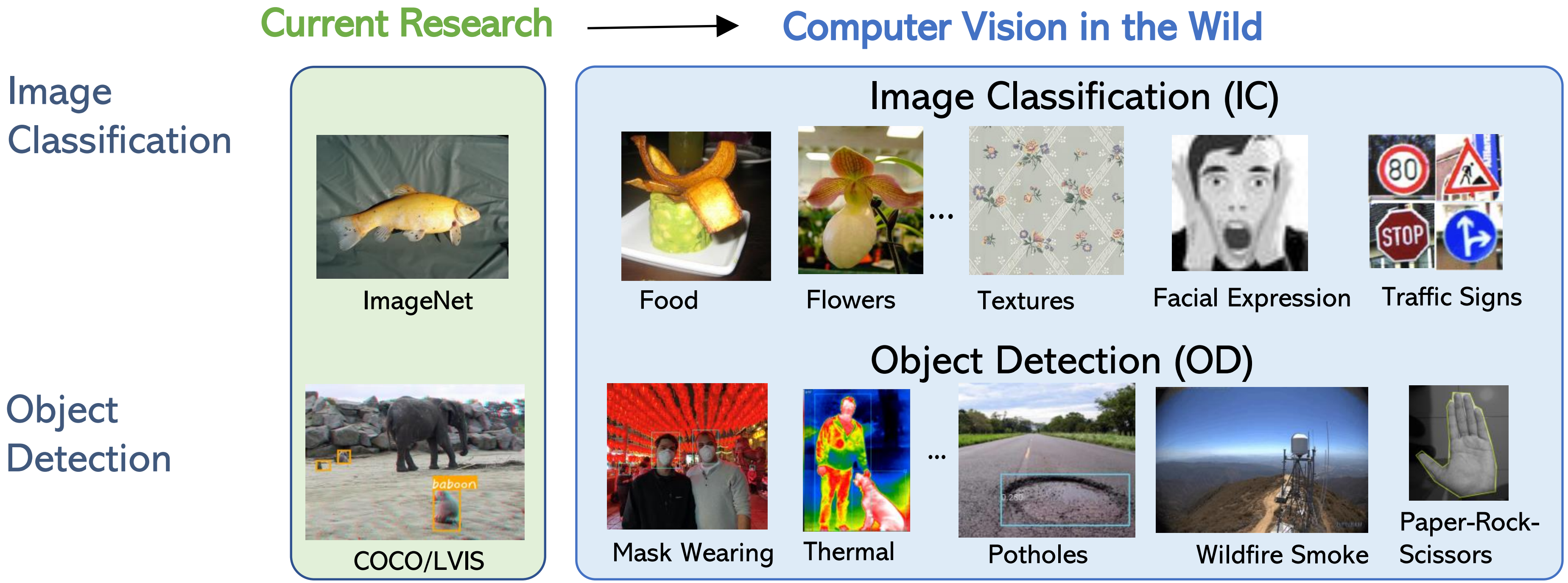
A Benchmark and Toolkit for Evaluating Language-Augmented Visual Models

https://arxiv.org/abs/2204.08790

- Data
- Toolkit
- Evaluation Playgrounds

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

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



Trend

- Building transferable systems that can effortlessly adapt to a wide range of CV tasks in the wild
- 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

