PromptCAL: Contrastive Affinity Learning via Auxiliary Prompts for Generalized Novel Category Discovery

Sheng Zhang[1] , Salman Khan[12], Zhiqiang Shen[13],
Muzammal Naseer[1], Guangyi Chen[14], Fahad Shahbaz Khan[15]

[1]Mohamed Bin Zayed of Artificial Intelligence, UAE
[2]Australian National University, Australia
[3]Hong Kong University of Science and Technology, China
[4]Carnegie Mellon University, US
[5]Linkoping University, Sweden

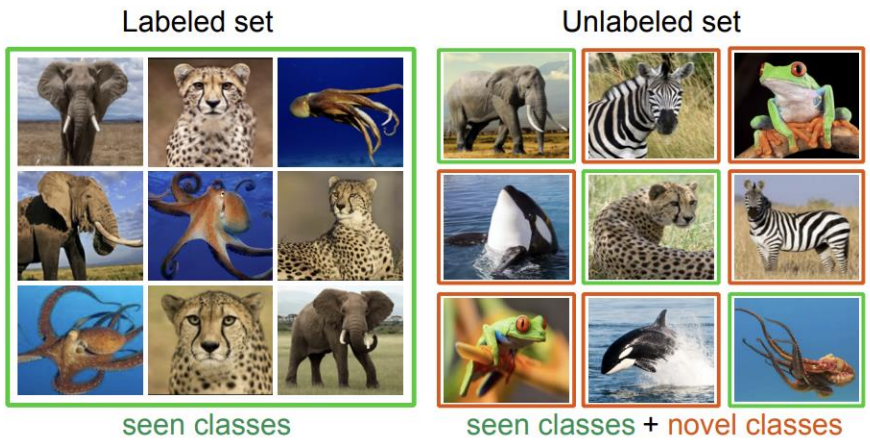
Poster:TUE-AM-331

Project Page: https://github.com/sheng-eatamath/PromptCAL

Paper: https://arxiv.org/abs/2212.05590



Problem



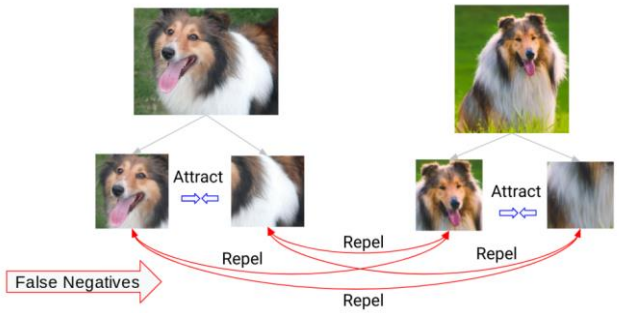
Generalized Novel Category Discovery (GNCD) aims to learn to categorize known and novel classes, given labeled-knowns and an unlabeled set with known and novel classes.

Challenge 1: Adaptability

Most layers of backbone are frozen.

