Reference-Limited Compositional Zero-Shot Learning

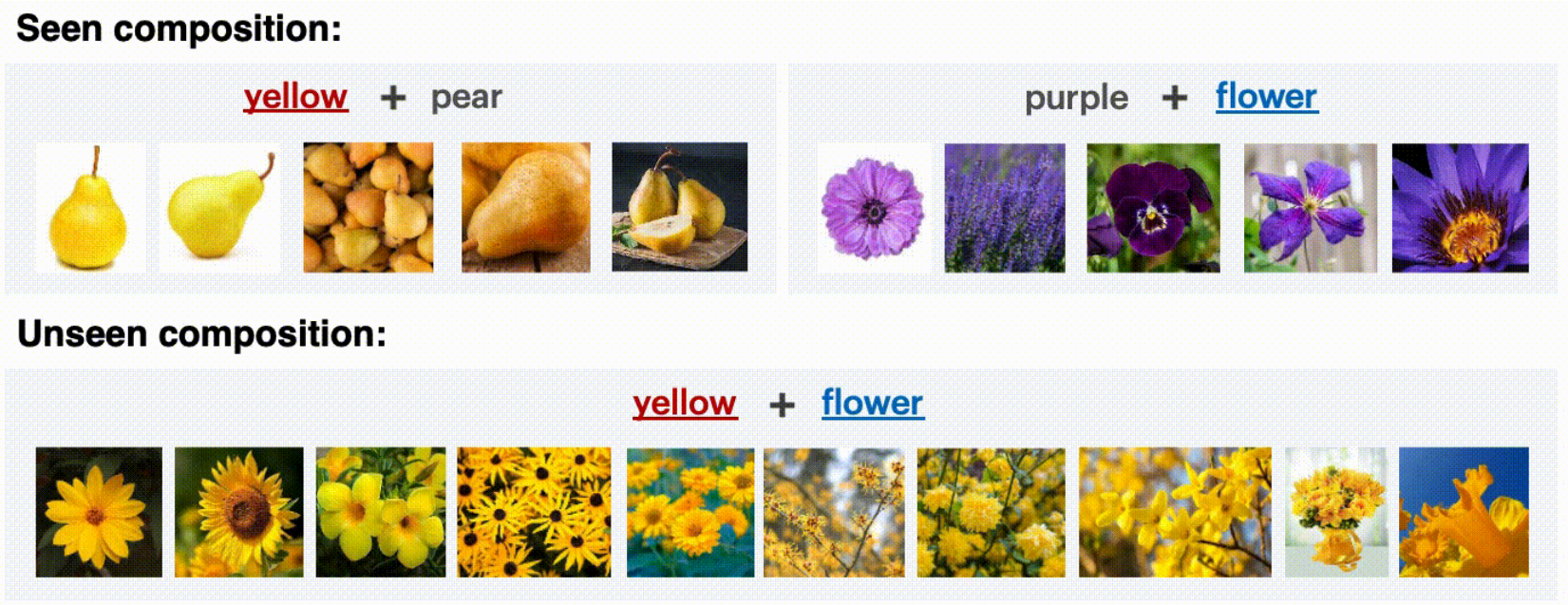
Siteng Huang[1], Qiyao Wei[2], Donglin Wang[3]✉

[1]Zhejiang University [2]University of Cambridge [3]Westlake University

ICMR 2023

Background

Compositional Zero-Shot Learning (CZSL)



[1]

- Given side information that describes novel composition pairs, CZSL aims to recognize **unseen** attribute-object compositions at test time while each constituent (i.e., **primitive**) exists in training samples.

- Relying on **sufficient seen compositions and training samples** to learn the compositionality.

