Context-based and Diversity-driven Specificity in Compositional Zero-Shot Learning

Yun Li[1] Zhe Liu[2] Hang Chen[3] Lina Yao[1]
[1] CSIRO's Data61 [2] Bytedance Ltd. [3] Snap Inc.
y.li@csiro.au zhe.liu01@bytedance.com chenhang386@gmail.com lina.yao@data61.csiro.au

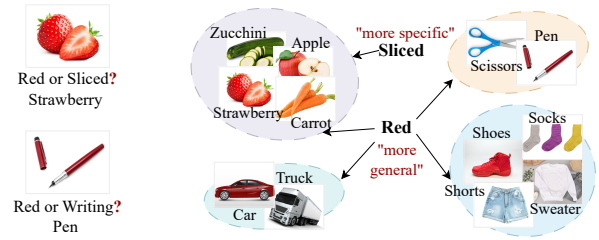
Abstract

Compositional Zero-Shot Learning (CZSL) aims to recognize unseen attribute-object pairs based on a limited set of observed examples. Current CZSL methodologies, despite their advancements, tend to neglect the distinct specificity levels present in attributes. For instance, given images of sliced strawberries, they may fail to prioritize 'Sliced-Strawberry' over a generic 'Red-Strawberry', despite the former being more informative. They also suffer from ballooning search space when shifting from Close-World (CW) to Open-World (OW) CZSL. To address the issues, we introduce the Context-based and Diversity-driven Specificity learning framework for CZSL (CDS-CZSL). Our framework evaluates the specificity of attributes by considering the diversity of objects they apply to and their related context. This novel approach allows for more accurate predictions by emphasizing specific attribute-object pairs and improves composition filtering in OW-CZSL. We conduct experiments in both CW and OW scenarios, and our model achieves state-of-the-art results across three datasets.

1. Introduction

Humans effortlessly combine known ideas, such as the *pink* of a rose and a blue *dolphin*, to recognize unseen concepts like a *pink dolphin*. This ability to learn compositionally is a hallmark of human intelligence [12], enabling us to infer vast knowledge from limited primitives without seeing every possible combination. Inspired by this capability, Compositional Zero-Shot Learning (CZSL) [8, 16, 24] emerged. In CZSL, the goal is to train models on images of seen attribute-object pairs (e.g., attributes like colors and objects like animals) so they can recognize unseen pairs, thereby minimizing the need for extensive training datasets.

Traditional CZSL methods often adopt one of two strategies: 1) they project attribute-object textual labels and images into a shared space for direct, similarity-based composition classification [24, 40, 41]; 2) they use dual modules



(a) Which is specific? (b) Diversity-driven specificity learning.

Figure 1. Specificity in CZSL. (a)For strawberries, *Sliced* is more specific than *Red*. Instead, *Red* is more specific than *Writing* for a pen, as *Writing* is its inherent function. (b) Clustering images based on their object features, *Red* spans multiple object clusters. In contrast, *Sliced*, though applicable to several objects, only links to the food cluster, indicating its greater specificity.

to classify attributes and objects separately, later fusing the results for the final composition [8, 15, 18]. Recently, their performance has been further improved by integrating large, pre-trained vision-language models like CLIP [32]. Harnessing the powerful visual-semantic aligning capabilities of CLIP, these methods [19, 27, 38] set new performance benchmarks that surpass traditional approaches.

While current CZSL methods show promise, they often focus on optimizing model performance while overlooking an inherent challenge in CZSL: specificity in attributes. Unlike objects, which generally have straightforward definitions, attributes can be multifaceted. Consider the example in Fig. 1a: a strawberry can be described as *Red*, a common descriptor, or *Sliced*, which is more specific. The value of such specific descriptors is supported by the Shannon Theory, which holds that rarer events offer more information than common ones [36]. However, in pursuit of overall accuracy and broader generalization to unseen pairs, CZSL models may favor general attributes like *Red*, which have wider applicability to almost all categories, over more specific yet valid descriptors like *Sliced*. The challenge of effectively prioritizing these specific attributes without compromising accuracy remains under-explored.

Beyond the challenge of specificity, transitioning from a Closed-World (CW) setting [16, 26, 30] to an Open-World (OW) setting [8, 21, 22] poses additional difficulties. In CW, test set compositions are predefined and given as prior knowledge, making it less realistic for real-world applications. The OW setting, free of such restrictions, faces a largely expanded output space. Direct composition classification methods, as a result, suffer significant performance drops in OW-CZSL [21]. To address this, some techniques narrow the search space by determining composition feasibility based on their similarity to seen pairs [19, 22, 27]. However, similarity-based feasibility estimation risks discarding specific attribute-object pairs in favor of more generic compositions. In contrast, methods that separately predict attributes and objects can avoid this expanding search space and achieve advanced learning of primitives, i.e., attributes and objects [7, 15, 18]. Yet, these methods might not grasp the contextual nuances present in the composition space, like how *Small* appears differently in contexts such as Small-Cat versus Small-House.

To address the aforementioned challenges in CZSL, we introduce the Context-based and Diversity-driven Specificity learning framework for CZSL (CDS-CZSL). Our framework employs a 3-branch structure: a composition-wise branch for contextual understanding through composition classification and two adapter-enhanced primitive-wise branches to extract attribute and object features effectively.

A distinguishing feature of our framework is the introduction of a context-based and diversity-driven specificity learner. Intuitively, the specificity of an attribute is linked to the range of objects it can describe. We thus cluster object features and drive the learning of an attribute's specificity using the diversity of clusters it covers, as depicted in Fig. 1b. We further incorporate the context of attributes into specificity learning. This is rooted in the observation that the specificity of an attribute can vary depending on the object it's paired with. For instance, while *Red* is general for strawberries, it is more specific than the inherent function attribute *Writing* for pen, as illustrated in Fig. 1a. Armed with the specificity insights, we refine attribute predictions to emphasize specific pairs.