Lack of Semantic Discriminativeness

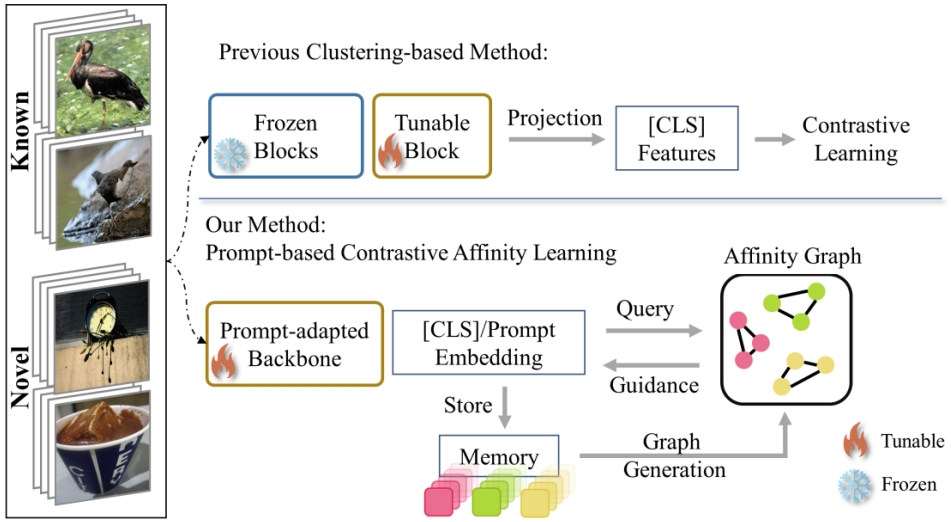
Challenge 2: Class Collision



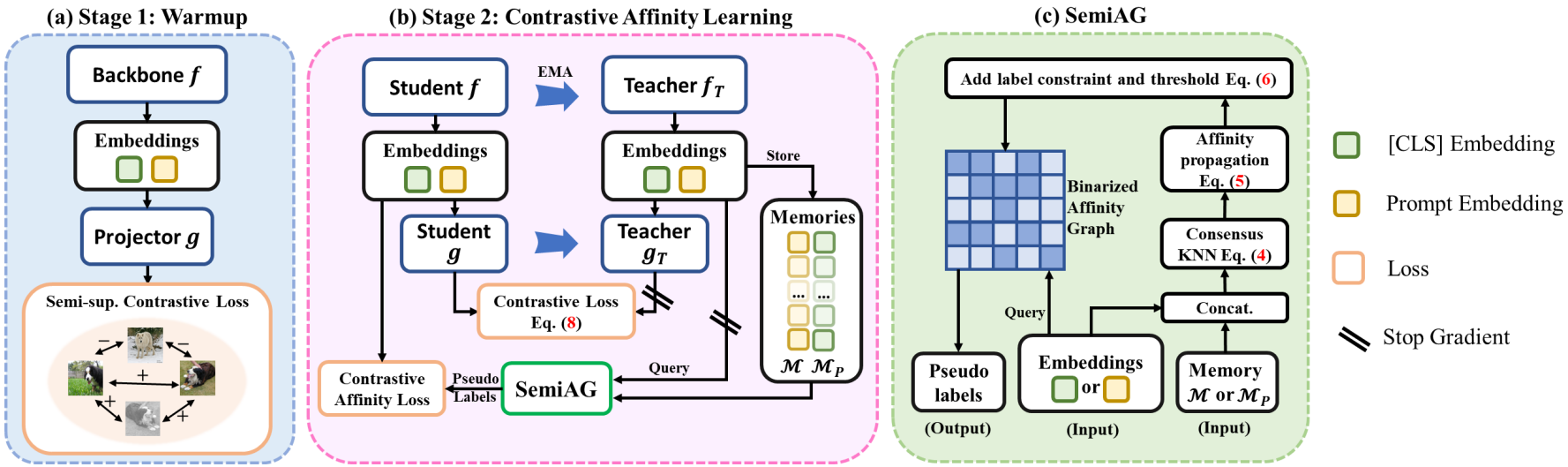
[1] *Cao, K., Brbic, M., & Leskovec, J. (2021). Open-world semi-supervised learning. arXiv preprint arXiv:2102.03526.*
[2] *Huynh, T., Kornblith, S., Walter, M. R., Maire, M., & Khademi, M. (2022). Boosting contrastive self-supervised learning with false negative cancellation. In Proceedings of the IEEE/CVF winter conference on applications of computer vision.*

Summary

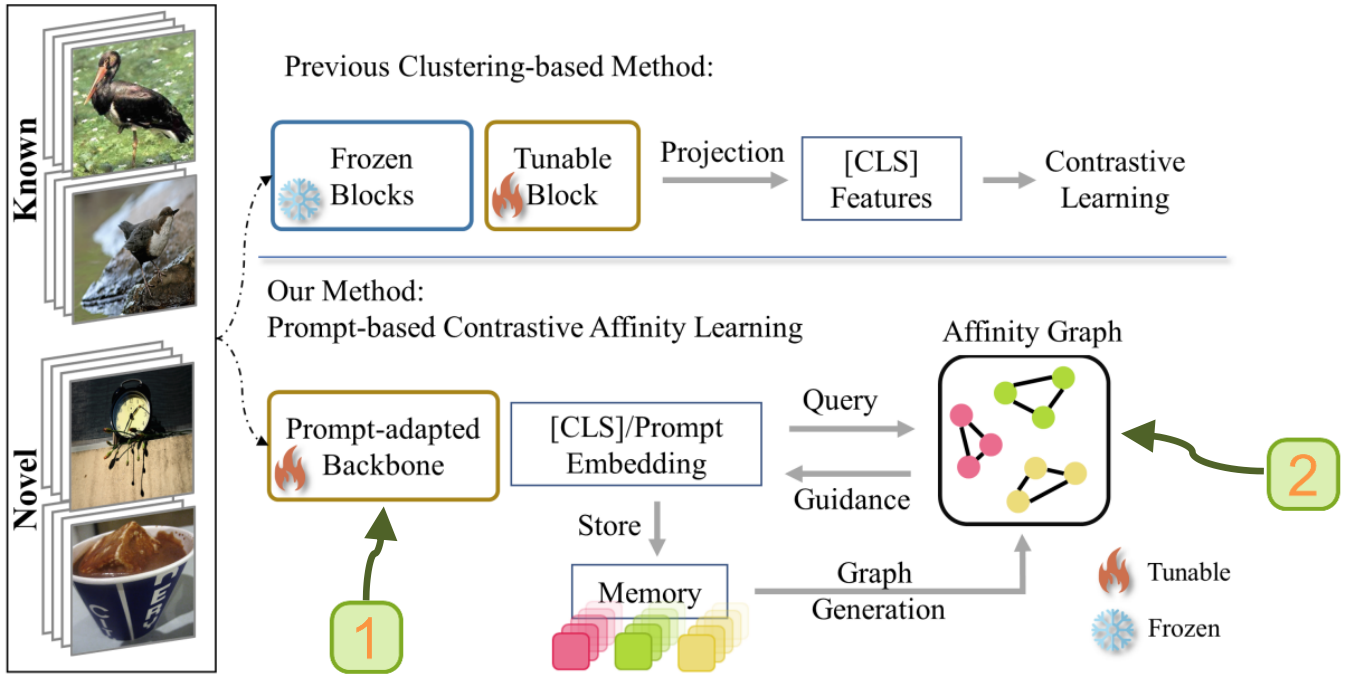
Method Overview



Overall Framework



Motivation



1 Adaptability

Frozen ViT shallow layers constrains the model semantic discriminativeness. Besides, we discover that visual prompt tuning leads to overfitting and needs regularization.

