An Efficient Alternative Framework for Generalized Category Discovery with Spatial Prompt Tuning

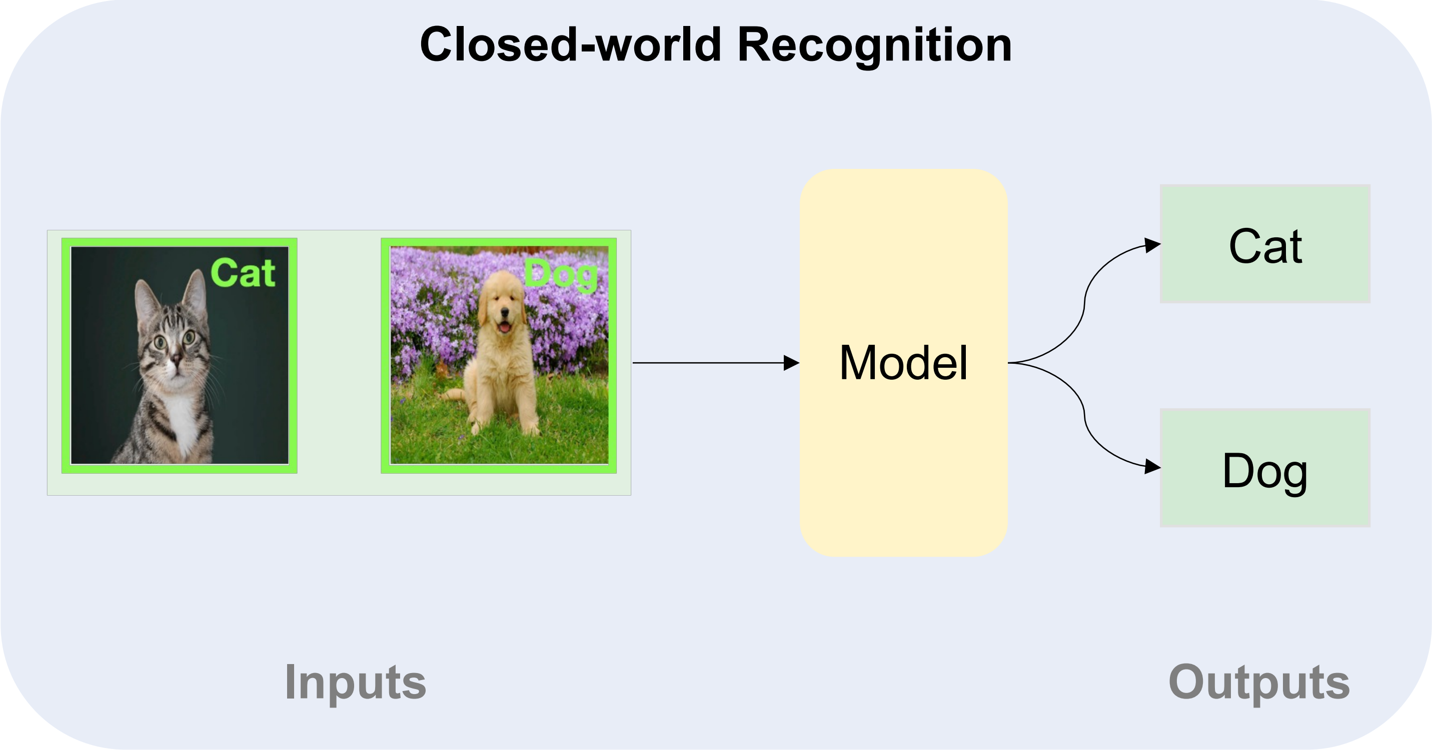
Hongjun Wang, Sagar Vaze, Kai Han

Contents

Introduction – Background

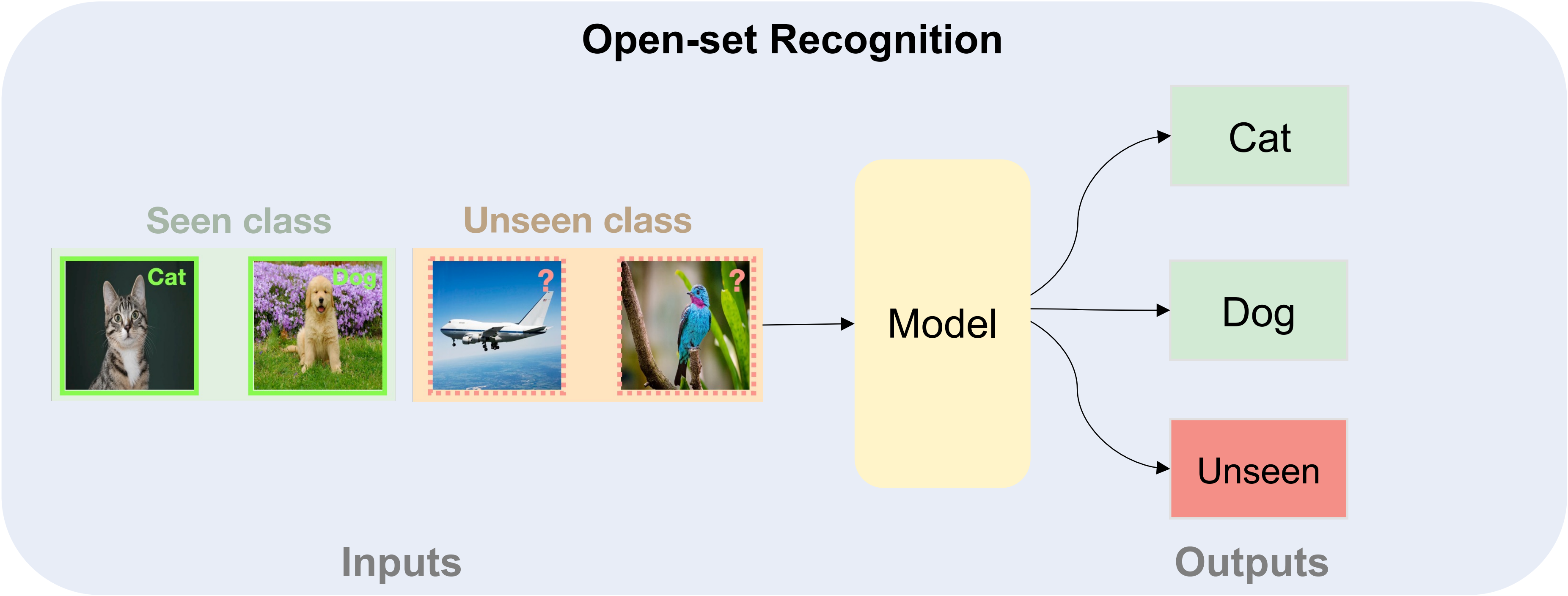
- **Closed-world Recognition**

Closed-world Recognition is the task of categorize the classes appearing in the training set.