This specificity also aids in pruning the composition space, filtering out both overly specific and overly generic pairs, alleviating the challenges of composition space explosion in the OW setting. Crucially, using specificity, our method can retain valid, specific pairs that other feasibility calibration techniques [19] might overlook.

In summary, our primary contributions include:
1) We propose CDS-CZSL, a novel CLIP-based CZSL method that employs a 3-branch structure to equip both contextual understanding and efficient attribute/object learning.
2) We introduce the specificity concept into attribute predictions for CZSL. The proposed specificity learner prioritizes specific attributes with accuracy holds. This specificity further enhances composition filtering in OW-CZSL setting.
3) Our model achieves state-of-the-art (SOTA) results on three benchmark datasets in both CW and OW scenarios.

2. Related Work

Compositional Zero-shot Learning (CZSL). CZSL [5, 9, 18, 25, 26, 43] has two main strategies for inferring unseen compositions. The first assumes that unseen and seen compositions share the same attribute and object scopes. Thus, this strategy first predicts primitive labels and then combines them to obtain the composition label [16, 18, 24, 26, 30]. For example, Liu *et al*. [18] separately predict attributes and objects based on contextual semantics to infer the compositions. Li *et al*. [15] disentangle attributes and objects with reversed attention. The second strategy directly predicts compositions by aligning images and textual labels in a shared space and searching for most similar compositions [13, 26, 35, 42]. For example, Nagarajan *et al*. [26] build a composition space by simulating all the visual changes of attributes performed on objects. Anwaar *et al*. [1] improve composition learning by building a composition graph. Recent approaches [19, 27, 38], rooted in Vision-Language Models (VLM), also adopt either of the two strategies, utilizing pre-trained VLM encoders to better encode and align images and texts. For example, Nayak *et al*. adapt prompt in CLIP [32] to fit the CZSL task. Lu *et al*. [19] further boost the performance by soft prompts and disentangling strategy. In this study, we unify two strategies and a fixed VLM backbone within a single model and further enhance the model with specificity-refined attribute learning and specificity-based composition filtering.

Vision-Language Model (VLM). VLM [14, 20, 31, 33, 34] demonstrates remarkable potential in addressing various vision and language tasks, such as visual question answering [28] and image captioning [44]. Recent approaches have enhanced VLM's compatibility with downstream tasks by incorporating small adapters into the network or customizing prompt engineering. Small adapters [19, 20, 33, 34] refer to additional layers added to VLMs. They can boost VLMs' performance on downstream tasks with minimal network parameters fine-tuned. For instance, Mahabadi *et al*. [20] introduce additional layers in each transformer block for multi-task fine-tuning. Meanwhile, prompt engineering [3, 14, 19, 27, 38] enhance large pre-trained models like CLIP [32] and GPT [2] by changing prompt guidance. Prompts can be static text or learnable word embeddings, aiming to help models quickly adapt to new tasks with little or none retraining. For instance, Wang *et al*. [38] utilize hierarchical prompts to enhance CLIP's performance in CZSL. In our model, we investigate both adapters and prompt engineering to improve the performance.

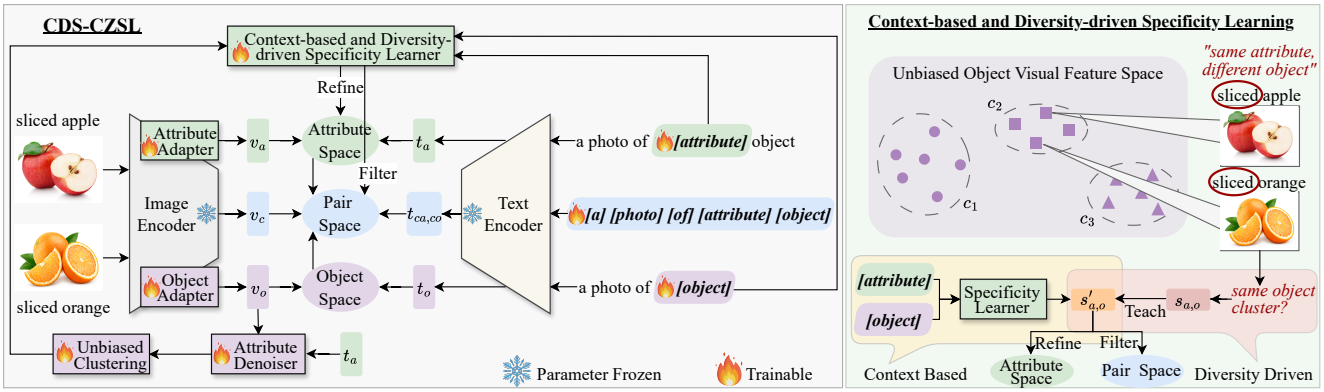


Figure 2. CDS-CZSL Overview and process of the context-based and diversity-driven specificity learning.

3. Method

Problem definitions and notations. In CZSL, images are modeled as compositions of primitives, i.e., attributes $a \in \mathcal{A}$ and objects $o \in \mathcal{O}$. This leads to a label space $Y = \mathcal{A} \times \mathcal{O}$, capturing all possible attribute-object combinations. Within this space, we distinguish between seen compositions $Y^{\mathcal{S}}$ and unseen compositions $Y^{\mathcal{U}}$, with $Y^{\mathcal{S}} \cap Y^{\mathcal{U}} = \emptyset$. During training, we have access to data from seen compositions: $\mathcal{S} = \{(x, y) | x \in X^{\mathcal{S}}, y \in Y^{\mathcal{S}}\}$, where each image x is labeled with an attribute-object pair $y = (a, o)$. Then, during testing, the model needs to predict labels for images from both seen and unseen compositions. Depending on the scope of the output label space, we have two test settings: CW-CZSL and OW-CZSL [21]. In CW, $Y^{\mathcal{U}}$ is given as prior knowledge, and the testing label space is restricted to $y \in Y^{\mathcal{S}} \cup Y^{\mathcal{U}}$, while in OW, the output space expands to all potential attribute-object pairs, i.e., $y \in Y$.