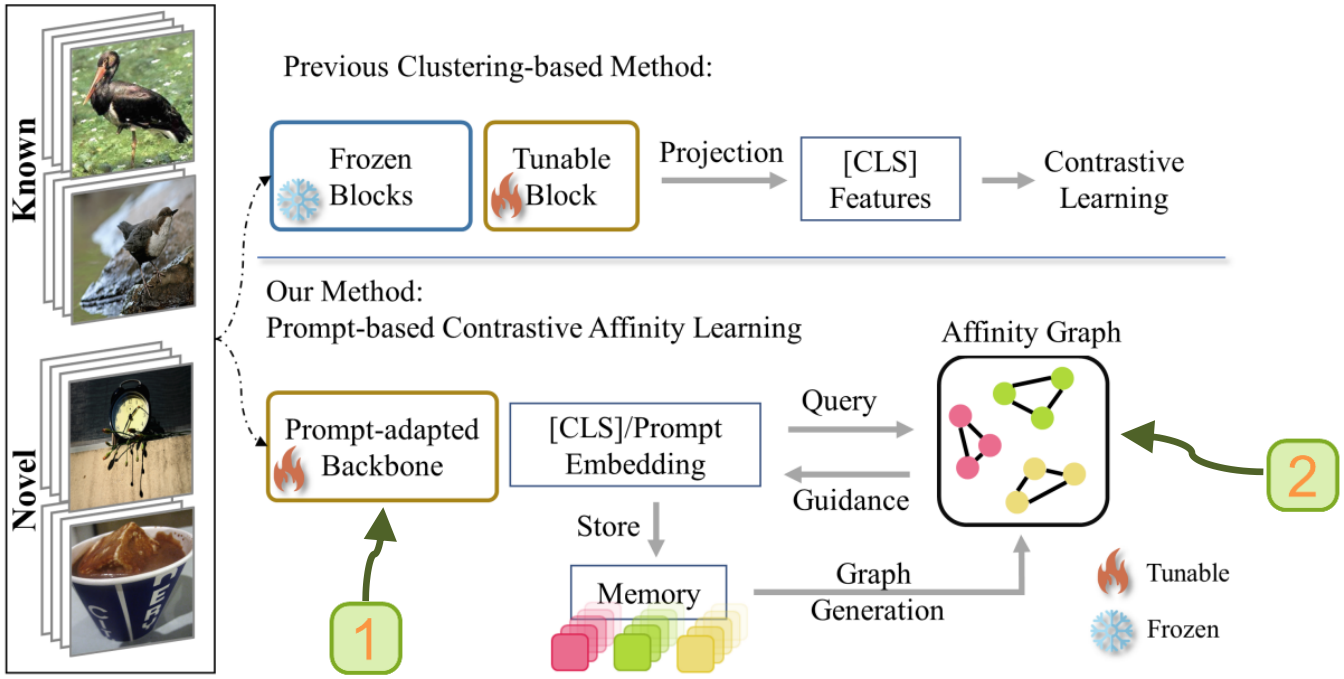
1 We propose ***Discriminative Prompt Regularization*** to provide extra flexibility to pre-trained backbone as well as alleviate overfitting.

2 Class Collision

Contrastive learning between unlabeled intra-class instances will significantly corrupt model semantic discriminativeness.

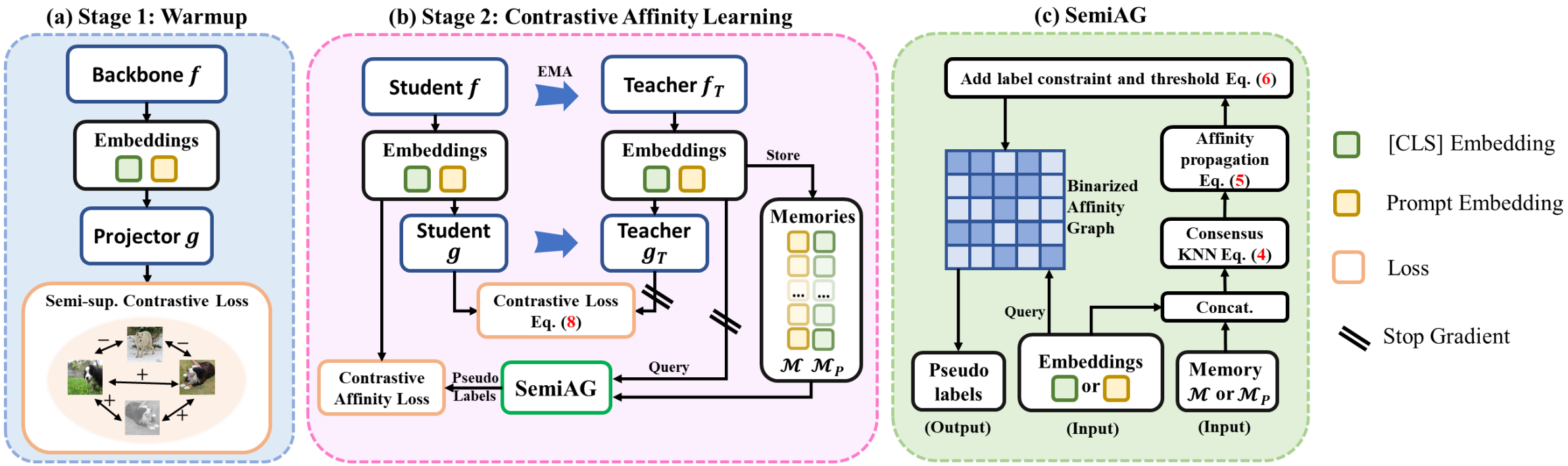
2 We propose ***Contrastive Affinity Learning*** to discover and calibrate false-negative pairs based on affinity graph for contrastive learning.