[1] Shaozhe Hao, et al., Learning Attention as Disentangler for Compositional Zero-shot Learning, CVPR 2023

Motivation

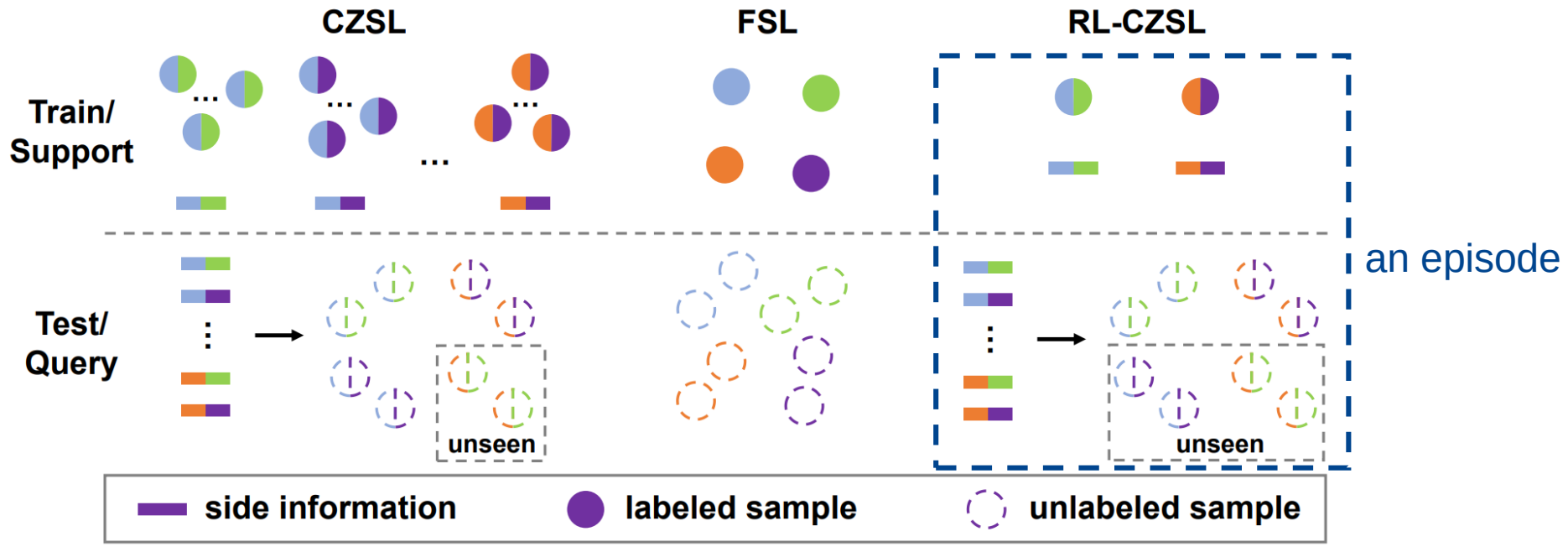
Two Core Challenges for Human-like Compositional Learning

- **Few-shot.** Humans can learn the compositionality of complex concepts with only a few examples.
- **Few referential compositions.** Humans can discover potential primitives from a few combinations, or even only one, based on prior knowledge.

- **Conclusion:** Existing CZSL setup is idealistic and inappropriate to simulate human-like compositional learning.

New Problem

Reference-Limited Compositional Zero-Shot Learning (RL-CZSL)



- Comparison among compositional zero-shot learning (**CZSL**), few-shot learning (**FSL**), and our proposed reference-limited compositional zero-shot learning (**RL-CZSL**).
- Different **colors** indicate different **categories of entities or primitives**.

Proposed Benchmarks

RL-CZSL-ATTR & RL-CZSL-ACT

Table 1: Basic statistics of our proposed datasets. The symbol # is used as an abbreviation for "number of".