Overview. To tackle CZSL, we propose the CDS-CZSL framework. It consists of three distinct branches, tailored for predictions within attribute, object, and pair spaces. Through pre-trained image f_i and text f_t encoders, we project images and composition labels into the common pair space for direct composition predictions. For attribute and object predictions, we enhance their visual representations with an attribute adapter f_a and an object adapter f_o, both anchored on the image encoder f_i, and improve their semantic understandings with two specialized prompts. We further design a context-based and diversity-driven specificity learner f_s to 1) refine the attribute learning process, thus enabling the prioritization of specific attributes, and 2) filter out undesired compositions in the OW pair space.

3.1. Composition-wise Learning

As mentioned before, contextuality is crucial for CZSL [23, 26] due to the diverse appearances of attributes and objects across varied compositions. Addressing this, we adapt CSP [27] as the composition-wise learning branch. The main idea is to create a unified space to which both images and composition labels are projected. A similarity search is conducted within this space to find the most compatible visual and semantic representations. This branch utilizes transformer-based encoders to project both visuals and labels. Notably, these encoders, pre-trained using Contrastive Language Image Pre-Training (CLIP) [32], are kept frozen throughout our training.

To effectively harness the power of CLIP for our CZSL task, we reformat composition labels into structured natural language prompts like *[a] [photo] [of] [**attribute**] [**object**]*. Different from CSP [27], our prompt is entirely soft in that each word in the prompt is modeled as learnable parameters $\theta_c = \{w_0, w_1, w_2, w_{ca}, w_{co}\}$, where the first three parameters represent prefix word embeddings. In contrast, the last two represent attribute and object word embeddings. Using the text encoder f_t, these embeddings yield a text representation $t_{ca,co}$; in parallel, the image encoder f_i processes the given image x to produce the image representation v_c:

$$t_{ca,co} = f_t(\theta_c), v_c = f_i(x) \tag{1}$$

Finally, we normalize both text and image representations using $\ell 2$-normalization: $t_{ca,co} = \frac{t_{ca,co}}{\|t_{ca,co}\|}$ and $v_c = \frac{v_c}{\|v_c\|}$, and obtain the probability for class $y = (ca, co)$ by:

$$p(y = (ca, co)|x) = \frac{exp(v_c \cdot t_{ca,co}/\tau)}{\sum_{(\acute{ca},\acute{co}) \in Y^{\mathcal{S}}} exp(v_c \cdot t_{\acute{ca},\acute{co}}/\tau)} \tag{2}$$

where τ is a temperature parameter from CLIP. During training, we optimize θ_c to produce better composition-level representations by minimizing the cross-entropy loss:

$$\min_{\theta_c} \mathcal{L}_{base} = -\frac{1}{|\mathcal{S}|} \sum_{(x,y) \in \mathcal{S}} \log p(y|x) \tag{3}$$

where, $|\mathcal{S}|$ denotes the number of instances in the seen set.

Though composition-wise learning captures the compositional contextuality and excels in extracting composition-level representations, it struggles when generalized to OW

seniors [15, 17]. Hence, we propose primitive-wise learning for distinct attribute and object learning.

3.2. Primitive-wise Learning

In this section, we introduce our primitive-wise learning branches. Distinct from composition-level insights of the composition-wise branch, primitive-wise learning ensures a deeper understanding of individual attributes and objects. It also lays the foundation for subsequent specificity learning.

For semantic primitive understanding, similar to HPL [38], we adopt two prompts: *a photo of **[attribute]** object* (additional suffix, i.e., *object*, is inserted to ensure sentence completeness) for attribute learning and *a photo of **[object]*** for object learning. We denote the prompts as $\theta_a = \{e_0, e_1, e_2, w_a, e_3\}$ for attributes and $\theta_o = \{e_0, e_1, e_2, w_o\}$ for objects. Unlike the full soft composition prompt θ_c, in θ_a and θ_o, only the attribute and object word embedding w_a and w_o are learnable during training, and the remainings are fixed. This ensures the two branches focus exclusively on attribute/object information. Primitive-level text representations are then obtained by:

$$t_a = f_t(\theta_a), t_o = f_t(\theta_o) \tag{4}$$

For image representations, a Multi-Head Self-Attention layer [37] is introduced after the image encoder, functioning as the attribute/object adapter. Distinct adapters, attribute adapter f_a and object adapter f_o, are employed, allowing the derivation of unique visual representations for attributes and objects respectively:

$$v_a = f_a(f_i(x)), v_o = f_o(f_i(x)) \tag{5}$$

Finally, we normalize the representations $t_a = \frac{t_a}{\|t_a\|}$, $t_o = \frac{t_o}{\|t_o\|}$, $v_a = \frac{v_a}{\|v_a\|}$, and $v_o = \frac{v_o}{\|v_o\|}$, and compute attribute and object probabilities as follows:

$$p(a|x) = \frac{exp(v_a \cdot t_a/\tau)}{\sum_{\hat{a}\in\mathcal{A}} exp(v_a \cdot t_{\hat{a}}/\tau)} \tag{6}$$

$$p(o|x) = \frac{exp(v_o \cdot t_o/\tau)}{\sum_{\hat{o}\in\mathcal{O}} exp(v_o \cdot t_{\hat{o}}/\tau)} \tag{7}$$

The learned visual representations for objects v_o and predicted attribute probabilities $p(a|x)$ are then fed into our Context-based and Diversity-driven Specificity Learner to learn specificity in attributes.