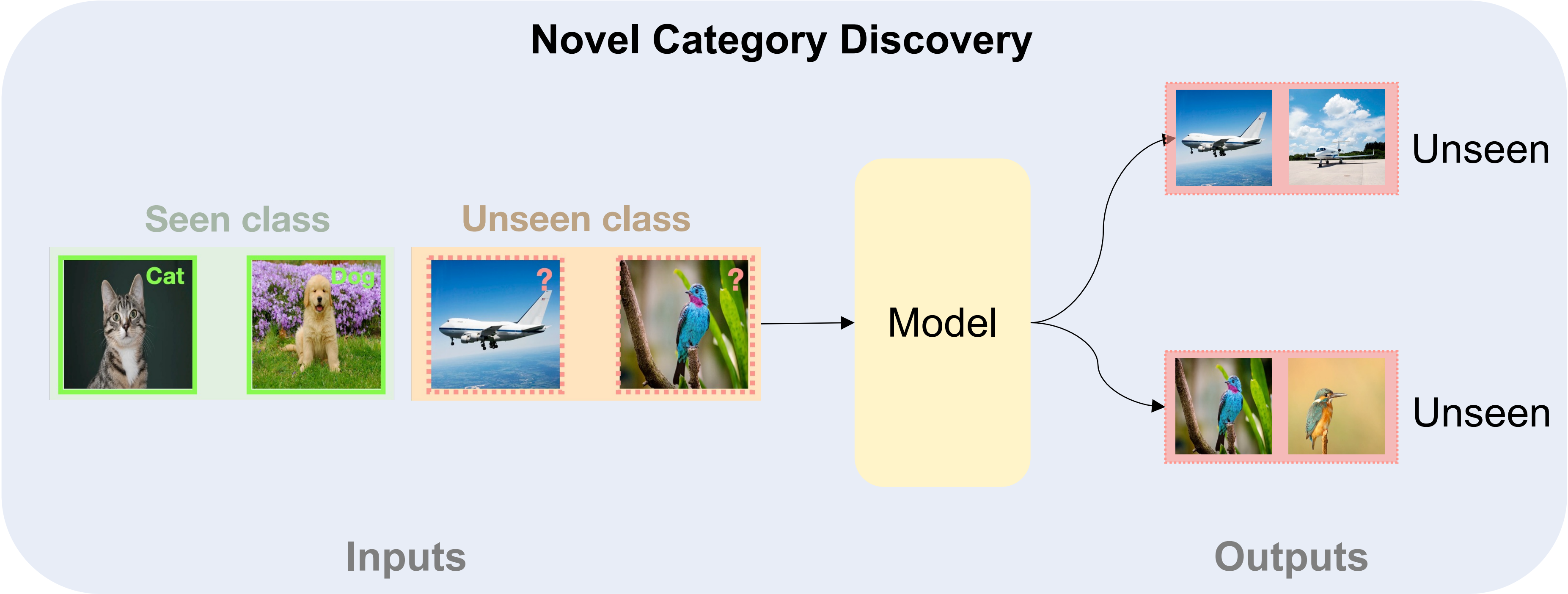
- **Open-set Recognition**

Open-set Recognition is the task of detecting whether a **test-time** image comes from a previously **'unseen' class**.



- **Novel Category Discovery**

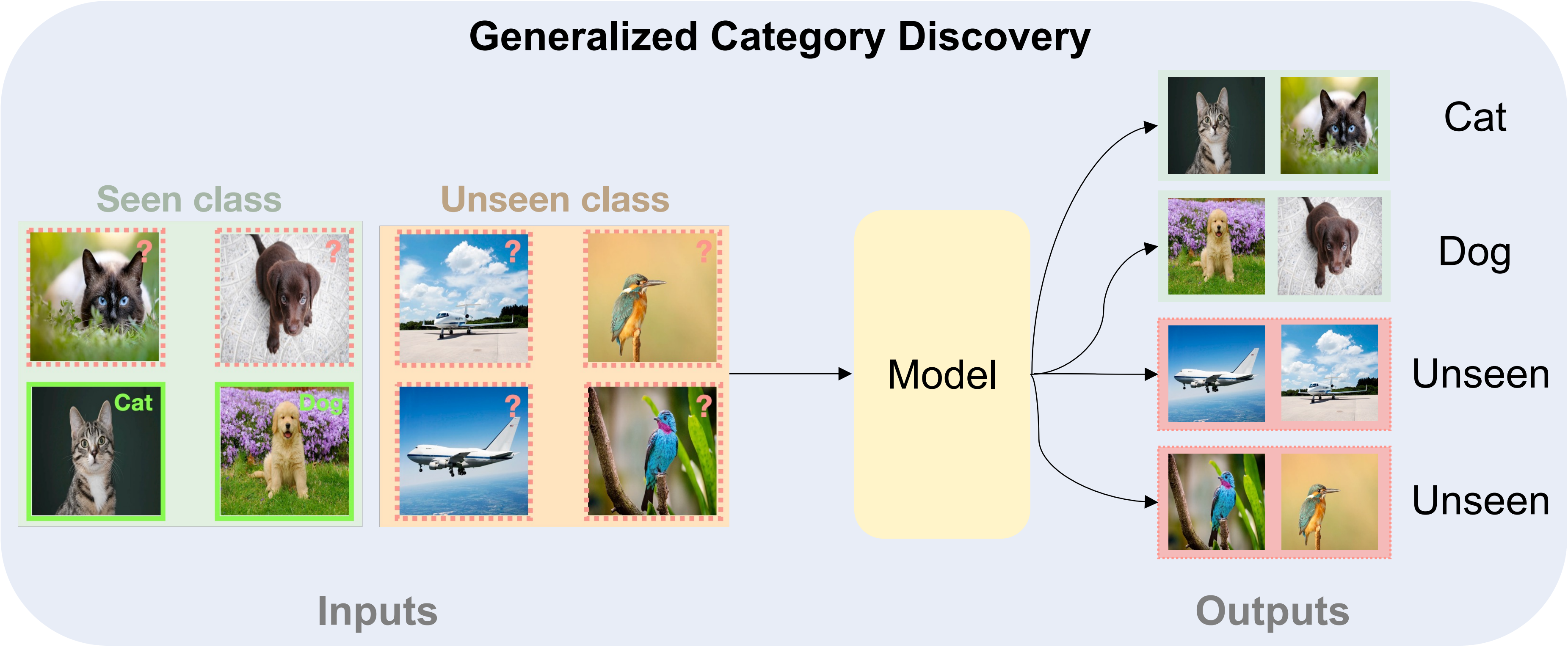
Novel Category Discovery (NCD) is the task of categorizing unlabelled images from **unseen classes** by transferring knowledge from **labelled data of seen classes**.