Dataset	RL-CZSL-ATTR	RL-CZSL-ACT
Composition type	attribute-object	action-object
Total #c	1,768	574
#c in train / val / test	1,076 / 136 / 556	214 / 22 / 338
Total #p^1	190	185
#p^1 in train / val / test	105 / 33 / 52	52 / 10 / 123
Total #p^2	488	154
#p^2 in train / val / test	281 / 12 / 195	59 / 11 / 84
Total #samples	99,771	30,420
#samples in train / val / test	51,928 / 29,922 / 17,921	20,604 / 1,207 / 8,609

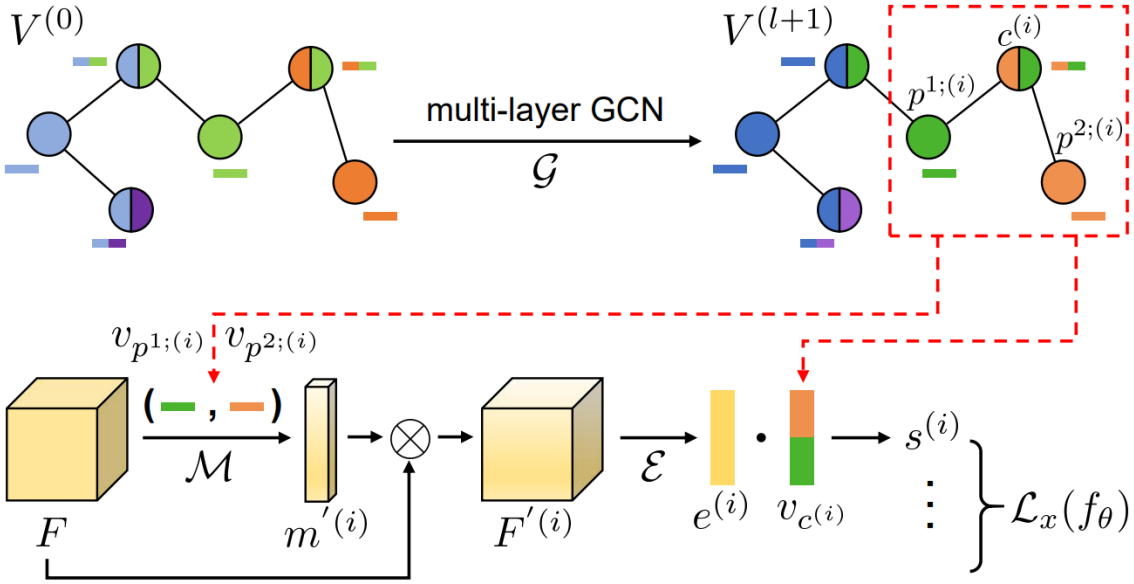
(a) RL-CZSL-ATTR: train (b) RL-CZSL-ATTR: val (c) RL-CZSL-ATTR: test

(d) RL-CZSL-ACT: train (e) RL-CZSL-ACT: val (f) RL-CZSL-ACT: test

Meta Compositional Graph Learner (MetaCGL)

Learning Composition Embeddings

- A graph that contains **all primitives** and **potential compositions** as nodes is constructed.
- The features of nodes are initialized with the pre-trained word embeddings.
- The nodes of two primitives and a composition are connected one by one if the composition is formed by the two primitives, and a self-loop is added to each node.
- A multi-layer graph convolutional network (**GCN**) is used to update node features.



Meta Compositional Graph Learner (MetaCGL)

Learning Visual Embeddings

- The learned primitive embeddings is transformed into a prior **channel-wise correlation map** to estimate which features are related to the prediction target.