3.3. Context-based and Diversity-driven Specificity Learning

To determine attribute specificity, our approach is based on the intuition that an attribute's specificity inversely relates to the range of object clusters it describes. Note that we consider an attribute's descriptive diversity across object clusters rather than individual categories. This is because, taking Fig. 1b as an example, both 'Sliced' and 'Red' apply to several object, but 'Sliced' is more specific as it describes similar fruits, such as apples and strawberries, while 'Red' is less specific because it applies to a broader range of unrelated objects, like apples and cars.

When clustering objects, attributes entangled in their representations can lead to biased clusters—'Red Car' and 'Red Apple' might cluster together due to the shared attribute. To address this, we apply an attribute denoising step before clustering, ensuring that objects are grouped by their inherent characteristics rather than by shared attributes.

Attribute Denoiser. We design the Attribute Denoiser following the idea in [18] that image representation v_o can be seen as attribute-denoised if we cannot infer the attribute from v_o. Thus, we first infer the attribute using v_o by:

$$p(a|v_o) = \frac{exp(v_o \cdot t_a/\tau)}{\sum_{\hat{a}\in\mathcal{A}} exp(v_o \cdot t_{\hat{a}}/\tau)} \tag{8}$$

Then we introduce a denoising loss $\mathcal{L}_{den}$ calculated by the mean square error to guide the predicted attribute probability towards a uniform distribution:

$$\min_{f_o} \mathcal{L}_{den} = \frac{1}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} (\frac{1}{|\mathcal{A}|} - p(a|v_o))^2 \tag{9}$$

Note that we detach the gradient of t_a during the calculation of $\mathcal{L}_{den}$ to prevent the denoising loss from compromising the precision of the text representation of the attribute.

Diversity-driven Specificity Learner. After achieving unbiased object representations v_o, we aim to cluster v_o to infer attributes' descriptive diversity. Directly clustering the entire dataset every time v_o is updated during training would require storing all instances of v_o and recalculating clusters after each update, a process both computationally and temporally prohibitive. To address this, we opt for batch-level cluster updates (using K-Means clustering [29]), and infer diversity by each time randomly selecting pairs of images x and $\hat{x}$ that share attributes but differ in objects. And then, if their object representations v_o and $\hat{v}_o$ are in the same cluster, it indicates that objects linked to this attribute share visual similarities, resulting in low diversity. Conversely, different clusters signify high diversity. The probability of co-clustering serves as an indirect measure of attributes' descriptive diversity across the dataset.

We then estimate the specificity as binary: it is 1 if v_o and $\hat{v}_o$ cluster together, indicating low descriptive diversity of attribute a and thus high specificity, and 0 if they do not co-cluster, indicating low specificity. For an image x with the label pair (a, o), the specificity $s_{a,o}$ is quantified as follows:

$$s_{a,o} = \begin{cases} 1 & \text{if } \Gamma(v_o) = \Gamma(\hat{v}_o) \\ 0 & \text{if } \Gamma(v_o) \neq \Gamma(\hat{v}_o) \end{cases} \tag{10}$$

where Γ denotes the K-Means algorithm.

However, specificity is also context-dependent, as discussed in Sec. 1. Therefore, we do not use $s_{a,o}$ directly as the specificity. Instead, we introduce a specificity learner, f_s, which employs word embeddings w_a and w_o to predict the context-based specificity $s'_{a,o}$ and use $s_{a,o}$ as the target of f_s. This approach ensures that our specificity learner, f_s, is context-based and guided by diversity. We employ the cross-entropy loss function to optimize f_s:

$$\min_{f_s} \mathcal{L}_{div} = -\frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \log p(s_{a,o}|x) \tag{11}$$

Specificity-refined Primitive Learning. To encourage the model to prioritize attributes with higher specificity in its predictions, we introduce specificity predicted by f_s as a penalty term. Then, the specificity-refined attribute prediction and the specificity-refined primitive loss $\mathcal{L}_{prim}$ can be expressed as follows:

$$p'(a|x) = p(a|x) - f_s(w_a, w_o)/\gamma \tag{12}$$

$$\min_{w_a,w_o,f_a,f_o} \mathcal{L}_{prim} = -\frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} [\log p'(a|x) + \log p(o|x)] \tag{13}$$

where, $\gamma = \frac{1}{|\mathcal{A}|}$ adjusts the range of the penalty term.

During training, a larger specificity penalty term results in a lower probability of the refined attribute prediction $p'(a|x)$. Then, to minimize $\mathcal{L}_{prim}$, the model leans towards assigning a higher probability to $p(a|x)$ (Eq. (6)) to remedy the penalty term. By optimizing $\mathcal{L}_{prim}$, this specificity penalty term amplifies $p(a|x)$ if (a, o) exhibits high specificity. Consequently, when we remove this penalty term during testing, the probability $p(a|x)$ is elevated for attributes with high specificity, increasing the likelihood of predicting specific attribute a. Conversely, if the attribute a is general, its smaller specificity penalty terms during training lead to lower prediction probabilities for a when we remove the penalty terms during testing.

3.4. Refined Prediction

During training, we utilize the specificity-refined attribute and object probabilities to enhance the base model, yielding the fused prediction of our model:

$$\begin{aligned} p'(y|x) &= \alpha p(y = (ca, co)|x) + (1-\alpha)p'(y = (a, o)|x) \\ &= \alpha p(y = (ca, co)|x) + (1-\alpha)p'(a|x)p(o|x) \end{aligned} \tag{14}$$

The refined prediction loss for our model is given by:

$$\min_{\theta_c,w_a,w_o,f_a,f_o} \mathcal{L}_{refine} = -\frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \log p'(y|x) \tag{15}$$

Dataset	a	o	Training sp	Training i	Testing sp	Testing up	Testing i
MIT-States	115	245	1262	30k	400	400(27k)	13k
UT-Zappos	16	12	83	23k	18	18(109)	3k
C-GQA	413	674	5592	27k	888	923(272k)	5k

Table 1. Statistics of datasets: a, o, i, sp, and up are the number of attributes, objects, images, seen pairs, and unseen pairs. Numbers in brackets are the unseen search space in OW.