Introduction – Background

- **Generalized Category Discovery**

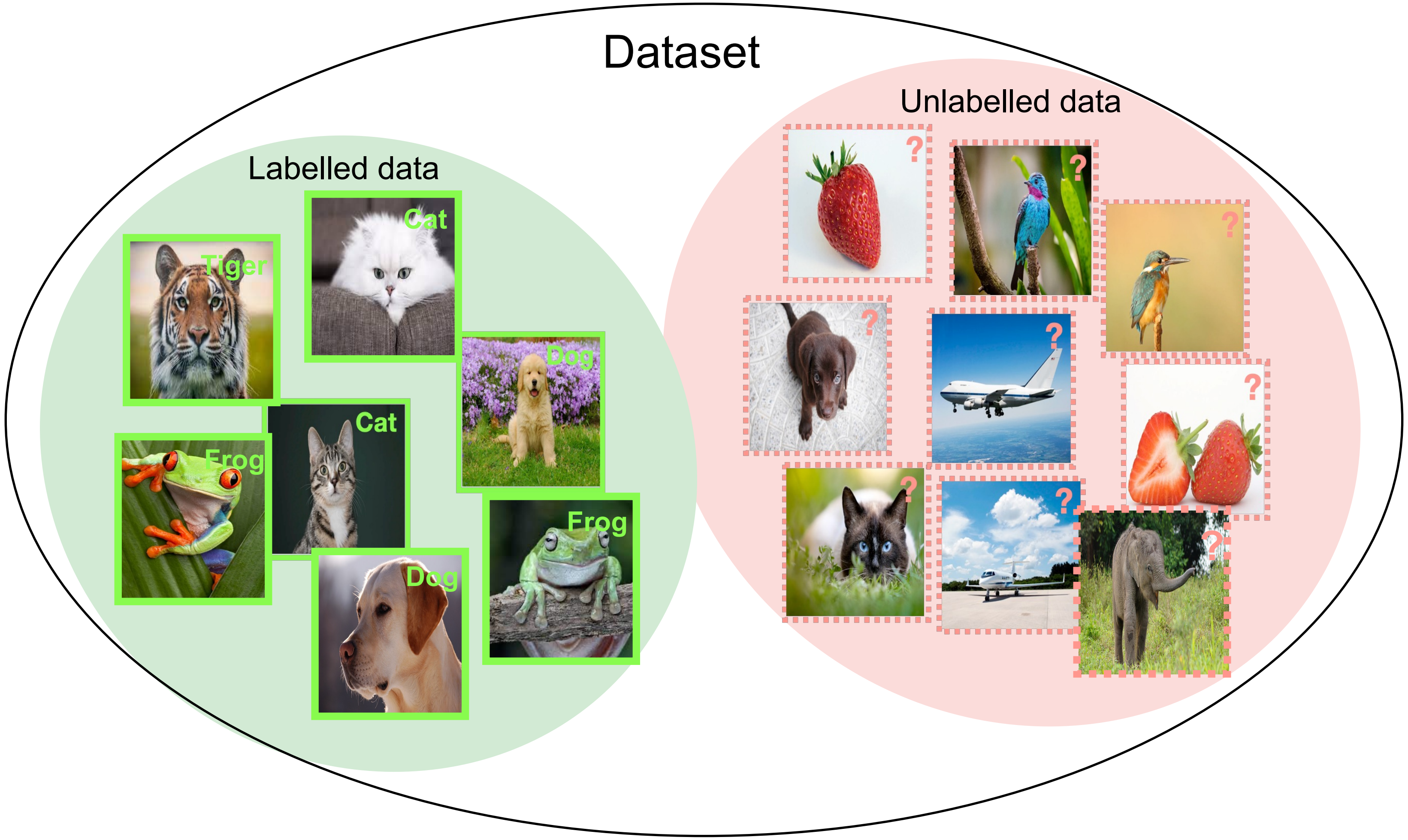
Generalized Category Discovery (GCD) extends NCD by categorizing unlabelled images from **both seen and unseen categories**.



Introduction – Background

Problem statement

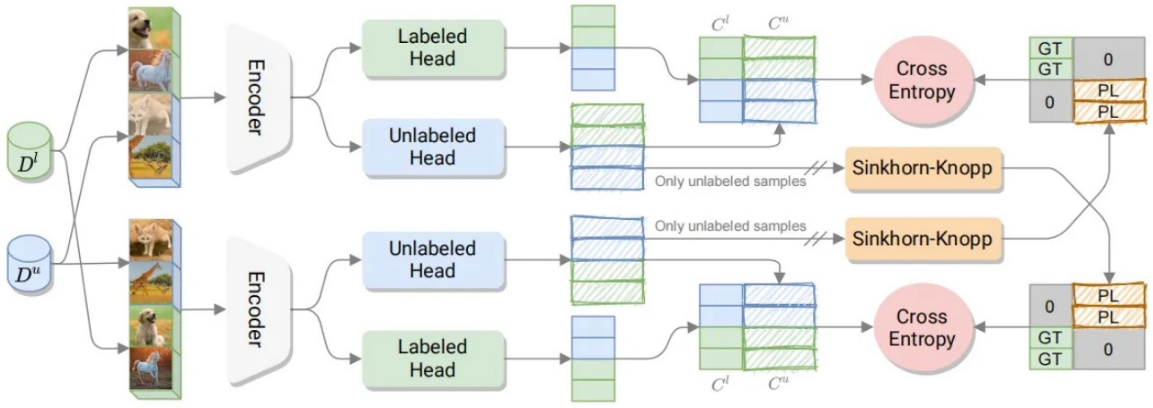
Given a dataset, a subset of which has class labels, categorize all unlabelled images in the dataset.