1x1 convolution

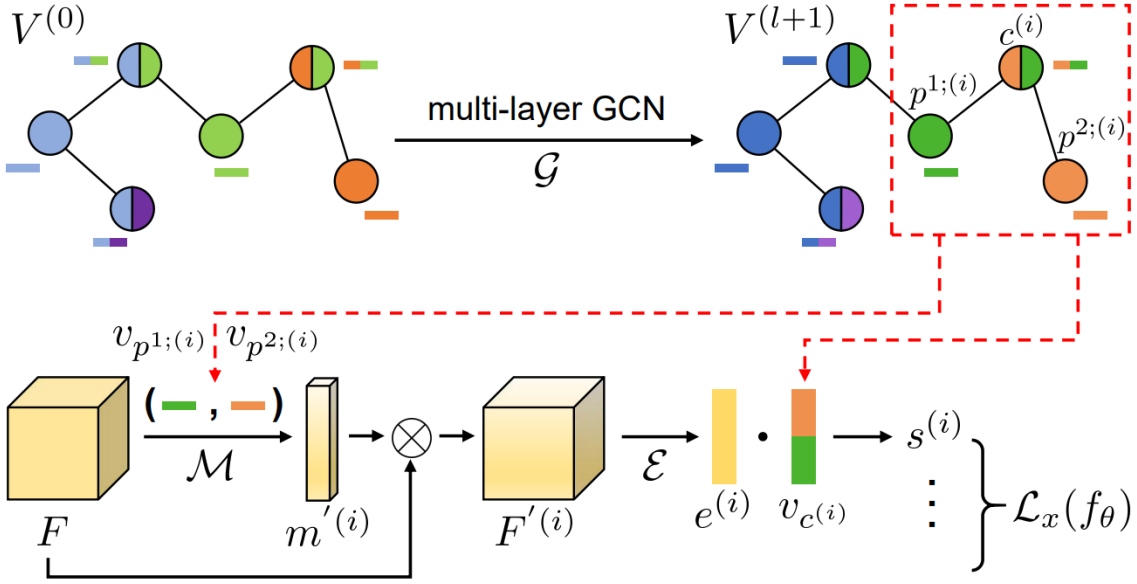
$$m'^{(i)} = \sigma(\mathcal{M}^1([\mathrm{GAP}(F); v_{p^{1;(i)}}]) + \mathcal{M}^2([\mathrm{GAP}(F); v_{p^{2;(i)}}])),$$

$$F'^{(i)} = m'^{(i)} \otimes F,$$

- A **compatibility score** can be calculated with the visual embedding and the learned composition embedding:

$$s^{(i)} = e^{(i)} \odot v_{c^{(i)}},$$

- The **inference** prediction can be made by applying a argmax operation on all compatibility scores.



- The **training loss**:

$$\mathcal{L}_x(f_\theta) = -\log\left(\frac{\exp(s^{(i)})}{\sum_j^{|C|}\exp(s^{(j)})}\right),$$

Meta Compositional Graph Learner (MetaCGL)

Training Strategy

Compositional Mixup Data Augmentation

- A new **augmented example** of image can be formed by a weighted linear interpolation of two random sampled query samples:

$$\tilde{x}_q^{(i)} = \lambda x_q^{(i)} + (1-\lambda) x_q^{(j)},$$

where $\lambda \in [0, 1]$ is randomly drawn from Beta(1, 1).

- The **augmented compositional label**:

$$\tilde{c}_q^{(i)} = \lambda^2 c_q^{(i)} + \lambda(1-\lambda) c_q^{(ij)} + \lambda(1-\lambda) c_q^{(ji)} + (1-\lambda)^2 c_q^{(j)},$$

where

$$c_q^{(ij)} = (p^{1;(i)}, p^{2;(j)}),\ c_q^{(ji)} = (p^{1;(j)}, p^{2;(i)}).$$

Bi-level Optimization

- All the trainable parameters can be updated as

$$\theta' = \theta - \epsilon \nabla_\theta \mathcal{L}_{\mathcal{S}}(f_\theta),$$
$$\theta \leftarrow \theta - \gamma \nabla_\theta \mathcal{L}_{\tilde{\mathcal{Q}}}(f_{\theta'}),$$

ϵ : the step size hyperparameter
γ : the meta step size hyperparameter
$\mathcal{S}$: support samples
$\tilde{\mathcal{Q}}$: augmented query samples