3.5. Training and Inference

Training Objectives. To ensure that specificity learning does not affect the optimization of representations, our training process is divided into two phases: representation learning and specificity learning. The representation learner encompasses adapters and learnable word embeddings, denoted as θ_c, w_a, w_o, f_a, and f_o. The specificity learner is represented as f_s. The overall training loss is defined as follows:

$$\min_{\theta_c,w_a,w_o,f_a,f_o} \mathcal{L}_{base}+\mathcal{L}_{prim}+\mathcal{L}_{den}+\mathcal{L}_{refine}; \qquad \min_{f_s} \mathcal{L}_{div} \tag{16}$$

Initially, we optimize the representation learner exclusively for one epoch, ensuring reliable representations for subsequent specificity learning. All object image representations are retained as initial inputs for the K-Means clustering. Subsequently, we iteratively optimize the representation learner, K-means kernel, and specificity learner in a batch-wise manner.

Inference. During testing, we exclude the penalty term in Eq. (12) to obtain attribute probability (Eq. (6)) and fuse primitive predictions and composition predictions as the final results:

$$y' = \arg\max_{y\in Y^{\mathcal{T}}} \alpha p(y = (ca, co)|x) + (1-\alpha)p(y = (a, o)|x) \tag{17}$$

where, $Y^{\mathcal{T}} = Y^{\mathcal{S}} \cup Y^{\mathcal{U}}$ in the CW-CZSL, and $Y^{\mathcal{T}} = \mathcal{A}\times\mathcal{O}$ in the OW-CZSL.

Additionally, we employ specificity to filter out less likely pairs, reducing the search space for compositions in the open world. This strategy is not used in the closed-world scenario where all pairs are feasible. Specifically, we introduce two thresholds to retain compositions with moderate specificity, such that $Y^{\mathcal{T}} = \{y = (a, o) : s_{a,o} \in [t_l, t_h]\}$, excluding compositions that are overly specific or overly general, where $t_l < t_h$.

4. Experiments

4.1. Experiment Settings

Datasets and evaluation metrics. We evaluate our model on three benchmark datasets: 1) MIT-States [6] features

	Closed-World	MIT-States				UT-Zappos				C-GQA			
		S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
w/o CLIP	CGE [24]	32.8	28.0	21.4	6.5	64.5	71.5	**60.5**	33.5	31.4	14.0	14.5	3.6
	CompCos [21]	25.3	24.6	16.4	4.5	59.8	62.5	43.1	28.7	28.1	11.2	12.4	2.6
	Co-CGE [22]	32.1	28.3	20.0	6.6	62.3	66.3	48.1	33.9	33.3	14.9	14.4	4.1
	SCEN [13]	29.9	25.2	18.4	5.3	63.5	63.1	47.8	32.0	28.9	25.4	17.5	5.5
	CANet [39]	29.0	26.2	17.9	5.4	61.0	66.3	47.3	33.1	30.0	13.2	14.5	3.3
	CoT [10]	34.8	31.5	23.2	7.8	-	-	-	-	34.0	18.8	17.5	5.1
	ADE [4]	-	-	-	-	63.0	64.3	51.1	35.1	35.0	17.7	18.0	5.2
w CLIP	CLIP [32]	30.2	46.0	26.1	11.0	15.8	49.1	15.6	5.0	7.5	25.0	8.6	1.4
	CLIP-based Co-CGE [22]	46.7	45.9	33.1	17.0	63.4	71.3	49.7	36.3	34.1	21.2	18.9	5.7
	CoOP [45]	34.4	47.6	29.8	13.5	52.1	49.3	34.6	18.8	20.5	26.8	17.1	4.4
	CSP [27]	46.6	49.9	36.3	19.4	64.2	66.2	46.6	33.0	28.8	26.8	20.5	6.2
	HPL [38]	47.5	50.6	37.3	20.2	63.0	68.8	48.2	35.0	30.8	28.4	22.4	7.2
	DFSP [19]	46.9	52.0	37.3	20.6	**66.7**	71.7	47.2	36.0	38.2	32.0	27.1	10.5
	CDS-CZSL (ours)	**50.3**	**52.9**	**39.2**	**22.4**	63.9	**74.8**	52.7	**39.5**	**38.3**	**34.2**	**28.1**	**11.1**

Table 2. Model performance in CW. We use 'w' and 'w/o' to distinguish models adopting CLIP as visual and language encoders or not. The best results are in **bold**. The second best results are in blue.

natural objects with diverse attributes. Its relatively noisy data and fine-grained attributes make it challenging to learn; 2) UT-Zappos is a specialized, small-scale dataset centered on footwear, encompassing a limited set of 16 attributes and 12 objects; 3) C-GQA stands out with its expansive vocabulary, 413 attributes and 674 objects. Such breadth presents a formidable challenge, particularly in OW scenarios. Datasets are split into seen and unseen compositions following the split in previous works [24, 30] to ensure fair comparisons. The splits are based on generalized CZSL setting [24] to ensure both seen and unseen compositions appear during testing. Split details are in Tab. 1.

We evaluate our model using the protocol from [21, 30]. Varied biases are added to unseen pairs to adjust testing results, during which the best-seen accuracy (S), best-unseen accuracy (U), best harmonic mean (HM), and Area Under the Curve (AUC) of the HM-bias curve are recorded as performance indicators. Among these, AUC is the core metric as it evaluates the model comprehensively.