Benchmarks: ELEVATER

- Dataset Suite

- Image Classification: **20** datasets

- Object Detection: **35** datasets

- External Knowledge

WordNet, Wiktionary, GPT-3

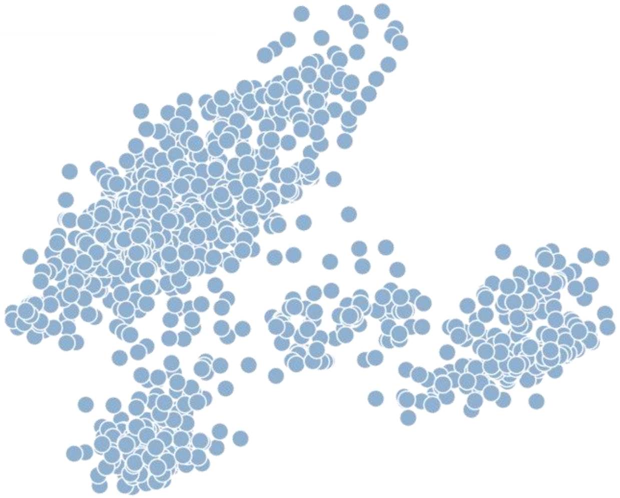
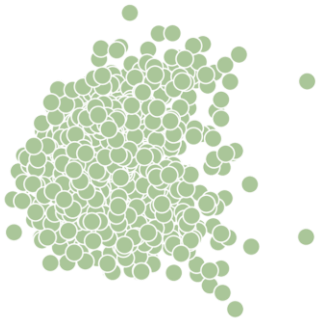
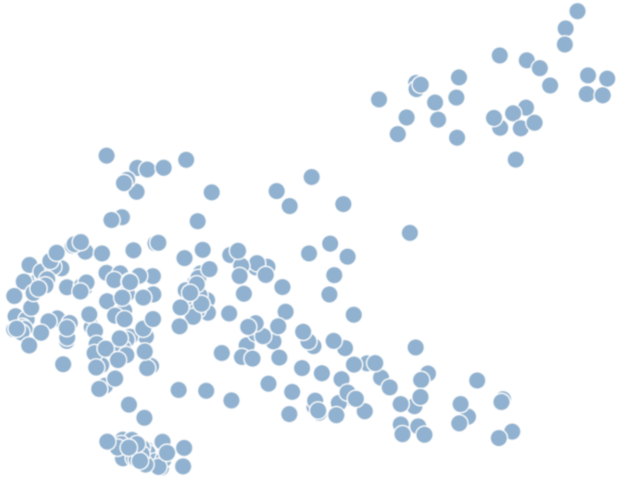
- Concept name: risotto
- Def_wik: An Italian savoury dish made with rice and other ingredients
- Def_wn: rice cooked with broth and sprinkled with grated cheese
- Path_wn: [risotto, dish, nutriment, food, substance, matter, physical_entity, entity]
- GPT3: A rice dish made with arborio rice and typically served with meat or fish

Benchmarks: Review

Problem		Benchmark Statistics				Evaluation Settings		
		#Datasets	#Image	#Concepts	Knowledge Source	Zero	Few	Full
IC	AwA [31]	1	30337 / 6985	40 / 10	Attributes	✓		
	CUB [64]	1	8855 / 2933	150 / 50	Attributes	✓		
	SUN [47]	1	12900 / 1440	645 / 72	Attributes	✓		
	aPY [14]	1	12695 / 2644	20 / 12	Attributes	✓		
	ZS-ImageNet [52]	1	1.2M / 54K	1K / 360	WordNet	✓		
	ImageNet-1K [9]	1	1.2M / 50K	1K	WordNet	✓		✓
	VTAB [73]	19	2.2M / -	940	-		✓	✓
	ELEVATER (Ours)	20	638K / 193K	1151$^{\diamond}$	WordNet, Wiki, GPT-3	✓	✓	✓
OD	LVIS [21]	1	120k / 40K	1723	WordNet			✓
	ELEVATER (Ours)	35	132K / 20K	314$^{\diamond}$	WordNet, Wiki, GPT-3	✓	✓	✓

- **Trend**: Benchmarking fast-developing techniques **from class-level transfer to task-level transfer**, with language-augmented visual models