Experiments

Main Results

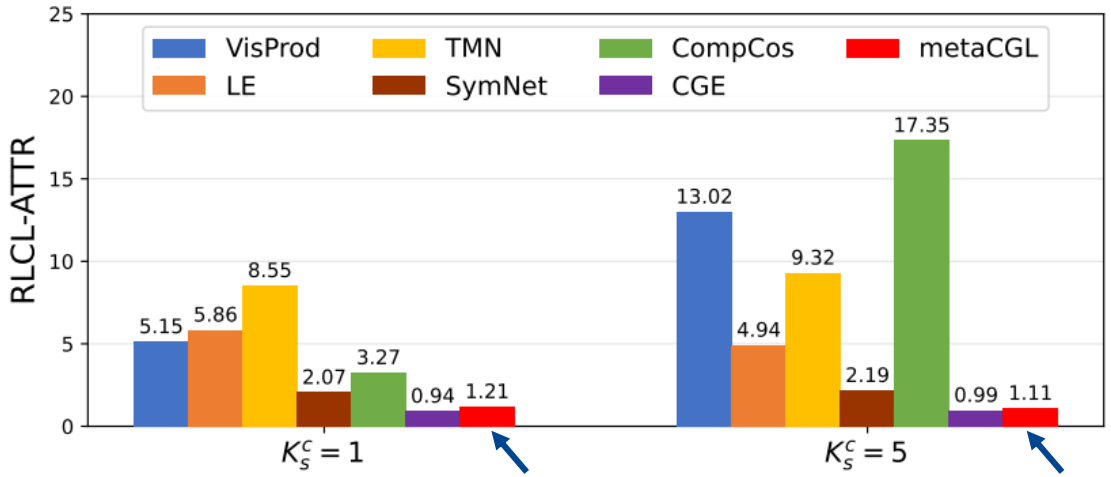
Table 2: Comparison with the baselines on two proposed benchmarks. Detailed results (%) are reported with 95% confidence intervals. Best results are displayed in boldface.