Implement details. We follow prior practices [19, 27] to adopt CLIP [32] as our image/text encoder. Object and attribute adapters are one-layer Multi-head Attention. The specificity learner is a four-layer Fully-Connected Network(FCN). We use K-means to cluster object representations, and the clustering is batch-wise to save computation costs. The cluster number is calculated using the base-2 logarithm of the object classes, resulting in numbers of 8, 4, and 10 for three datasets, respectively. The model is trained end-to-end with Adam optimizer [11]. Hyper-parameters, such as learning rate and fusion weight (α), are determined based on validation set performance. The supplementary provides codes with detailed parameters.

4.2. Comparisons with SOTAs

We compare our CDS-CZSL with most recent CZSL methods [4, 8, 10, 13, 15, 18, 19, 21, 22, 22, 27, 32, 39, 45] in both CW and OW settings. Given the same data splits and evaluation metrics, the reported performances from the original publications are directly used for competitors. The results of all CLIP-based methods are run with ViT-L/14 with CLIP fixed during training.

The CW results, presented in Tab. 2, reveal that CDS-CZSL achieves the best results on all datasets. Specifically, it yields improvements of 1.8%, 3.2%, and 0.6% in AUC over the second-best methods on three datasets. It also attains considerable gains in HM on MIT-States and C-GQA, with increases of 1.9% and 1%, respectively. Although CGE outperforms it in HM on UT-Zappos, the overall performance of CDS-CZSL surpasses that of CGE, particularly in AUC, with a 6% improvement.

Comparing methods with and without CLIP as backbones, we observe that CLIP-based methods possess higher performance ceilings. However, our enhancements are not solely due to CLIP. Compared with other CLIP-based methods, our performance still prevails. Firstly, unlike CSP [27] and DFSP [19], which only predict in the composition space, our approach also leverages separate learning of attributes and objects, thereby exhibiting better generalization capabilities. This is also proved by our model achieving the highest unseen accuracy (U) across all datasets. Secondly, we introduce context-based and diversity-driven specificity learning that prioritizes informative specific attributes, thus facilitating more accurate predictions.

The OW results, depicted in Tab. 3, indicate a significant performance drop for all methods transitioning from CW to

	Open-World	MIT-States				UT-Zappos				C-GQA			
		S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
w/o CLIP	CGE [24]	32.4	5.1	6.0	1.0	61.7	47.7	39.0	23.1	32.7	1.8	2.9	0.47
	CompCos [21]	25.4	10.0	8.9	1.6	59.3	46.8	36.9	21.3	28.4	1.8	2.8	0.39
	Co-CGE [22]	30.3	11.2	10.7	2.3	61.2	45.8	40.8	23.3	32.1	3.0	4.8	0.78
	KG-SP [8]	28.4	7.5	7.4	1.3	61.8	52.1	42.3	26.5	31.5	2.9	4.7	0.78
	SAD-SP [18]	29.1	7.6	7.8	1.4	63.1	54.7	44.0	28.4	31.0	3.9	5.9	1.00
	DRANet [15]	29.8	7.8	7.9	1.5	65.1	54.3	44.0	28.8	31.3	3.9	6.0	1.05
	ADE [4]	-	-	-	-	62.4	50.7	44.8	27.1	35.1	4.8	7.6	1.42
w CLIP	CLIP [32]	30.1	14.3	12.8	3.0	15.7	20.6	11.2	2.2	7.5	4.6	4.0	0.27
	CLIP-based Co-CGE [22]	38.1	20.0	17.7	5.6	59.9	56.2	45.3	28.4	33.2	3.9	5.3	0.91
	CoOP [45]	34.6	9.3	12.3	2.8	52.1	31.5	28.9	13.2	21.0	4.6	5.5	0.70
	CSP [27]	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.20
	HPL [38]	46.4	18.9	19.8	6.9	63.4	48.1	40.2	24.6	30.1	5.8	7.5	1.37
	DFSP [19]	47.5	18.5	19.3	6.8	**66.8**	60.0	44.0	30.3	**38.3**	7.2	10.4	2.40
	CDS-CZSL w filtering (ours)	**49.4**	**21.8**	**22.1**	**8.5**	**64.7**	**61.3**	**48.2**	**32.3**	37.6	**8.2**	**11.6**	**2.68**

Table 3. Model performance in OW. 'w filtering' indicates that the reported CDS-CZSL uses specificity to filter compositions.

OW. Yet, CDS-CZSL continues to outperform in all criteria (bar the seen accuracy in C-GQA) across datasets. Particularly, it achieves improvements of 1.6%, 2%, and 0.28% in AUC. While the absolute increments are modest compared to the CW scenario, the relative improvements are higher on MIT-States (23.2% vs. 8.7%) and C-GQA (11.7% vs. 5.7%), confirming the effectiveness of our model. In addition to the aforementioned reasons in CW analysis, our specificity-based filtering strategy also contributes to the improvements in the OW setting. The lower relative improvements on UT-Zappos for OW (6.7%) compared to CW (8.8%) may be due to its small-scale and fine-grained nature, making our specificity learning and filtering less effective. However, the model still benefits from our design of composition-wise and primitive-wise learning, hence outperforming other methods. Notably, our model achieves the best U in OW as well; this demonstrates that our model's specificity learning does not compromise the model's generalization ability to unseen compositions.

4.3. Ablation Study

Module ablation. In our ablation study, we evaluate the efficacy of each component in our proposed CDS-CZSL by comparing it against four variants. The variant *SPM* contains only the composition-wise learning branch, while *3branch* extends *SPM* by incorporating primitive-wise learning branches for both attributes and objects but without the specificity learning component. To assess the impact of specificity learning, we introduce *CBS*, which integrates context-based specificity into *3branch*, and *DDS*, which incorporates diversity-driven specificity. Fig. 3 shows differences among CBS, DDS, and CDS.