Introduction – Literature review

GCD baselines modified from NCD

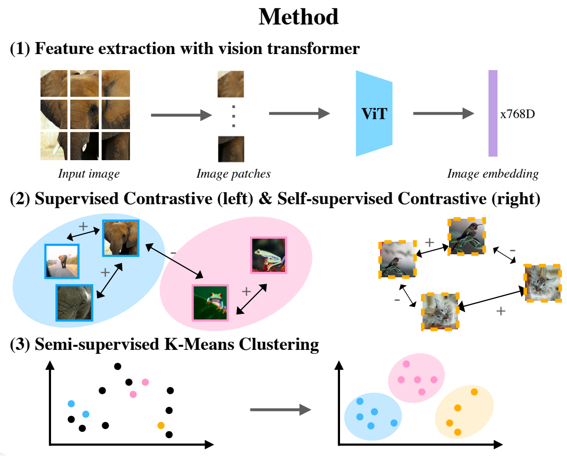
Han et al. (TPAMI'21)



Fini et al. (ICCV'21)

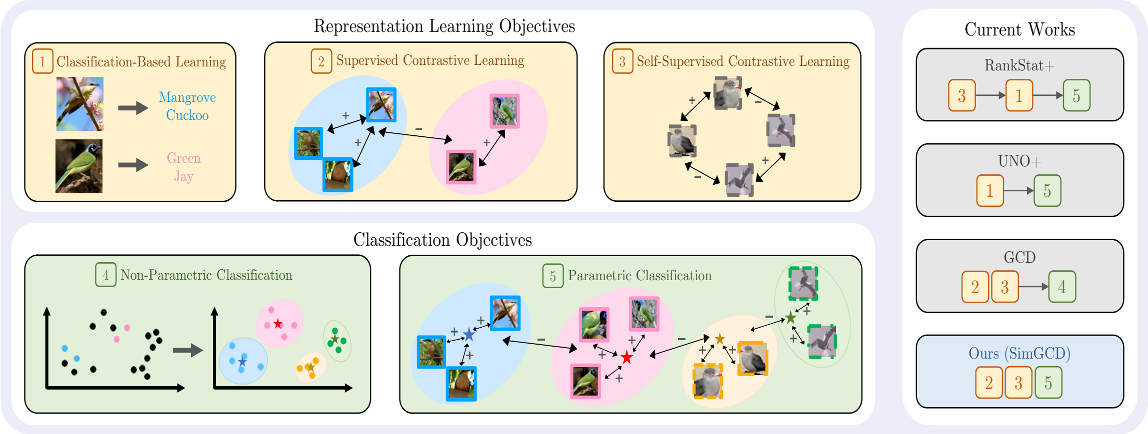
GCD baselines

Non-parametric approach



Vaze et al. (CVPR'22)

Parametric approach

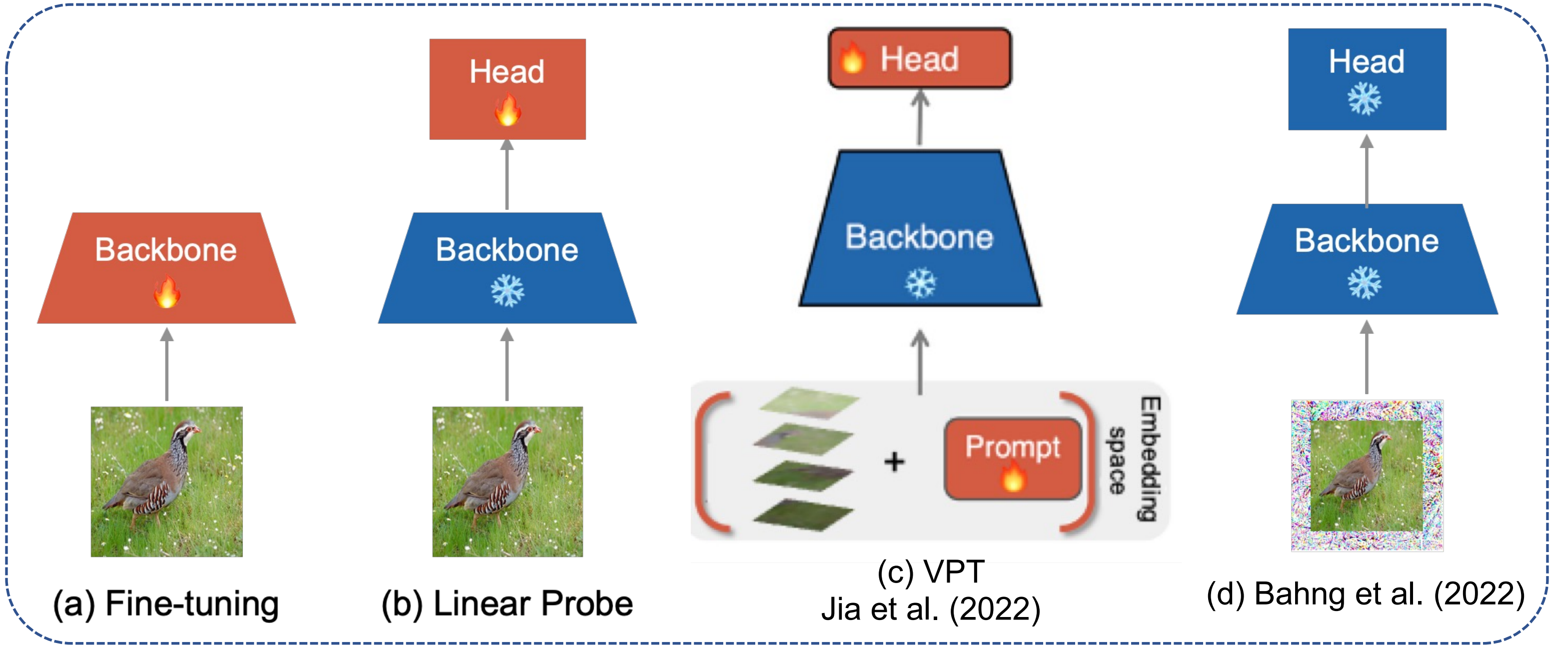


Wen et al. (ICCV'23)

Introduction – Research gap

Research gap

- Previous studies of GCD focused on model parameters, **overlooking the potential of data itself**
- Previous studies modifying the input or intermediate features through the addition of extra learnable tokens. They **do not improve representations for generalization**



Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In ECCV, 2022.
Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. ArXiv e-prints, 2022.