Benchmarks: A more diverse set of tasks

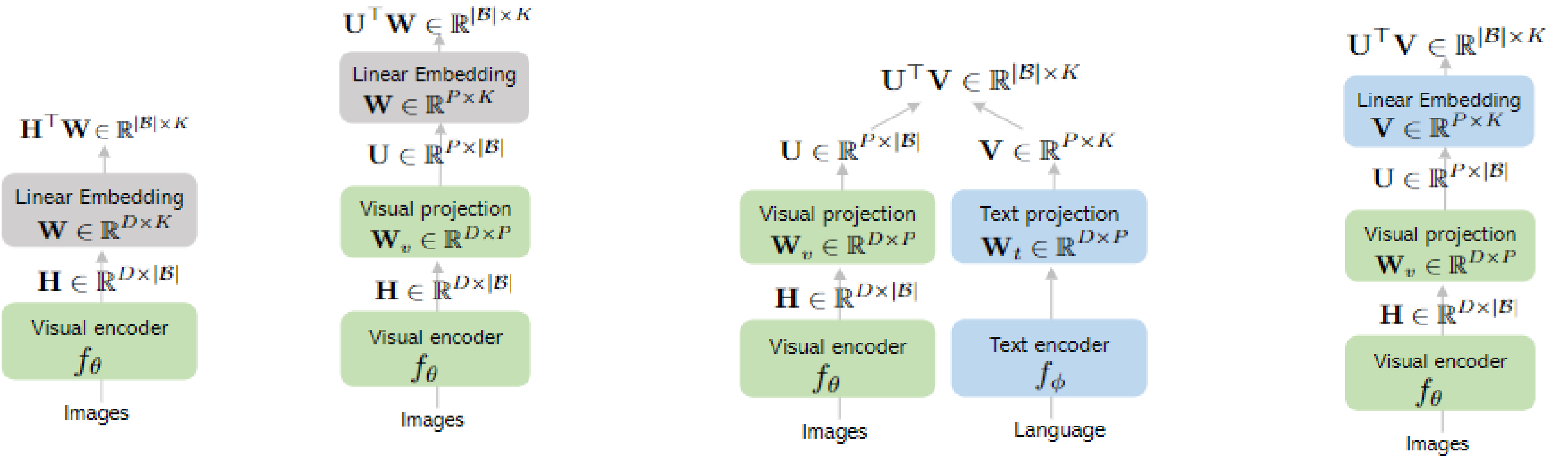
	ImageNet →	Image Classification in the Wild (20 datasets)	LVIS →	Object Detection in the Wild (35 datasets)
Semantic Space using PCA				
Task diversity using std	0.610	0.680	0.533	0.619

Toolkits

- **Automatic hyper-parameter tuning pipeline** (eg, learning rate, weight decay)