Contributions



- We propose **PromptCAL**, a novel two-stage framework, to address the GNCD problem.
- We design two synergistic objectives, *Discriminative Prompt Tuning* 1 and *Contrastive Affinity Learning* 2 which is based on proposed *Semi-supervised Affinity Generation* process, to enhance ViT semantic discriminativeness for known and novel classes.
- Achieve SOTA on six benchmarks and generalize to more challenging GNCD setups.

Method



Pre-trained ViT backbone with deep visual prompts f, projection head g, EMA teachers (f_T, g_T), [CLS]/prompt embedding memory $\mathcal{M}, \mathcal{M}_P$.

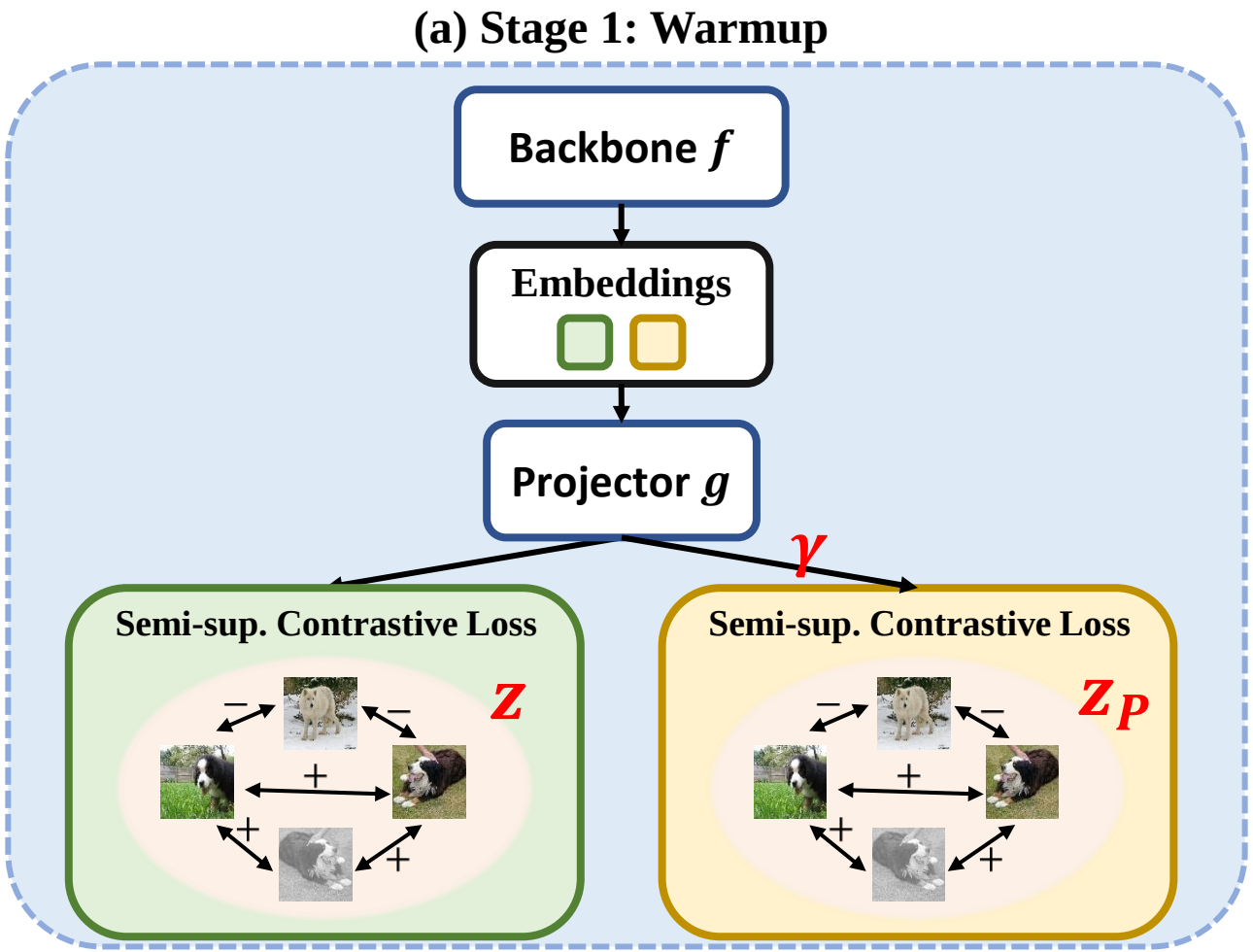
PromptCAL — Discriminative Prompt Regularization (DPR)

To regularize visual prompts from overfitting, we also add supervision on visual prompts at last ViT block.