Introduction – Motivation

Prior Insight (Vaze et al. (2022))

- **Representations** with strong generalization properties achieve better GCD performance
- **Object parts** are an effective vehicle to transfer knowledge between 'seen' and 'unseen' categories

Our target

(1) Integrate advantages of **both model parameters and data parameters learning** for GCD, and improve representation from prompted data

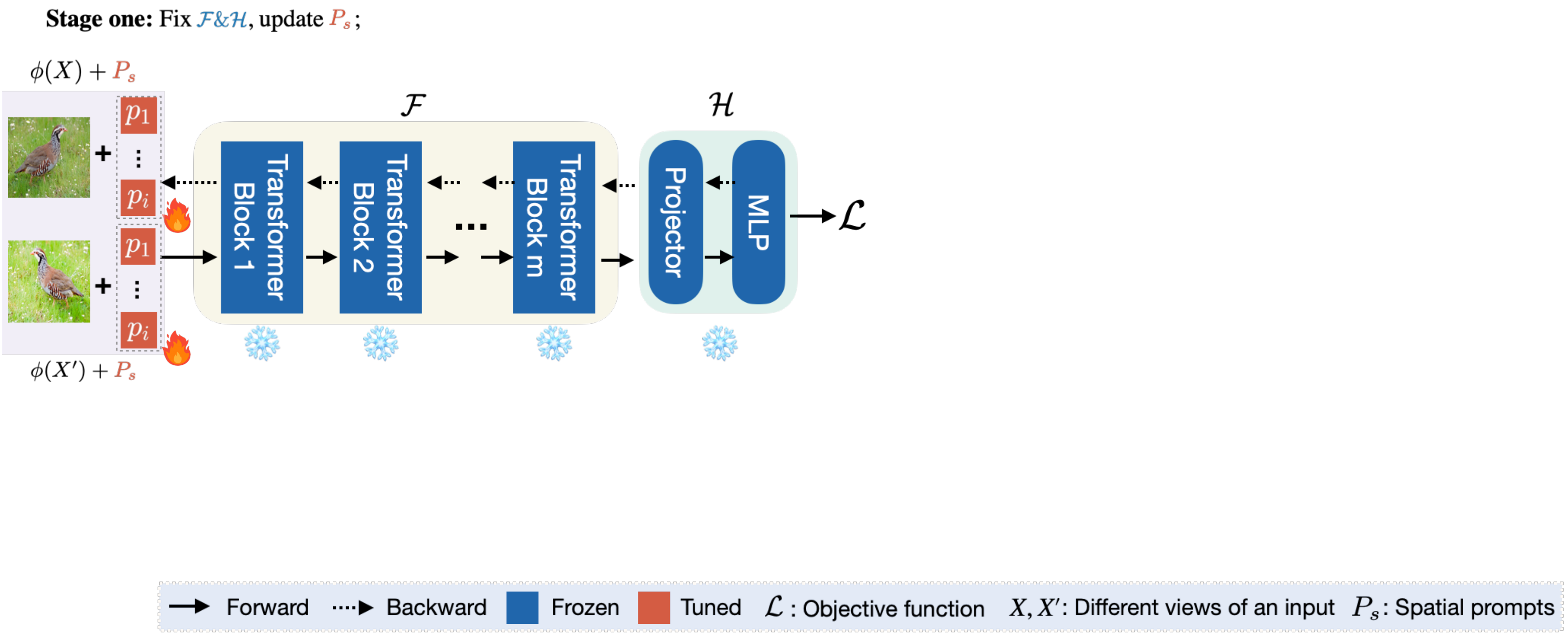
(2) Propose data parameters that enable the model to **focus on local image object regions**

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In CVPR, 2022.

Methodology

Framework

Stage one: Fix F&H and update Ps

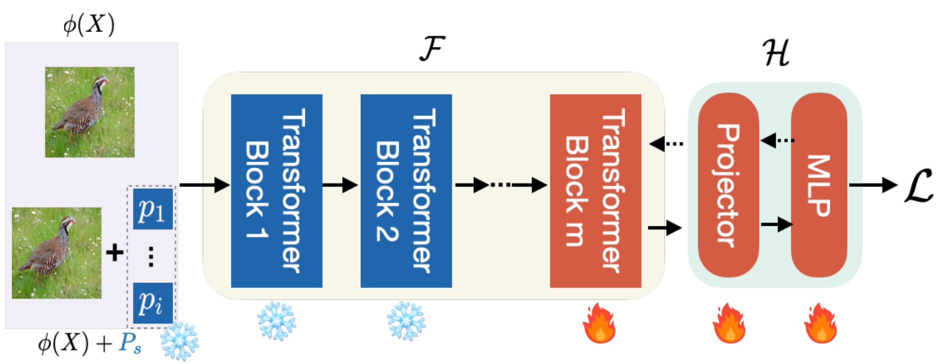


Methodology

Framework

Stage two: Fix Ps and update F&H

Stage two: Fix P_s , update $\mathcal{F}\&\mathcal{H}$.