The results, presented in Tab. 4, illustrate several key findings. Firstly, the comparison between SPM and CSP [27]—which differs from SPM in terms of whether the prompt is fully learnable—highlights the efficiency gains afforded by fully learnable prompts within the composition-wise branch. Furthermore, 3branch demonstrates an improvement over SPM, validating the importance of including primitive-wise learning. Significantly, our model outperforms HPL [38], which also employs a three-branch structure. This proves the advantages of our unique prompt designs and the integration of adapters in primitive learning.

Introducing specificity learning through both CBS and DDS leads to further performance gains over 3branch, confirming the effectiveness of specificity. The superior performance of CBS over DDS suggests that context-based adaptability provides additional advantages over the cluster-based, less-learnable diversity strategy of DDS. However, DDS's success indicates the importance of descriptive di-

		MIT-States				UT-Zappos			
		S	U	HM	AUC	S	U	HM	AUC
CW	SPM	44.1	51.7	35.8	19.1	64.3	66.3	47.3	33.8
	3branch	47.1	**53.2**	37.5	21.1	64.5	71.3	49.0	36.2
	CBS	50.0	52.5	38.4	22.0	**65.1**	**75.0**	52.5	39.3
	DDS	49.6	52.5	38.2	21.8	63.2	73.5	52.2	38.6
	CDS	**50.3**	52.9	**39.2**	**22.4**	63.9	74.8	**52.7**	**39.5**
OW	SPM	45.8	16.7	18.3	6.1	63.2	52.0	44.1	27.1
	3branch	45.0	21.1	20.6	7.4	61.9	60.0	45.2	29.9
	CBS	48.3	**22.1**	21.8	8.3	63.5	59.0	47.9	31.0
	DDS	**50.1**	21.2	21.7	8.3	**66.8**	56.0	45.4	29.5
	CDS	49.4	21.8	**22.1**	**8.5**	64.7	**61.3**	**48.2**	**32.3**

Table 4. Module ablation study. CDS denotes our CDS-CZSL with '-CZSL' removed to save space.

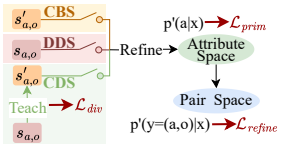


Figure 3. Module ablation.

	UT-Zappos					C-GQA				
	feasible pairs	S	U	HM	AUC	feasible pairs	S	U	HM	AUC
w/o filtering	192	64.7	60.3	47.9	32.0	278362	-	-	-	-
w similarity-based filtering [19]	189	64.7	60.4	48.0	32.1	67293	38.2	8.0	10.9	2.61
w our filtering	176	64.7	61.3	48.2	32.3	30332	37.6	8.2	11.6	2.68

Table 5. Ablation on filtering strategy.

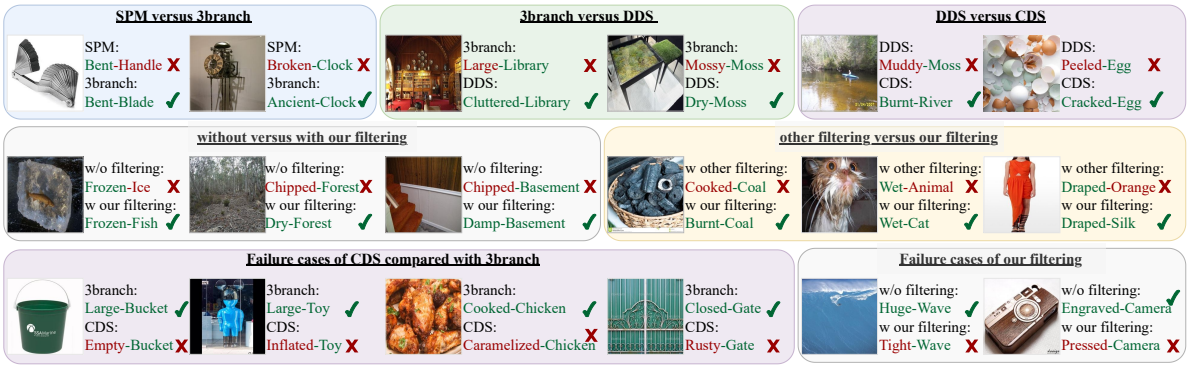


Figure 4. Qualitative Results of varying network structure (first row), changing filtering method (second row), and failure cases (third row).

versity in ascertaining attribute specificity. Combining the strengths of both context and diversity, our complete model, CDS-CZSL, achieves the highest results in terms of HM and AUC. This validates that both context and descriptive diversity are vital components of specificity learning in CZSL.

Filtering ablation. To assess the effectiveness of our specificity-based filtering strategy in OW, we compare it against no filtering and similarity-based filtering [19]. Filtering thresholds are chosen on validation sets. Due to GPU limitations, we excluded unfiltered results for C-GQA. The results, detailed in Tab. 5, show that while both filtering reduces the search space and increases performance, our specificity-based approach retains fewer but more relevant pairs, leading to superior HM and AUC. This proves its ability to effectively discard both overly general and highly specific pairs, unlike the similarity-based method, which may preserve generic pairs and omit unique, specific ones.

4.4. Qualitative Results

Module ablation. We study the qualitative results to explore the effects of varying network structures in the first row of Fig. 4. Initially, SPM is misled by seen pairs (e.g., Bent-Handle), while 3branch overcomes this through primitive-wise learning but struggles with too-general pairs like Mossy-Moss. DDS learning corrects this, refining such pairs to more specific Dry-Moss. CDS-CZSL further improves DDS with contexts, as seen when Peeled-Egg is adjusted to Cracked-Egg.

Filtering. The efficacy of our filtering strategy is shown in the second row; it can exclude overly general pairs, such as Frozen-Ice, and too specific ones, like Chipped-Basement. This contrasts with similarity-based filtering that might settle for broader categories, resulting in predictions like Wet-Animal instead of the more precise Wet-Cat.