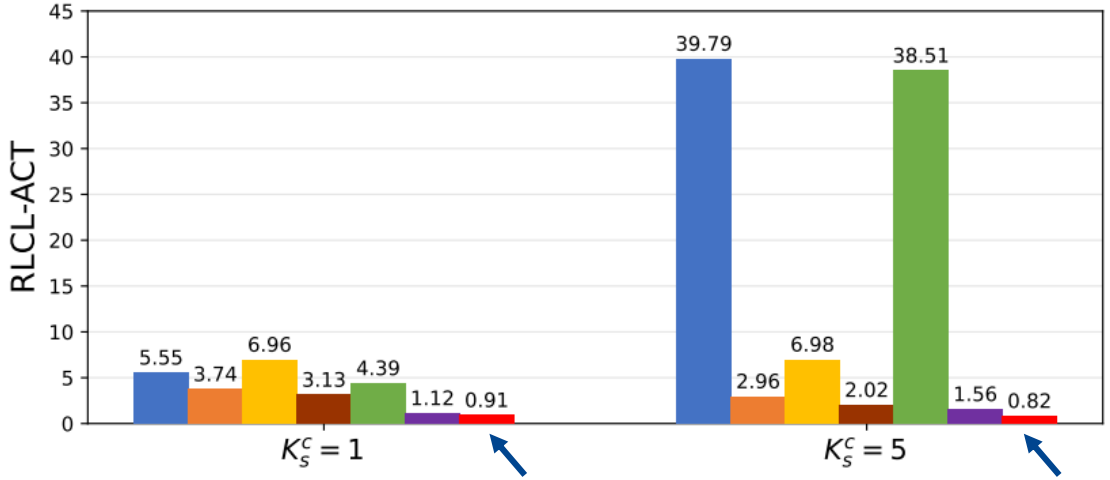
Method	RL-CZSL-ATTR					RL-CZSL-ACT				
	UA	SA	HM	Prim 1 HM	Prim 2 HM	UA	SA	HM	Prim 1 HM	Prim 2 HM
$K_s^c = 1$										
VisProd [21]	1.24 ±0.43	20.99 ±0.18	2.34 ±0.76	25.44 ±2.15	27.41 ±1.79	0.88 ±0.19	21.97 ±0.56	1.68 ±0.36	26.12 ±1.78	25.43 ±1.15
LE [21]	1.01 ±0.96	14.98 ±0.52	1.89 ±1.67	22.25 ±0.90	22.45 ±0.53	1.27 ±0.85	14.02 ±0.18	2.32 ±1.44	17.90 ±3.28	22.45 ±4.56
TMN [25]	0.52 ±0.50	**28.31** ±0.39	1.02 ±0.97	24.86 ±1.59	32.41 ±0.50	0.62 ±0.34	**28.85** ±0.10	1.21 ±0.65	31.76 ±0.37	26.03 ±1.55
SymNet [18]	1.94 ±0.08	17.34 ±0.80	3.48 ±0.12	27.01 ±1.05	23.95 ±2.87	2.28 ±1.71	17.90 ±0.56	4.01 ±2.72	27.35 ±1.59	23.02 ±2.82
CompCos [19]	2.57 ±0.55	25.14 ±0.70	4.66 ±0.93	26.84 ±0.72	33.53 ±2.39	3.02 ±0.34	28.19 ±0.55	5.45 ±0.56	**32.07** ±2.46	**28.51** ±2.57
CGE [22]	4.65 ±1.12	15.40 ±0.54	7.13 ±1.29	25.97 ±3.30	31.56 ±1.26	4.05 ±0.78	15.51 ±0.91	6.41 ±0.91	28.56 ±2.09	26.39 ±1.50
MetaCGL (Ours)	**10.44** ±0.42	19.01 ±1.78	**13.47** ±0.77	**30.22** ±1.58	**38.37** ±2.29	**7.76** ±0.31	15.95 ±1.24	**10.44** ±0.41	31.19 ±1.38	26.68 ±2.76
$K_s^c = 5$										
VisProd [21]	0.55 ±0.45	15.83 ±0.60	1.07 ±0.83	22.80 ±2.59	22.90 ±1.31	0.18 ±0.14	16.19 ±0.32	0.35 ±0.28	13.72 ±1.80	21.88 ±2.73
LE [21]	0.72 ±0.49	13.79 ±0.02	1.37 ±0.89	21.14 ±0.89	20.45 ±2.64	1.23 ±1.01	12.67 ±0.62	2.23 ±1.66	18.18 ±2.41	18.88 ±2.24
TMN [25]	0.27 ±0.16	**32.02** ±0.51	0.54 ±0.32	26.53 ±0.60	35.22 ±1.78	0.30 ±0.29	31.28 ±0.37	0.59 ±0.57	**34.03** ±1.07	27.35 ±1.43
SymNet [18]	1.96 ±0.95	18.47 ±0.68	3.54 ±1.54	27.24 ±2.03	24.47 ±2.57	2.28 ±1.71	17.90 ±0.56	4.01 ±2.72	27.35 ±1.59	23.02 ±2.82
CompCos [19]	1.05 ±0.23	31.62 ±0.54	2.03 ±0.44	27.91 ±2.68	34.71 ±2.69	1.16 ±0.55	**34.32** ±0.64	2.25 ±1.02	32.59 ±0.37	26.66 ±2.82
CGE [22]	4.10 ±1.09	17.03 ±0.13	6.61 ±1.43	25.06 ±0.20	31.04 ±0.84	2.73 ±0.78	19.12 ±0.65	4.78 ±1.17	25.57 ±2.48	23.05 ±1.11
MetaCGL (Ours)	**11.85** ±2.55	20.70 ±1.21	**15.05** ±1.81	**31.88** ±2.68	**40.41** ±1.25	**8.01** ±0.23	17.48 ±0.98	**10.99** ±0.23	32.93 ±1.73	**28.01** ±1.34

UA (Unseen Accuracy): Accuracy of Unseen Compositions. **SA** (Seen Accuracy): Accuracy of Seen Compositions. **HM** (Harmonic mean) = 2 (SA $*$ UA) / (SA + UA).

Experiments

Error Analysis





$\frac{U \rightarrow S}{U \rightarrow U}$: The ratios of whether error cases from unseen compositions are confused for seen pairs (U→S) or incorrect unseen pairs (U→U). The **lower** the value, the **smaller** the trend of overfitting.