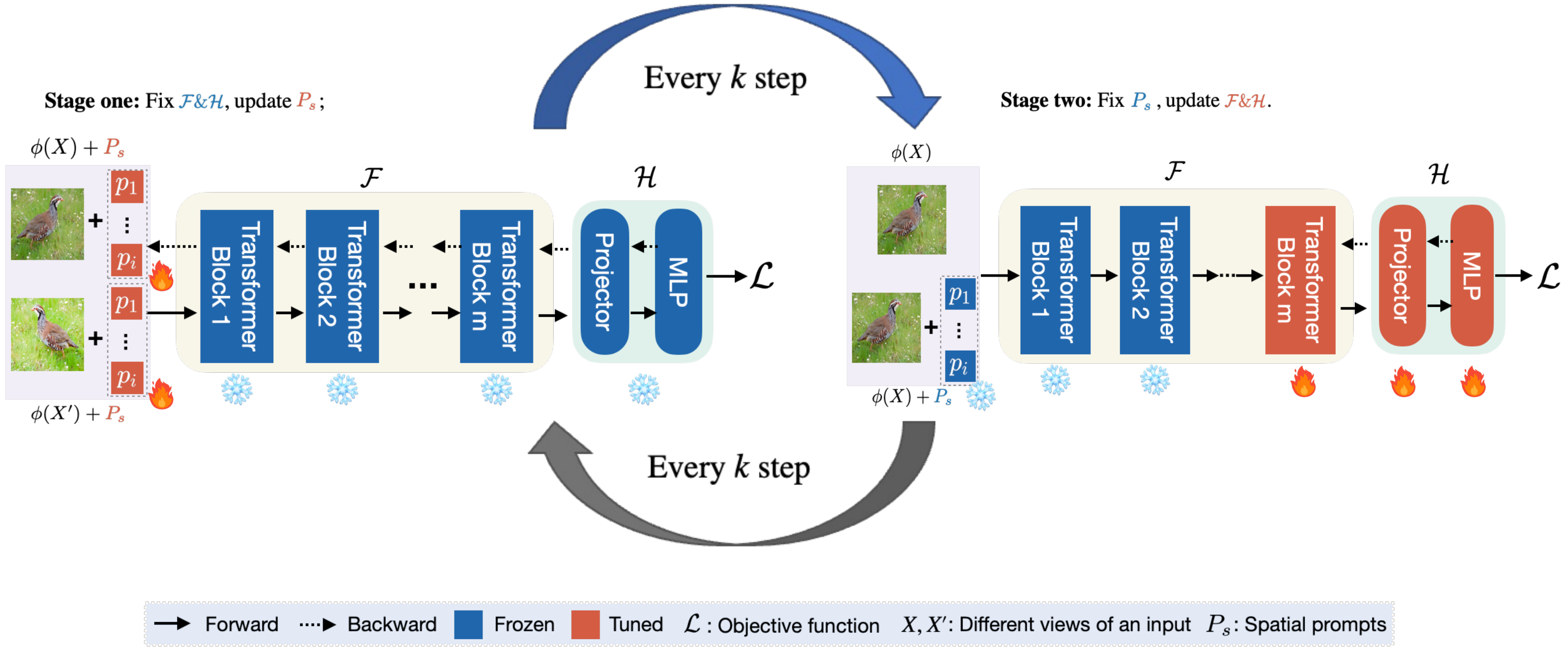


Forward ····▶ Backward Frozen Tuned $\mathcal{L}$: Objective function X, X': Different views of an input P_s: Spatial prompts

Methodology

Framework

Each stage optimizes the parameters for k iterations.

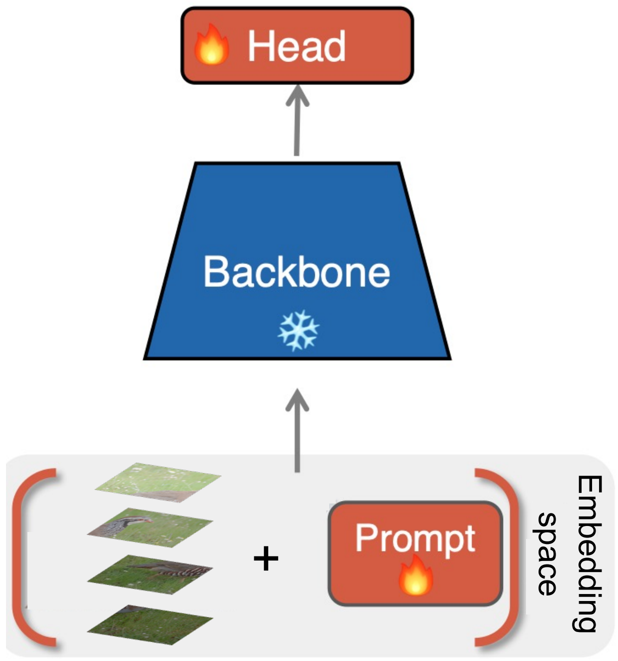


Methodology

Spatial Prompt Tuning (SPT)

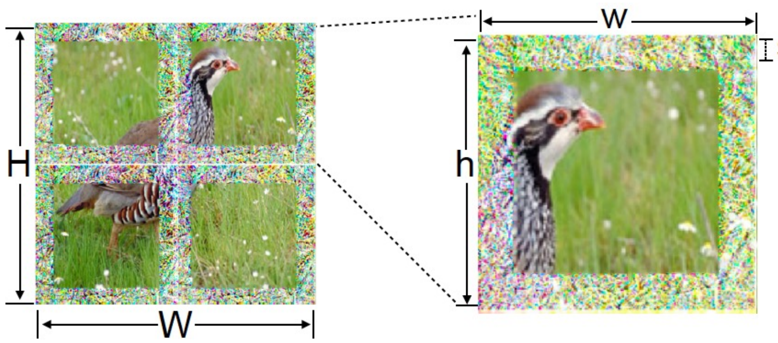
<u>Recall</u>: object parts are an effective vehicle to transfer knowledge between ‘seen’ and ‘unseen’ categories

<u>SPT</u>: enables the model to <u>focus on local image object regions</u>, while serving as <u>a learned data augmentation</u> for model parameters updating



VPT
Jia et al. (2022)

Bahng et al. (2022)



SPT

SPT & Global

Results and Discussion

Dataset statistics

- Generic datasets
  - i.e. CIFAR-10, CIFAR-100, and ImageNet-100
- Fine-grained datasets
  - i.e. CUB, Stanford Cars, FGVC-Aircraft, and Herbarium-19

Table 1: Dataset statistics and training configurations.

	Labelled		Unlabelled		Configs					
Dataset	#Num	#Class	#Num	#Class	$\texttt{lr}_b$	$\texttt{wd}_b$	$\texttt{lr}_p$	$\texttt{wd}_p$	k	m
CIFAR10 Krizhevsky et al. (2009)	12.5K	5	37.5K	10	3e-3	5e-4	1.0	0	20	1
CIFAR100 Krizhevsky et al. (2009)	20.0K	80	30.0K	100	1e-3	5e-4	1.0	0	20	1
ImageNet-100 Tian et al. (2020)	31.9K	50	95.3K	100	3e-3	5e-4	10.0	0	20	1
Herbarium 19 Tan et al. (2019)	8.9K	341	25.4K	683	3e-3	5e-4	10.0	0	20	1
CUB Welinder et al. (2010)	1.5K	100	4.5K	200	0.05	5e-4	25.0	0	20	1
Stanford Cars Krause et al. (2013)	2.0K	98	6.1K	196	0.05	5e-4	25.0	0	20	1
FGVC-Aircraft Maji et al. (2013)	1.7K	50	5.0K	50	0.05	5e-4	25.0	0	20	1

Results and Discussion

Generic datasets

- SPTNet consistently outperforms previous SOTA methods
- Limited gains (i.e. CIFAR-10 / CIFAR-100) caused by extremely low-resolution

Table 2: Evaluation on three generic image recognition datasets. Bold values represent the best results, while underlined values represent the second best results.