Limitations and potentials. The third row in Fig. 4 shows cases where CDS-CZSL fails compared to 3branch and non-filtered version. It occasionally predicts specific pairs unsupported by the image content—for instance, we cannot tell if the Large-Bucket is Empty. Moreover, our filtering method is not unmistakable, retaining specific yet invalid pairs such as Pressed-Camera. However, CDS-CZSL, while not always accurate, often provides more descriptive labels than the original—refining Cooked-Chicken to Caramelized-Chicken, for example. This suggests that our model may have the potential to aid in refining labels.

5. Conclusion

In this work, we propose a Context-based and Diversity-driven Specificity learning framework for Compositional Zero-Shot Learning (CDS-CZSL). We incorporate composition-wise and primitive-wise learning to capture compositional contextuality and enhance attribute/object learning simultaneously. We then design the context-based and diversity-driven specificity learning to prioritize specific attributes that are more informative and use the learned specificity to filter compositions in Open-World scenarios. Through comprehensive experiments, we demonstrate the effectiveness of our model and achieve SOTA performance on three datasets. Furthermore, we discuss the limitations and potentials of our model leaning towards specific pairs.

References

[1] Muhammad Umer Anwaar, Zhihui Pan, and Martin Kleinsteuber. On leveraging variational graph embeddings for open world compositional zero-shot learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4645–4654, 2022. 2

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[3] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. Ppt: Pre-trained prompt tuning for few-shot learning. *arXiv preprint arXiv:2109.04332*, 2021. 2

[4] Shaozhe Hao, Kai Han, and Kwan-Yee K Wong. Learning attention as disentangler for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15315–15324, 2023. 6, 7

[5] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14646–14655, 2021. 2

[6] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015. 5

[7] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Revisiting visual product for compositional zero-shot learning. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. 2

[8] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9336–9345, 2022. 1, 2, 6, 7

[9] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–251, 2018. 2

[10] Hanjae Kim, Jiyoung Lee, Seongheon Park, and Kwanghoon Sohn. Hierarchical visual primitive experts for compositional zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5675–5685, 2023. 6

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[12] Brenden M Lake. *Towards more human-like concept learning in machines: Compositionality, causality, and learning-to-learn*. PhD thesis, Massachusetts Institute of Technology, 2014. 1

[13] Xiangyu Li, Xu Yang, Kun Wei, Cheng Deng, and Muli Yang. Siamese contrastive embedding network for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9335, 2022. 2, 6

[14] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 2

[15] Yun Li, Zhe Liu, Saurav Jha, and Lina Yao. Distilled reverse attention network for open-world compositional zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1782–1791, 2023. 1, 2, 4, 6, 7

[16] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11316–11325, 2020. 1, 2

[17] Zhe Liu, Yun Li, Lina Yao, Julian McAuley, and Sam Dixon. Rethink, revisit, revise: A spiral reinforced self-revised network for zero-shot learning. *arXiv preprint arXiv:2112.00410*, 2021. 4

[18] Zhe Liu, Yun Li, Lina Yao, Xiaojun Chang, Wei Fang, Xiaojun Wu, and Abdulmotaleb El Saddik. Simple primitives with feasibility- and contextuality-dependence for open-world compositional zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2023. 1, 2, 4, 6, 7

[19] Xiaocheng Lu, Song Guo, Ziming Liu, and Jingcai Guo. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23560–23569, 2023. 1, 2, 6, 7, 8

[20] Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. *arXiv preprint arXiv:2106.04489*, 2021. 2

[21] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5222–5230, 2021. 2, 3, 6, 7

[22] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Learning graph embeddings for open world compositional zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 6, 7

[23] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801, 2017. 3

[24] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 953–962, 2021. 1, 2, 6, 7

[25] Muhammad Ferjad Naeem, Evin Pınar Örnek, Yongqin Xian, Luc Van Gool, and Federico Tombari. 3d compositional zero-shot learning with decompositional consensus. In *European Conference on Computer Vision*, pages 713–730. Springer, 2022. 2

[26] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185, 2018. 2, 3

[27] Nihal V. Nayak, Peilin Yu, and Stephen H. Bach. Learning to compose soft prompts for compositional zero-shot learning. In *International Conference on Learning Representations*, 2023. 1, 2, 3, 6, 7

[28] Maria Parelli, Alexandros Delitzas, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. Clip-guided vision-language pre-training for question answering in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5606–5611, 2023. 2

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 4

[30] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3593–3602, 2019. 2, 6

[31] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6, 7

[33] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017. 2

[34] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127, 2018. 2

[35] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13658–13667, 2022. 2

[36] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 1

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[38] Henan Wang, Muli Yang, Kun Wei, and Cheng Deng. Hierarchical prompt learning for compositional zero-shot recognition. IJCAI, 2023. 1, 2, 4, 6, 7

[39] Qingsheng Wang, Lingqiao Liu, Chenchen Jing, Hao Chen, Guoqiang Liang, Peng Wang, and Chunhua Shen. Learning conditional attributes for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2023. 6

[40] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3741–3749, 2019. 1

[41] Guangyue Xu, Parisa Kordjamshidi, and Joyce Y Chai. Zero-shot compositional concept learning. *arXiv preprint arXiv:2107.05176*, 2021. 1

[42] Ziwei Xu, Guangzhi Wang, Yongkang Wong, and Mohan S Kankanhalli. Relation-aware compositional zero-shot learning for attribute-object pair recognition. *IEEE Transactions on Multimedia*, 2021. 2

[43] Muli Yang, Chenghao Xu, Aming Wu, and Cheng Deng. A decomposable causal view of compositional zero-shot learning. *IEEE Transactions on Multimedia*, 2022. 2

[44] Youyuan Zhang, Jiuniu Wang, Hao Wu, and Wenjia Xu. Distinctive image captioning via clip guided group optimization. In *European Conference on Computer Vision*, pages 223–238. Springer, 2022. 2

[45] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 6, 7