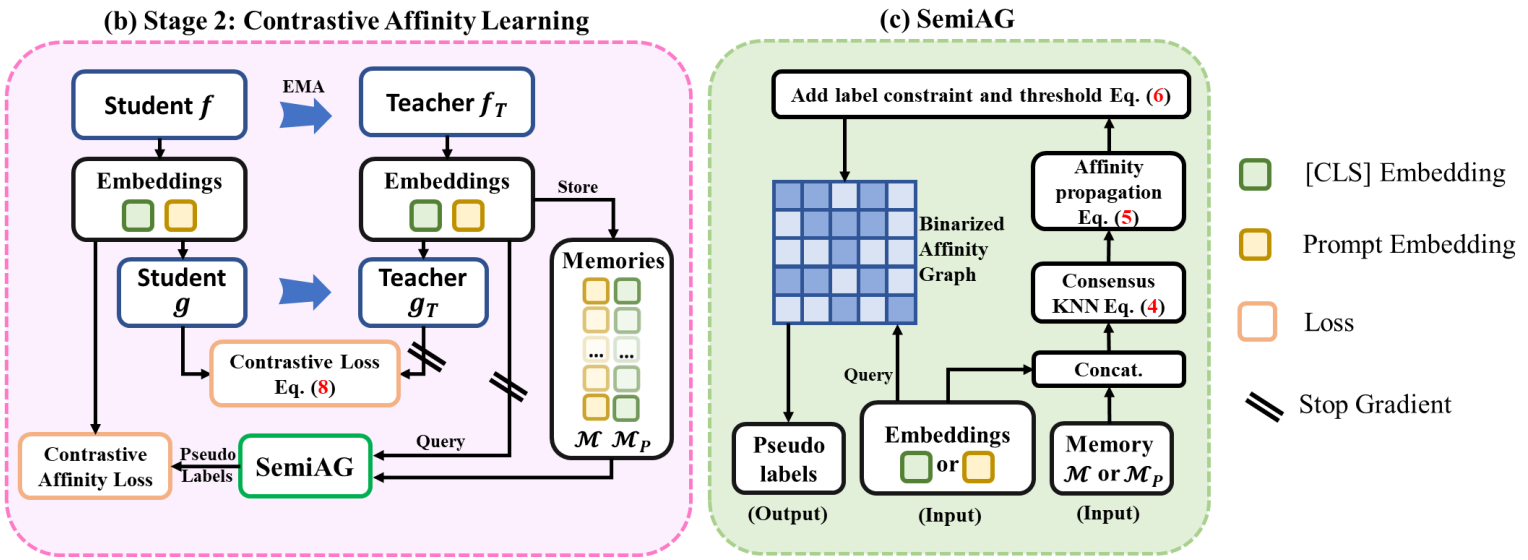
Prompt regularization loss are enforced in both stages, e.g., in the 1st stage,

$$L_1(\mathbf{x}) = L_{\text{semi}}^{\text{CLS}}(\mathbf{z}) + \gamma L_{\text{semi}}^{\text{P}}(\mathbf{z}_{\text{P}})$$

Discriminative Prompt Regularization loss: apply same loss form on prompts



PromptCAL — Contrastive Affinity Learning (CAL)



During contrastive affinity learning (2nd stage), we design the following loss for CLS token,

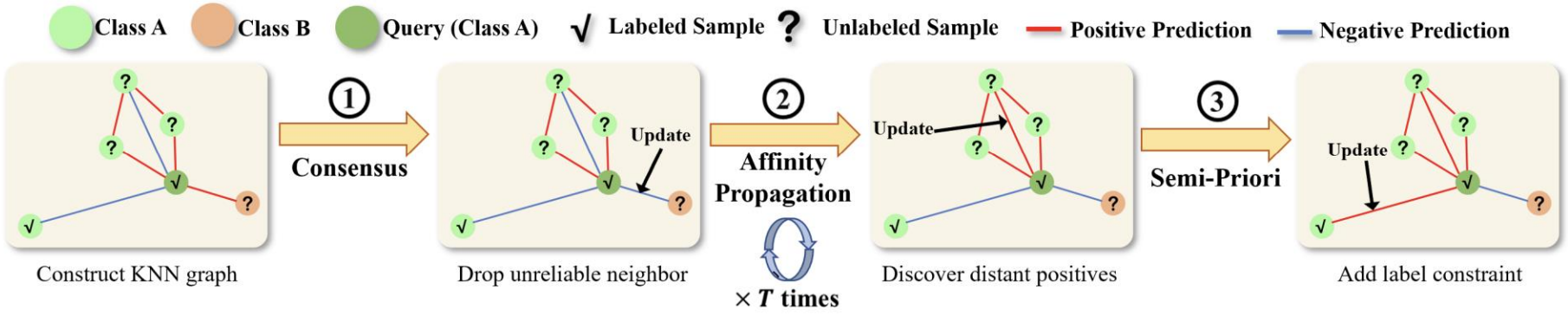
$$L_2^{\text{CLS}} = (1-\alpha)L_{\text{sup}}^{\text{CLS}} + \alpha\left(\beta L_{\text{CAL}}^{\text{CLS}} + (1-\beta)L_{\text{self}}^{\text{CLS}}\right)$$

- L_2^{CLS}: Loss on [CLS] token
- $L_{\text{sup}}^{\text{CLS}}$: Supervised contrastive loss
- $L_{\text{CAL}}^{\text{CLS}}$: **Contrastive Affinity Loss (CAL Loss)**
- $L_{\text{self}}^{\text{CLS}}$: Self-supervised contrastive loss

After discriminative prompt regularization, the total loss of the 2nd stage is:

$$L_2 = L_2^{\text{CLS}} + \gamma L_2^{\text{P}}$$

PromptCAL — Semi-supervised Affinity Generation (SemiAG)