  Avoid human-in-the-loop tuning

- **Language-augmented model adaptation methods**



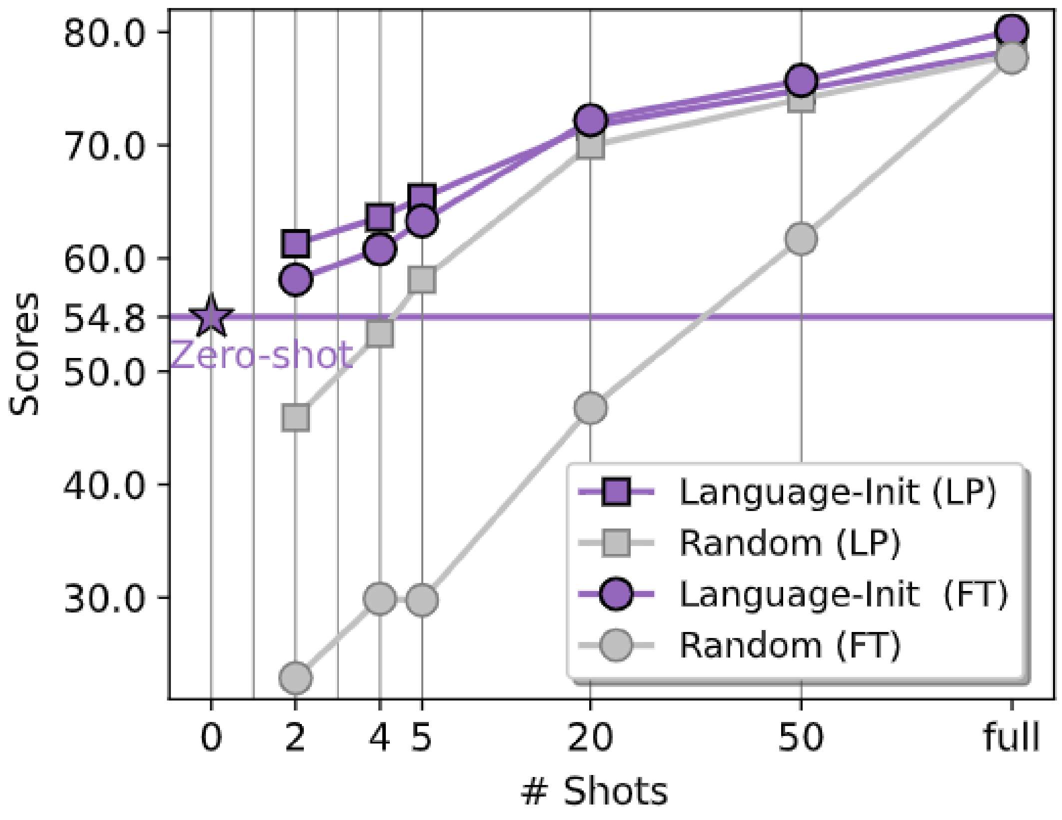
(a) IC Adaptation (b) CLIP Adaptation (c) CLIP Zero-shot (d) Language-Init

Missing language encoder power

Leverage language or knowledge

Toolkits

- **Effectiveness of Language-initialized model adaptation methods**



- Language-initialized strategy consistently improves the baseline random initialization, for both linear probing (LP) and fine-tuning (FT)

- **Few-shot performance is always better than zero-shot**, in contrast to the discovery in original CLIP paper

Baseline Pre-trained Visual Models

	Checkpoints	Taxonomy		Pre-training Settings	
		Language	Knowledge	Training Objective	Dataset
	Image Classification				
Language-free	MoCo-v3 [6]	✗	✗	Self-Supervised	ImageNet-1K (1.2M)
	MAE [23]	✗	✗	Self-Supervised	ImageNet-1K (1.2M)
	DeiT [66]	✗	✗	Supervised	ImageNet-1K (1.2M)
	ViT [13]	✗	✗	Supervised	ImageNet-22K (14M)
Language-augmented: Knowledge-free	CLIP [55]	✓	✗	Image-Text Contrast	WebImageText (400M)
	UniCL [73]	✓	✗	Image-Text Contrast	ImageNet-21K (13M)
Language-augmented: Knowledge-augmented	K-LITE [59]	✓	✓	Image-Text Contrast	ImageNet-21K (13M)
	Object Detection				
Language-free	DyHead [9]	✗	✗	Supervised	Object365
Language-augmented: Knowledge-free	GLIP [38]	✓	✗	Supervised	Object365 & Grounding
	GLIP-A [38]	✓	✗	Supervised	Object365
Language-augmented: Knowledge-augmented	K-LITE [59]	✓	✓	Supervised	Object365