	CIFAR-10			CIFAR-100			ImageNet-100		
Method	All	Old	New	All	Old	New	All	Old	New
k-means Arthur & Vassilvitskii (2006)	83.6	85.7	82.5	52.0	52.2	50.8	72.7	75.5	71.3
RankStats+ Han et al. (2021)	46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8
UNO+ Fini et al. (2021)	68.6	**98.3**	53.8	69.5	80.6	47.2	70.3	**95.0**	57.9
GCD Vaze et al. (2022)	91.5	<u>97.9</u>	88.2	73.0	76.2	66.5	74.1	89.8	66.3
ORCA Cao et al. (2022)	96.9	95.1	97.8	74.2	82.1	67.2	79.2	93.2	72.1
SimGCD Wen et al. (2023)	97.1	95.1	98.1	80.1	81.2	**77.8**	83.0	93.1	77.9
DCCL Pu et al. (2023)	96.3	96.5	96.9	75.3	76.8	70.2	80.5	90.5	76.2
PromptCAL Zhang et al. (2023)	**97.9**	<u>96.6</u>	98.5	<u>81.2</u>	84.2	75.3	83.1	92.7	<u>78.3</u>
SPTNet (Ours)	<u>97.3</u>	95.0	**98.6**	**81.3**	**84.3**	<u>75.6</u>	**85.4**	<u>93.2</u>	**81.4**

Results and Discussion

Fine-grained datasets

- SPTNet achieves an average proportional improvement of ~10% across all evaluated datasets in SSB
- SPT assists the model in focusing on details that dominate correctness in fine-grained recognition in GCD

Table 3: Evaluation on the Semantic Shift Benchmark (SSB) and Herbarium 19. Bold values represent the best results, while underlined values represent the second best results.

	CUB			Stanford Cars			FGVC-Aircraft			Herbarium19		
Method	All	Old	New	All	Old	New	All	Old	New	All	Old	New
k-means Arthur & Vassilvitskii (2006)	34.3	38.9	32.1	12.8	10.6	13.8	12.9	12.9	12.8	13.0	12.2	13.4
RankStats+ Han et al. (2021)	33.3	51.6	24.2	28.3	61.8	12.1	27.9	55.8	12.8	27.9	55.8	12.8
UNO+ Fini et al. (2021)	35.1	49.0	28.1	35.5	70.5	18.6	28.3	53.7	14.7	28.3	53.7	14.7
GCD Vaze et al. (2022)	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9	35.4	51.0	27.0
ORCA Cao et al. (2022)	36.3	43.8	32.6	31.9	42.2	26.9	31.6	32.0	31.4	20.9	30.9	15.5
SimGCD Wen et al. (2023)	60.3	65.6	57.7	53.8	<u>71.9</u>	45.0	<u>54.2</u>	59.1	51.8	<u>43.0</u>	58.0	<u>35.1</u>
DCCL Pu et al. (2023)	63.5	60.8	<u>64.9</u>	43.1	55.7	36.2	-	-	-	-	-	-
PromptCAL Zhang et al. (2023)	62.9	64.4	62.1	50.2	70.1	40.6	52.2	52.2	52.3	37.0	52.0	28.9
SPTNet (Ours)	**65.8**	<u>68.8</u>	**65.1**	**59.0**	**79.2**	<u>49.3</u>	**59.3**	**61.8**	**58.1**	**43.4**	**58.7**	**35.2**

Results and Discussion

Ablation objective: Effect of prompt-related techniques

- Existing prompt tuning methods does not yield satisfactory performance, while SPT gives a relatively larger improvement of 1.8% on ‘All’ classes
- Alternate training can effectively improve the performance
- After further introducing the global prompts, the performance is further improved

Table 4: Comparison on effectiveness of different prompting methods on SSB. We report the average test accuracy score over all component datasets of SSB (*i.e.*, CUB, Stanford Cars and FGVC-Aircraft). ‘Shared’ and ‘Alter’ refer to a single *shared* prompt for all patches and *alternative* learning. Row (9) represents SPTNet and rows (6) and (7) represent its two variants SPTNet-P and SPTNet-S.

No	Method config	Prompt config	All	Old	New
(1)	SimGCD Wen et al. (2023)	None (baseline)	56.1	65.5	51.5
(2)		+VPT Jia et al. (2022)	$54.4^{-1.7}$	$64.7^{-0.8}$	$49.1^{-2.4}$
(3)		+Global Bahng et al. (2022)	$56.7^{+0.6}$	$64.6^{-0.9}$	$53.5^{+2.0}$
(4)		+SPT	$57.9^{+1.8}$	$67.2^{+1.7}$	$53.3^{+1.8}$
(4)	+Alter	+Global Bahng et al. (2022)	$57.8^{+1.7}$	$66.3^{+0.8}$	$53.8^{+2.3}$
(5)		+Shared	$60.5^{+4.4}$	$68.6^{+3.1}$	$56.5^{+5.0}$
(6)		+SPT	$59.1^{+3.0}$	$68.5^{+3.0}$	$54.5^{+3.0}$
(7)	+Alter	+Shared & Global Bahng et al. (2022)	$60.9^{+4.8}$	$69.0^{+3.5}$	$57.3^{+5.8}$
(8)		+SPT & Global Bahng et al. (2022)	$61.4^{+5.3}$	$69.9^{+4.4}$	$57.5^{+6.0}$

Results and Discussion

Ablation objective: Effect of different training strategies

a) Finetune: continue finetuning pretrained SimGCD model

b) End-to-end: both the data parameters and the model parameters are jointly trained

c) Data first / model first: the prompt / model parameters are optimized first, followed by the model / prompt parameters