Step 1: Build Consensus Graph

Sample embeddings from the token memory,

For any $\mathbf{h}_i, \mathbf{h}_j$ from all embeddings $\mathcal{V}$,

$$g_{i,j} = \begin{cases} |\{\mathbf{h}_c | \mathbf{h}_i, \mathbf{h}_j \in \mathcal{O}_K(\mathbf{h}_c)), \forall \mathbf{h}_c \in \mathcal{V}\}| & i \neq j \\ 0 & i = j, \end{cases}$$

$$\mathcal{O}_K(\mathbf{h}_c) = \texttt{argtopK}_{\mathbf{h}_j}(\{\mathbf{h}_j \cdot \mathbf{h}_c | \mathbf{h}_j \in \mathcal{V}\})$$

Step 2: Affinity Propagation

Consensus Graph

graph ↓ diffusion

Row-normalized diffused graph $\tilde{\mathbf{G}}_d$

Step 3: SemiPrior Calibration

Positive Pairs

$$\mathbf{G}_b(i,j) = \begin{cases} 1 & (y_i = y_j) \vee \left(\tilde{\mathbf{G}}_d(i,j) > q\right) \\ 0 & (y_i \neq y_j) \end{cases}$$

Negative Pairs

$\mathbf{G}_b$ provides pseudo-labels for contrastive affinity learning.

Experiments

- Datasets

Dataset	CIFAR-10	CIFAR-100	ImageNet-100	CUB-200	Aircraft	StanfordCars
#Images in $\mathcal{D}$	50k	50k	127.2k	6k	6.6k	8.1k
#Classes ($\lvert\mathcal{C}\rvert$)	10	100	100	200	100	196
#Known Classes ($\lvert\mathcal{C}_{kwn}\rvert$)	5	80	50	100	50	98

Table 4.1: **The dataset profiles of six benchmarks for evaluation.**

- Metrics: Clustering Accuracy

$$\text{Acc} = \max_{\rho \in \mathcal{P}(\mathcal{Y})} \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \mathbb{I}(y_i = \rho(\hat{y}_i))$$

- Implementation Details

Model Architecture: ViT-B/16, pre-trained DINO
Hyperparameters: $\alpha = 0.35, \beta = 0.6, \gamma = 0.35$, Memory size 4096, $K = |\mathcal{M}|/(4|\mathcal{C}|)$
Negative samples for CAL: 1024
Optimizer: SGD, initial lr=0.1, momentum=0.9, weight decay=5e-4
1st stage: training for 100 epochs for ImageNet-100 and 200 epochs for others.
2nd stage: training for 70 epochs for generic datasets and 100 epochs for fine-grained datasets.

Evaluation

Method	CIFAR-10			CIFAR-100			ImageNet-100		
	All	Known	New	All	Known	New	All	Known	New
KMeans [2]	83.6	85.7	82.5	52.0	52.2	50.8	72.7	75.5	71.3
RankStats+ [19]	46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8
UNO+ [14]	68.6	**98.3**	53.8	69.5	80.6	47.2	70.3	**95.0**	57.9
GCD [56]	91.5	97.9	88.2	73.0	76.2	66.5	74.1	89.8	66.3
ORCA† [6]	96.9	95.1	97.8	74.2	82.1	67.2	79.2	93.2	72.1
PromptCAL-1st (Ours)	97.1	97.7	96.7	76.0	80.8	66.6	75.4	94.2	66.0
PromptCAL-2nd (Ours)	**97.9**	96.6	**98.5**	**81.2**	**84.2**	**75.3**	**83.1**	92.7	**78.3**

Table 4.2: **Evaluation on three generic datasets.** Accuracy scores are reported. †denotes adapted methods. Both stages of PromptCAL are evaluated.

Method	CUB-200			StanfordCars			Aircraft		
	All	Known	New	All	Known	New	All	Known	New
KMeans [2]	34.3	38.9	32.1	12.8	10.6	13.8	12.9	12.9	12.8
RankStats+ [19]	33.3	51.6	24.2	28.3	61.8	12.1	27.9	**55.8**	12.8
UNO+ [14]	35.1	49.0	28.1	35.5	**70.5**	18.6	28.3	53.7	14.7
GCD [56]	51.3	56.6	48.7	39.0	57.6	29.9	45.0	41.1	46.9
ORCA† [6]	36.3	43.8	32.6	31.9	42.2	26.9	31.6	32.0	31.4
PromptCAL-1st (Ours)	51.1	55.4	48.9	42.6	62.8	32.9	44.5	44.6	44.5
PromptCAL-2nd (Ours)	**62.9**	**64.4**	**62.1**	**50.2**	70.1	**40.6**	**52.2**	52.2	**52.3**

Table 4.3: **Evaluation on three fine-grained datasets.** Accuracy scores are reported. †denotes adapted methods. Both stages of PromptCAL are evaluated.