Benchmarking Pre-trained Visual Models

Pre-training Settings			20 Image Classification Datasets			
Checkpoint	Method	Dataset	5-shot	20-shot	50-shot	Full-shot
Linear Probing						
CLIP‡	Image-Text Contrast	WebImageText (400M)	68.27 $\pm$ 0.97	74.76 $\pm$ 1.11	77.75 $\pm$ 0.81	81.17
ViT†	Supervised	ImageNet-22K (14M)	57.61 $\pm$ 3.62	69.93 $\pm$ 0.71	73.74 $\pm$ 0.79	77.60
DeiT	Supervised	ImageNet-1K (1.2M)	54.06 $\pm$ 3.02	68.57 $\pm$ 3.43	75.53 $\pm$ 0.72	79.56
MAE	Self-Supervised	ImageNet-1K (1.2M)	33.37 $\pm$ 1.98	48.03 $\pm$ 2.70	58.26 $\pm$ 0.84	68.70
MoCo-v3	Self-Supervised	ImageNet-1K (1.2M)	50.17 $\pm$ 3.43	61.99 $\pm$ 2.51	69.71 $\pm$ 1.03	74.92
Fine-tuning						
CLIP‡	Image-Text Contrast	WebImageText (400M)	69.12 $\pm$ 1.66	74.76 $\pm$ 2.34	78.21 $\pm$ 2.04	83.63
ViT†	Supervised	ImageNet-22K (14M)	57.18 $\pm$ 2.02	72.45 $\pm$ 2.85	78.53 $\pm$ 0.69	82.02
DeiT	Supervised	ImageNet-1K (1.2M)	54.06 $\pm$ 3.02	68.53 $\pm$ 3.47	75.57 $\pm$ 0.68	79.55
MAE	Self-Supervised	ImageNet-1K (1.2M)	36.10 $\pm$ 3.25	54.13 $\pm$ 3.86	65.86 $\pm$ 2.42	74.43
MoCo-v3	Self-Supervised	ImageNet-1K (1.2M)	39.30 $\pm$ 3.84	58.75 $\pm$ 5.55	70.33 $\pm$ 1.64	77.71

Image Classification Performance Ranking: MAE < MoCo < DeiT < ViT < CLIP

Note: The conclusion is obtained using our auto-tuning adaptation process, **without** model / dataset –specific tuning for the best performance

Benchmarking Pre-trained Visual Models

Industry Track

- Scaling success

Academic Track

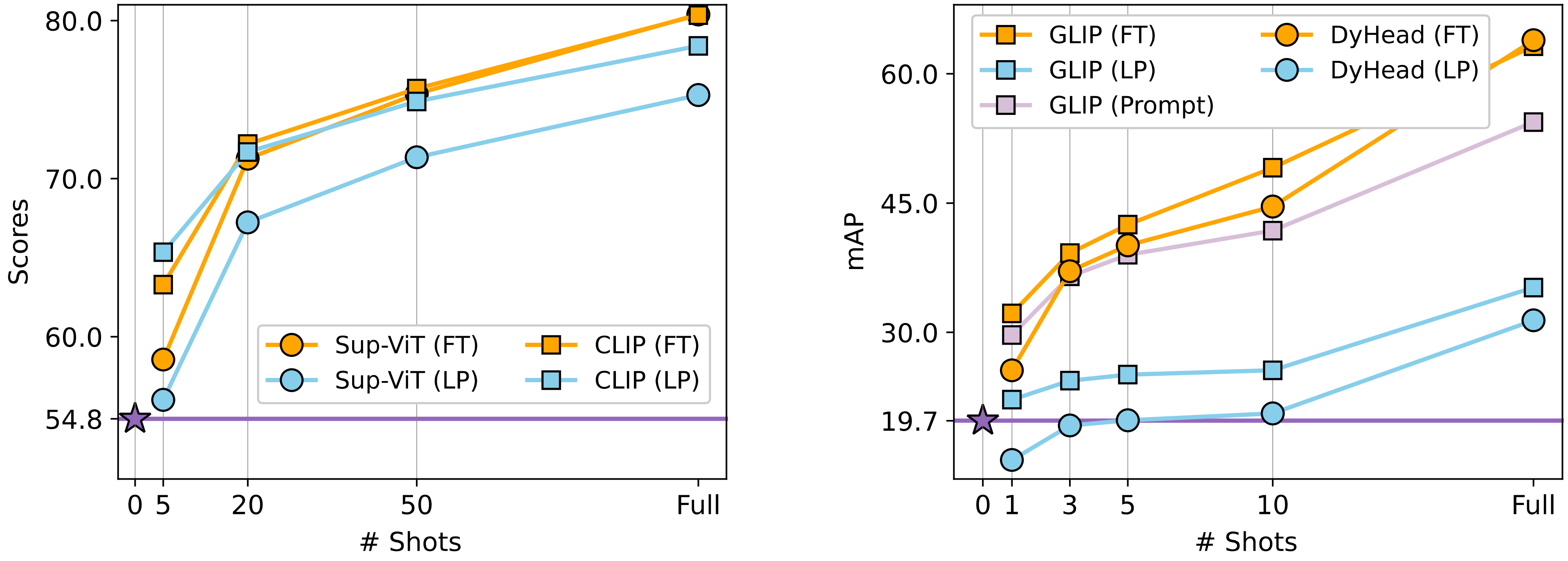
- Research Innovation
- limited to large public datasets