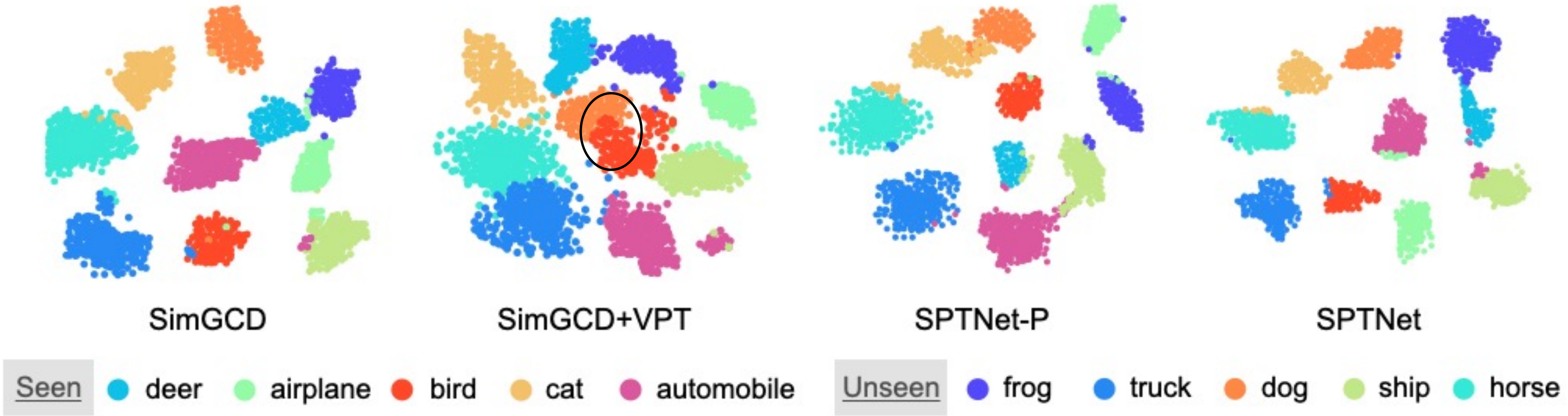
Table 5: Evaluation on ImageNet-100 and SSB using different training strategies.

		ImageNet-100			SSB		
No	Methods	All	Old	New	All	Old	New
(1)	SimGCD Wen et al. (2023)	83.0	93.1	77.9	56.1	65.5	51.5
(2)	SimGCD (further fine-tune)	84.3	93.1	79.7	57.0	66.0	52.3
(3)	SPTNet (end-to-end)	84.1	92.8	80.0	58.6	67.4	53.2
(4)	SPTNet (data first)	83.5	92.9	77.7	58.0	66.4	51.9
(5)	SPTNet (model first)	84.8	93.3	80.6	59.2	67.8	54.9
(6)	SPTNet (alternative)	**85.4**	**93.2**	**81.4**	**61.4**	**69.9**	**57.5**

Results and Discussion

Ablation objective: How do prompts affect the representations?

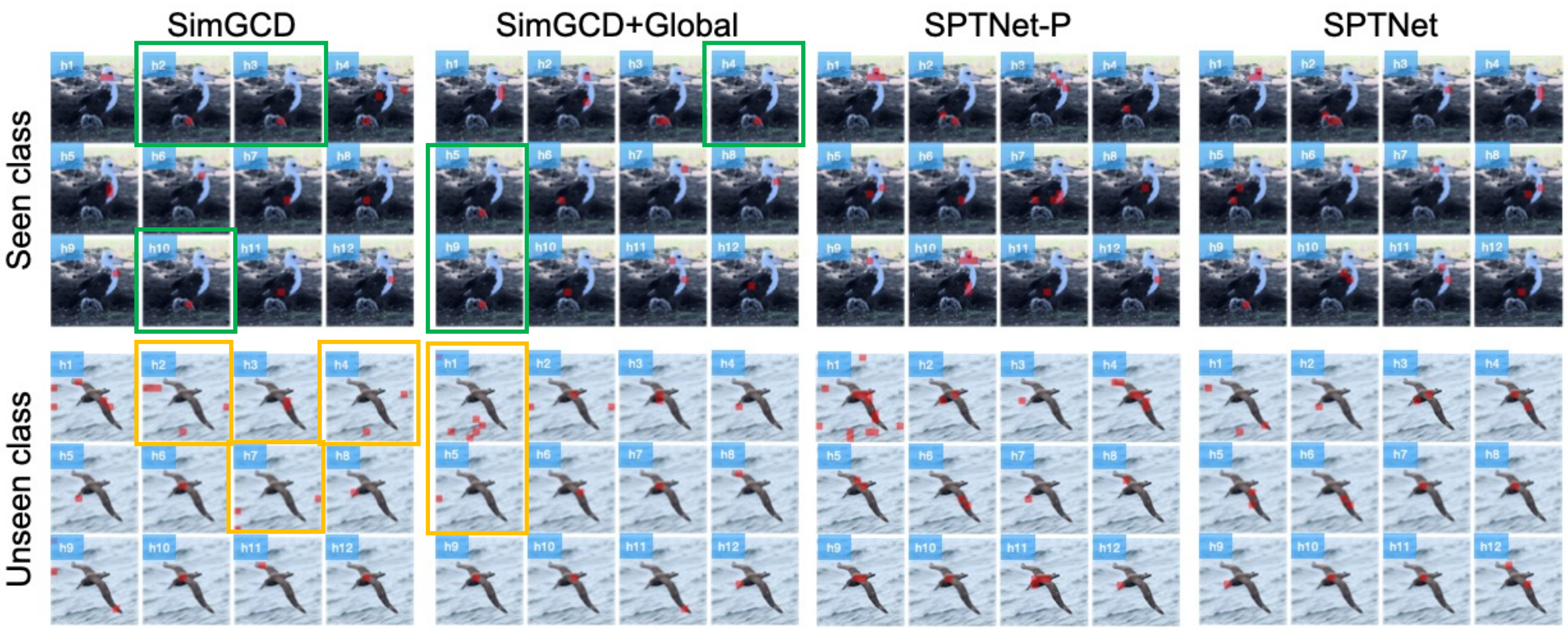
- VPT leads to clutter between seen and unseen classes
- SPTNet and its variant produce more discriminative features and more compact clusters



Results and Discussion

Ablation objective: How do prompts affect the model's attention?

- **Issue**: SimGCD and SimGCD+Global may focus on the same regions
- SPT and SPT&Global attend to more diverse regions of the object and focus more on the foreground object regions



Conclusion

- We propose **a two-stage alternative optimization scheme**, called SPTNet
  - Optimizing both model and data parameters, to enhance alignment between the pre-trained model and the target task.

- Additionally, we introduce **spatial prompt tuning (SPT)** as a method to
  - **Focusing on object parts** and facilitate knowledge transfer between seen and unseen classes
  - Yielding **extra parameters amounting to only 0.117%** of those in the backbone architecture.

Thanks for listening!