Experiments

Ablation Study

Model	RL-CZSL-ATTR			RL-CZSL-ACT		
	UA	SA	HM	UA	SA	HM
w/o $\mathcal{G}$	8.64	11.90	9.99	5.16	6.55	5.77
w/o $\mathcal{M}$	11.97	20.13	15.02	7.23	16.15	9.99
w/o BO	1.25	13.70	2.28	0.85	13.88	1.60
w/o CM	10.82	**20.81**	14.23	6.52	**19.61**	9.78
Full	**11.85**	20.70	**15.05**	**8.01**	17.48	**10.99**

Table 3: Ablation study with various model configurations of MetaCGL. $\mathcal{G}$: the compositional graph. $\mathcal{M}$: the correlation map generating network. BO: the bi-level optimization strategy. CM: the Compositional Mixup data augmentation.

- The removal of any component from MetaCGL generally results in a worse performance on UA and HM.
- The **compositional graph** plays an important role in recognizing samples of **seen** compositions.
- The **bi-level optimization** significantly contributes to recognizing **unseen** compositions.
- The **Compositional Mixup** data augmentation improves UA while sacrificing SA.

Experiments

Effect of Equipping MAML

Table 4: Comparison with CZSL baselines equipped with MAML. Our MetaCGL still achieves the best UA and HM.

Method	$K_s^c = 1$						$K_s^c = 5$					
	RL-CZSL-ATTR			RL-CZSL-ACT			RL-CZSL-ATTR			RL-CZSL-ACT		
	UA	SA	HM	UA	SA	HM	UA	SA	HM	UA	SA	HM
VisProd [21]	1.24	20.99	2.34	0.88	21.97	1.68	0.55	15.83	1.07	0.18	16.19	0.35
+MAML	3.53	20.76	6.03	1.72	21.50	3.18	3.76	29.71	6.67	2.23	28.60	4.13
LE [21]	1.01	14.98	1.89	1.27	14.02	2.32	0.72	13.79	1.37	1.23	12.67	2.23
+MAML	4.49	6.24	5.21	6.14	7.93	6.91	6.06	14.15	8.48	5.68	8.06	6.64
SymNet [18]	1.94	17.34	3.48	2.28	17.90	4.01	1.96	18.47	3.54	2.96	17.12	5.04
+MAML	3.62	4.48	4.00	4.91	4.47	4.65	3.40	4.14	3.69	3.61	4.74	4.10
CompCos [19]	2.57	**25.14**	4.66	3.02	**28.19**	5.45	1.05	**31.62**	2.03	1.16	**34.32**	2.25
+MAML	3.17	5.93	4.07	2.81	6.44	3.90	2.98	7.67	4.28	3.54	8.76	5.02
CGE [22]	4.65	15.40	7.13	4.05	15.51	6.41	4.10	17.03	6.61	2.73	19.12	4.78
+MAML	9.44	18.62	12.54	6.06	17.86	9.05	11.09	19.76	14.21	6.73	18.17	9.82
MetaCGL (Ours)	**10.44**	19.01	**13.47**	**7.76**	15.95	**10.44**	**11.85**	20.70	**15.05**	**8.01**	17.48	**10.99**

Take-home message

- We introduce a new problem named **reference-limited compositional zero-shot learning (RL-CZSL)**, where given only **a few samples of limited compositions**, the model is required to generalize to recognize unseen compositions. This offers a **more realistic and challenging** environment for evaluating compositional learners.

- We establish **two benchmark datasets** with diverse compositional labels and well-designed data splits, providing the required platform for systematically assessing progress on the task.

- We propose a novel method, **Meta Compositional Graph Learner (MetaCGL)**, for the challenging RL-CZSL problem. Experimental results show that MetaCGL consistently outperforms popular baselines on recognizing unseen compositions.

Thank you for listening!

Reference-Limited Compositional Zero-Shot Learning

Siteng Huang, Qiyao Wei, Donglin Wang

Contact: huangsiteng@westlake.edu.cn

arXiv

Project page

Github