Settings: Checkpoint	Settings: Adaptation	Init Method	20 IC datasets: Zero-shot	20 IC datasets: Few-shot (5, 20, 50)	20 IC datasets: Full
Industry Track *(No pre-train data scale limit)*					
CLIP (ViT-B32)	LP	Random	56.64	58.09 $\pm$ 2.80, 69.97 $\pm$ 1.30, 74.09 $\pm$ 0.69	78.38
	LP	Language-S		65.35 $\pm$ 1.24, 71.69 $\pm$ 0.93, 74.89 $\pm$ 0.79	78.40
	LP	Language-M		65.88 $\pm$ 0.79, 72.05 $\pm$ 0.85, 75.08 $\pm$ 0.73	78.96
	FT	Random		29.75 $\pm$ 6.64, 46.76 $\pm$ 11.9, 61.70 $\pm$ 9.97	77.77
	FT	Language-S		63.29 $\pm$ 3.18, 72.19 $\pm$ 1.31, 75.70 $\pm$ 1.14	80.35
Supervised (ViT-B32)	LP	Random	-	56.00 $\pm$ 2.67, 67.23 $\pm$ 1.66, 71.35 $\pm$ 1.17	75.29
	FT	Random		58.55 $\pm$ 2.58, 71.27 $\pm$ 1.25, 75.36 $\pm$ 1.42	80.39
Academic Track *(Pre-trained on large established public datasets)*					
UniCL (Swin-Tiny)	LP	Language-S	27.15	54.31 $\pm$ 4.15, 66.42 $\pm$ 2.08, 70.49 $\pm$ 1.01	74.75
	FT	Language-S		44.75 $\pm$ 5.42, 56.53 $\pm$ 5.37, 67.90 $\pm$ 5.31	78.48
K-LITE (Swin-Tiny)	LP	Language-S	33.44	55.06 $\pm$ 2.36, 66.26 $\pm$ 1.56, 70.16 $\pm$ 1.09	74.47
	FT	Language-S		48.41 $\pm$ 2.84, 58.06 $\pm$ 4.30, 71.66 $\pm$ 2.02	78.05

Language-free vs Language-augmented

- Language-augmented model (CLIP) consistently outperforms language-free model (Supervised ViT) in most settings, especially for the limited data settings.
- language-augmented models enables zero-shot task transfer