- Consistently and significantly surpasses SOTA
- 2nd stage gains more improvements
- More favorable gains on New
- Remarkable advantages on FG datasets

Ablations

Dataset	Setup	All	Known	New
CUB-200	w/o prompt	60.3	64.8	58.0
CUB-200	w/o DPR	59.3	63.3	57.4
CUB-200	KNN w/ S.P.	60.1	**70.1**	55.1
CUB-200	R.S.	55.6	66.0	50.3
CUB-200	PromptCAL	**62.9**	64.4	**62.1**
CIFAR-100	w/o prompt	78.1	83.0	68.4
CIFAR-100	w/o DPR	79.0	83.4	70.3
CIFAR-100	KNN w/ S.P.	78.7	85.3	65.4
CIFAR-100	R.S.	75.9	**87.1**	53.4
CIFAR-100	PromptCAL	**81.2**	84.2	**75.3**
ImageNet-100	w/o prompt	81.8	94.7	75.3
ImageNet-100	w/o DPR	80.7	94.8	73.6
ImageNet-100	KNN w/ S.P.	81.9	95.0	75.3
ImageNet-100	R.S.	78.1	**95.2**	69.4
ImageNet-100	PromptCAL	**83.1**	92.7	**78.3**

Table 4.8: **Further ablation study on CUB-200 [57], CIFAR-100 [35], and ImageNet-100 [36] datasets.** We investigate four setups: the first is PromptCAL removing all prompt related components; the second is PromptCAL without DPR loss; the third is replacing SemiAG with naive KNN incorporated with SemiPriori; the last one is replacing our SemiAG with RankingStats [19] pseudo labeling.

- No prompt lead to suboptimal performance
- Prompt tuning without regularization degrades performance
- Naïve KNN cannot learn good affinities among novel class instances since it is not robust to noises

Ablations — Wilder Scenarios

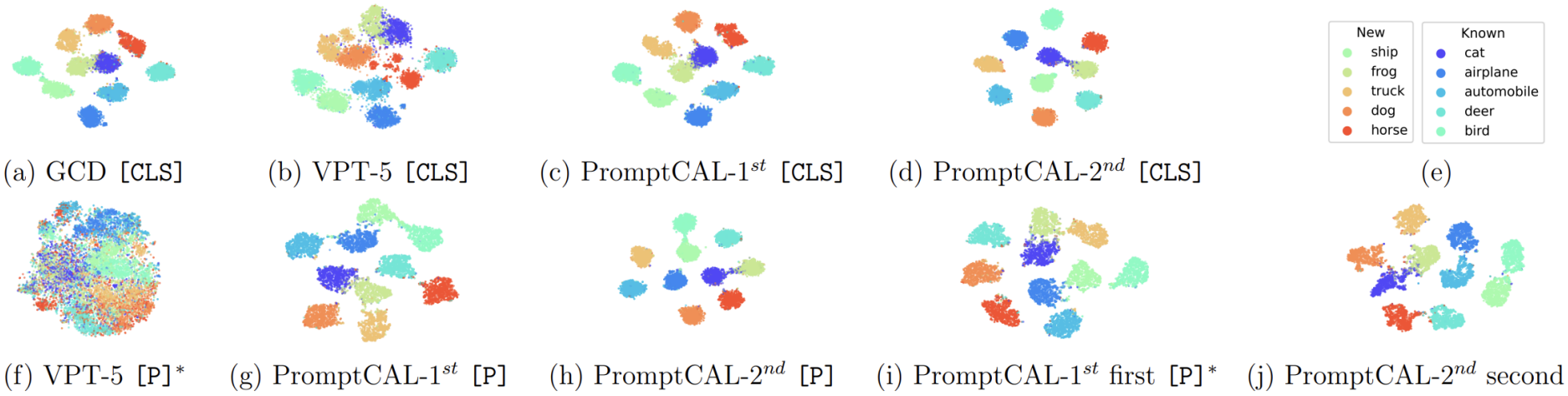
Method	C50-L10			C25-L50			C10-L50		
	All	Known	New	All	Known	New	All	Known	New
GCD [56]	60.2	68.9	55.8	56.8	67.6	55.0	48.3	65.1	47.3
ORCA (ResNet) [6]	39.4	55.1	31.2	37.0	64.1	31.7	30.1	64.3	27.1
ORCA† (ViT) [6]	60.3	66.0	55.3	58.2	**79.9**	57.5	51.7	78.0	50.2
PromptCAL-1st (Ours)	62.7	74.7	56.6	60.2	70.7	58.5	48.7	68.4	47.6
PromptCAL-2nd (Ours)	**68.9**	**77.5**	**64.7**	**65.7**	76.9	**63.8**	**53.2**	**79.3**	**51.7**

Table 4.6: **Ablation study on few-annotation GNCD on CIFAR-100 [35] dataset.** Digits following 'C' and 'L' stand for percentages of known classes and labeling ratios. †denotes adapted methods. Scores reported in accuracy.

Method	CUB-200					CIFAR-100					ImageNet-100				
	All	Known	New	Known*	New*	All	Known	New	Known*	New*	All	Known	New	Known*	New*
GCD [56]	57.5	64.5	50.6	69.2	57.6	70.1	76.8	43.5	78.7	58.2	79.7	92.7	66.7	92.7	66.9
ORCA (DINO) [6]	40.7	61.2	20.2	**76.3**	38.3	77.7	83.6	53.9	83.6	66.6	81.3	**94.5**	68.0	**94.5**	71.1
PromptCAL (our)	**62.4**	**68.1**	**56.8**	70.1	**60.1**	**81.6**	**85.3**	**66.9**	**86.2**	**71.3**	**84.8**	94.4	**75.2**	94.4	**75.3**

Table 4.12: **Evaluation in the inductive GCD setting[6] on three benchmarks.** The results are reported in accuracy scores on the test set. Here, we also adopt the task-informed evaluation protocol in [14, 6], *i.e.*, `Known`* and `New`* are evaluated by separate clustering and Hungarian assignment.