Playground I: Sample-efficiency



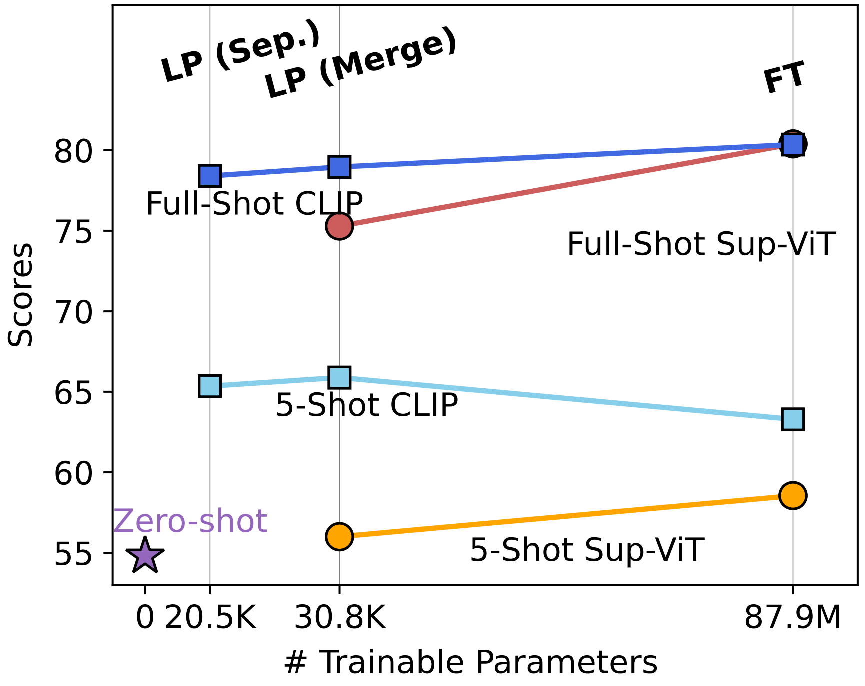
Zero-shot and Few-shot

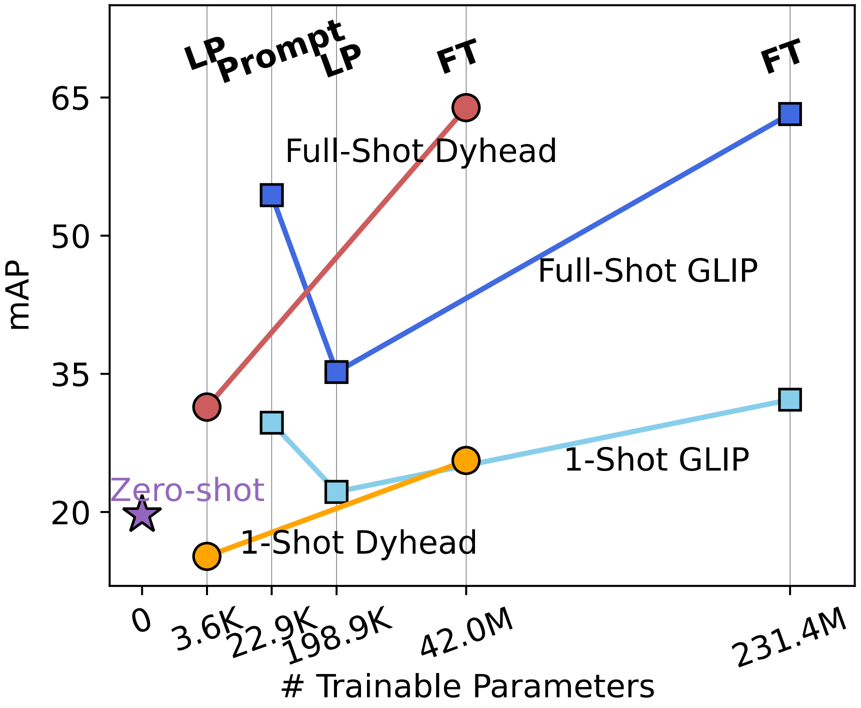
- Quick and more afford settings to assess pre-trained model quality

Existing & new VLP-for-CV models are welcome

- AGLIN, CoCa, DeCLIP, FILIP, SLIP, OpenCLIP, etc.

Playground II: Parameter-efficiency





Few-shot with VLP

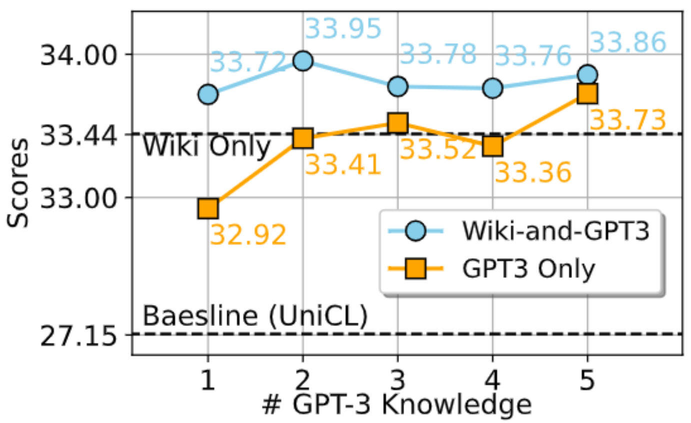
- A more meaning setting to study parameter-efficient methods

Existing & new VLP parameter-efficient model adaptation methods are welcome

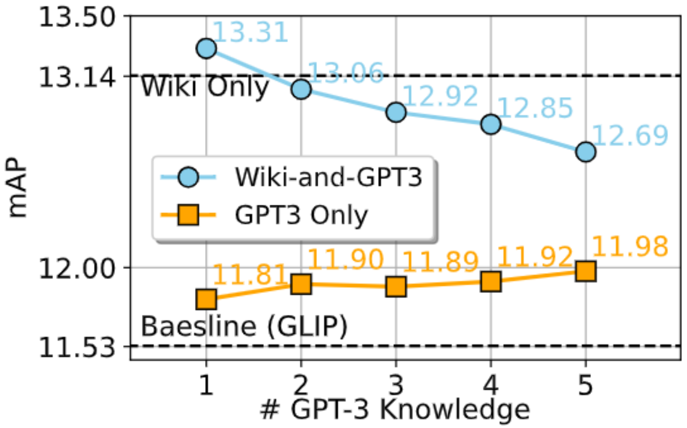
- Color Prompt Tuning, VL-Adapter, CLIP-Adapter, Conditional Prompt Learning etc.

Playground III: Benefit of external knowledge

- Combination of both explicit and implicit knowledge sources improves performance (K-Lite vs UniCL/GLIP)



(a) Image classification



(b) Object detection

- The collected knowledge benefits knowledge-free pre-trained models (eg, CLIP)

Adaptation Methods	5-shot LP	5-shot FT	Full-shot LP	Full-shot FT
Knowledge-free adaptation	65.35 $\pm$ 1.24	63.29 $\pm$ 3.18	78.40	79.97
Knowledge-augmented adaptation	**65.83** $\pm$ 1.50	**65.10** $\pm$ 2.08	**78.75**	**80.32**
Gain	+0.48	+1.81	+0.35	+0.35
# win / tie / lose	7 / 8 / 5	8 / 8 / 4	12 / 4 / 4	10 / 5 / 5

ECCV
TEL AVIV 2022

ECCV Workshop & Challenges: Computer Vision in the Wild

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

Benchmark Website:

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

Advisory Committee


Trevor Darrell


Lei Zhang


Jenq-Neng Hwang


Yong Jae Lee


Ce Liu


Xuedong Huang

Organizers


Pengchuan Zhang, Microsoft


Chunyuan Li, Microsoft


Jyoti Aneja, Microsoft


Ping Jin, Microsoft


Jianwei Yang, Microsoft


Xin Wang, Microsoft


Houdong Hu, Microsoft


Zicheng Liu, Microsoft


Haotian Liu, Univ. of Wisconsin at Madison


Liunian Li, UCLA


Kai-Wei Chang, UCLA


Jianfeng Gao, Microsoft

Thanks

Next: Text-to-Image Generation