Ablations — Role of DPR



(a) GCD [CLS] (b) VPT-5 [CLS] (c) PromptCAL-1st [CLS] (d) PromptCAL-2nd [CLS] (e)

(f) VPT-5 [P]* (g) PromptCAL-1st [P] (h) PromptCAL-2nd [P] (i) PromptCAL-1st first [P]* (j) PromptCAL-2nd second [P]*

Figure 4.1: **The t-SNE [54] visualization of ViT embeddings on CIFAR-10 test set** for GCD [56], naive VPT model [28], and PromptCAL-1st stage and 2nd stage, Here, [CLS], [P], and [P]* denote embeddings from ViT class token, ensembled prompts supervised by DPR loss, and unsupervised prompts, respectively. The embedding clustering shows that DPR reinforces the semantic discriminativeness of [P], and for [P]* despite no explicit supervision. (e) exhibits the class name each color denotes. All figures share the same axis scale.

Qualitative Results — Nearest Neighbors Query

Visualization of 8-NN predictions for randomly sampled query images from ImageNet-100.

Qualitative Results — Confusion Matrix

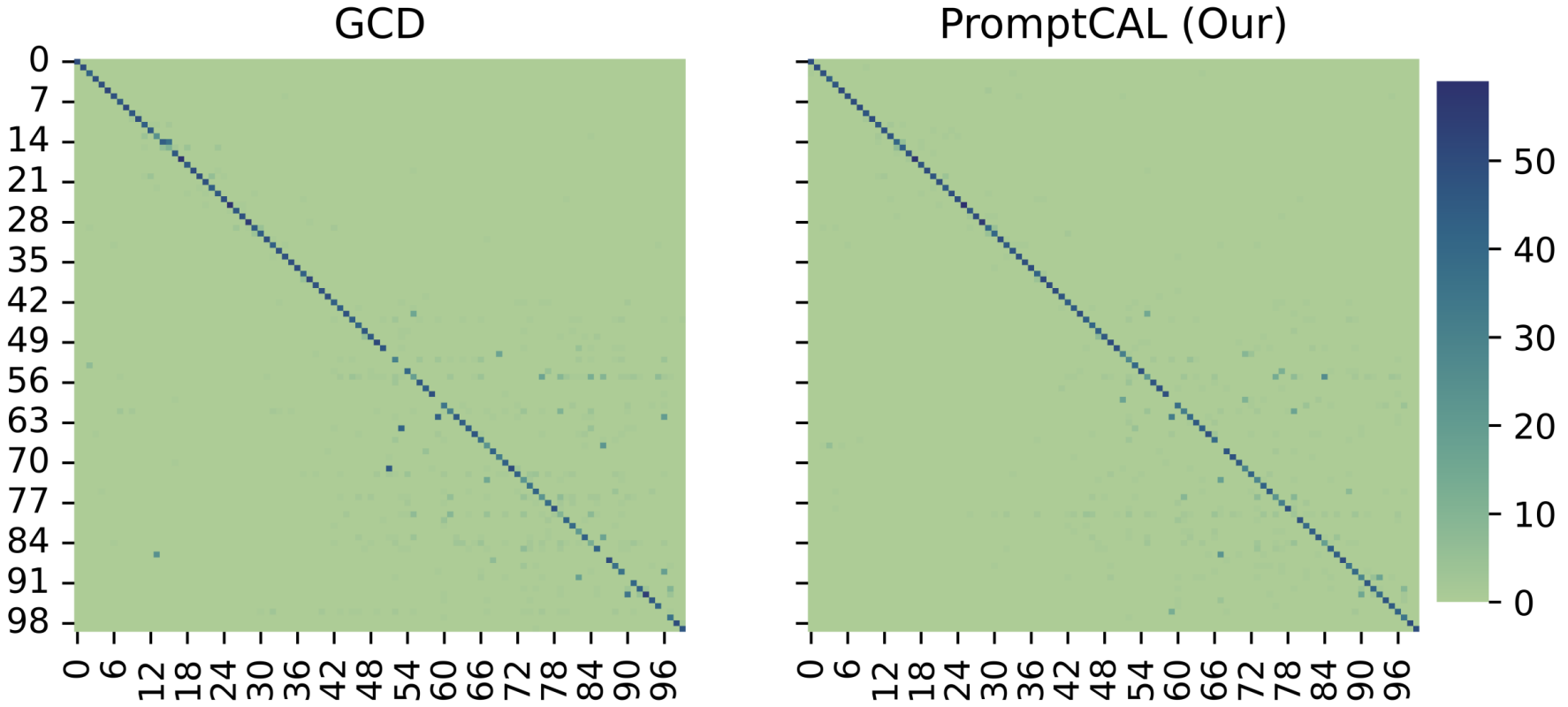


Figure 4.3: **Confusion matrix of PromptCAL on ImageNet-100 [36] test set.** The labels on the x-axis and y-axis denotes the class index of our generated split. The first 50 classes are `Known`, and the last 50 classes are `New`.

Conclusions

1. We propose a novel two-stage framework, PromptCAL, for the generalized novel category discovery problem.

2. We propose two synergistic learning objectives, discriminative prompt regularization and contrastive affinity learning.

3. We comprehensively validate the effectiveness of our method on multiple benchmarks, achieving state-of-the-art performance.

4. We further showcase generalization ability of PromptCAL in challenging few-annotation scenarios and inductive setup.