

# TsCA: On the Semantic Consistency Alignment via Conditional Transport for Compositional Zero-Shot Learning

Miaoge Li<sup>1</sup>, Jingcai Guo<sup>1\*</sup>, Richard Yi Da Xu<sup>3</sup>, Dongsheng Wang<sup>2</sup>, Xiaofeng Cao<sup>4</sup>,  
Zhijie Rao<sup>1</sup>, Song Guo<sup>5</sup>

<sup>1</sup>Department of COMP/LSGI, The Hong Kong Polytechnic University, Hong Kong SAR

<sup>2</sup>College of Computer Science and Software Engineering, Shenzhen University, China

<sup>3</sup>Department of Mathematics, Hong Kong Baptist University, Hong Kong SAR

<sup>4</sup>School of Computer Science and Technology, Tongji University, China

<sup>5</sup>Department of CSE, Hong Kong University of Science and Technology, Hong Kong SAR  
jc-jingcai.guo@polyu.edu.hk

## Abstract

Compositional Zero-Shot Learning (CZSL) aims to recognize novel *state-object* compositions by leveraging the shared knowledge of their primitive components. Despite considerable progress, effectively calibrating the bias between semantically similar multimodal representations, as well as generalizing pre-trained knowledge to novel compositional contexts, remains an enduring challenge. In this paper, our interest is to revisit the conditional transport (CT) theory and its homology to the visual-semantics interaction in CZSL and further, propose a novel **Tri**sets **C**onsistency **A**lignment framework (dubbed TsCA) that well-addresses these issues. Concretely, we utilize three distinct yet semantically homologous sets, i.e., *patches*, *primitives*, and *compositions*, to construct pairwise CT costs to minimize their semantic discrepancies. To further ensure the consistency transfer within these sets, we implement a cycle-consistency constraint that refines the learning by guaranteeing the feature consistency of the self-mapping during transport flow, regardless of modality. Moreover, we extend the CT plans to an open-world setting, which enables the model to effectively filter out unfeasible pairs, thereby speeding up the inference as well as increasing the accuracy. Extensive experiments are conducted to verify the effectiveness of the proposed method. The code is available at <https://github.com/keepgoingjkg/TsCA>.

## 1 Introduction

Identifying new concepts from a set of seen primitives is one of the fundamental challenges for AI systems to imitate the human learning process. Imagine a new cuisine—*Spicy Chocolate Cake*. Although it may sound like an odd combination, one can still perceive its appearance and taste based on existing knowledge of ‘*spiciness*’ and ‘*Chocolate*’. Likewise, compositional zero-shot learning (CZSL) [Misra *et al.*, 2017;

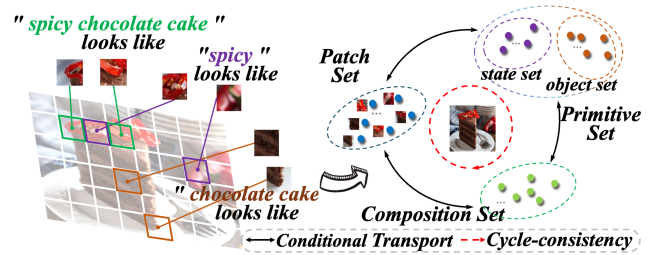


Figure 1: We represent each image as a set of patch embeddings and two sets of textual embeddings and employ semantic consistency conditional transport to align such cross-modal distribution trio.

Naeem *et al.*, 2021] emerges as a compelling paradigm that seeks to endow machines with similar cognitive abilities, allowing them to understand the composed primitives (*i.e.*, states and objects) during training and generalize to unseen compositions for inference.

Technically, studies on CZSL have revolved around achieving fine-grained alignment across the *image* and composed *state-object text* domains [Huynh and Elhamifar, 2020]. Pioneer methods typically load pre-trained image encoders and word embeddings to extract multi-modal features. Various alignment tricks are then applied to constrain the shared latent space, including concept learning [Xu *et al.*, 2021], geometric properties [Li *et al.*, 2020], semantic transformation [Nagarajan and Grauman, 2018], and graph embedding [Mancini *et al.*, 2022]. Moving beyond these post-processing alignments, a series of prompt-tuning methods have been developed to explore pre-trained vision-language models (VLMs) for the CZSL scenario [Xu *et al.*, 2022; Nayak *et al.*, 2023], achieving state-of-the-art prediction performance. Pre-trained on massive amounts of image-text pairs, VLMs (*e.g.*, CLIP [Radford *et al.*, 2021]) show impressive abilities to match the visual image with its textual label (prompt) in the multimodal space. Those models often employ a multi-path paradigm for fine-grained textual features, *e.g.*, one composition and two primitive branches. One of the core challenges ensues: *How to align the image with such multiple textual labels?* Recent studies attempt to

\*Corresponding author: Jingcai Guo.

address it from different fields, such as hierarchical prompt searching [Wang *et al.*, 2023], decomposed feature fusion [Lu *et al.*, 2023], and multi-step observation [Li *et al.*, 2023a]. Although showing attractive results, these methods focus primarily on *image-to-composition* (or *to-primitive*) alignments, ignoring intrinsic relationships within the composition and its primitives. This may fail to capture semantic consistency among the *image-composition-primitive* interactions, leading to suboptimal alignments.

To address the above issues, this paper proposes the **Trisets Consistency Alignment** framework (dubbed  $T_sCA$ ), which is built upon a novel Consistency-aware Conditional Transport (CCT) derived from the new view of CZSL. As shown in Fig. 1, we represent an image in three directions, *i.e.*,  $P_1$ , a distribution over all visual patches;  $P_2$ , a distribution over the composition set; and  $P_3$ , a distribution over the primitive set. Concretely,  $P_1$  captures the detailed visual features of an image, while  $P_2$  and  $P_3$  denote the global and local textual concepts of the same content. Therefore, the task of CZSL can be viewed as aligning these three discrete distributions as closely as possible. Accordingly, it is indeed key to properly measure the distance between empirical distributions with different supports. Fortunately, conditional transport (CT), well-examined in recent research, offers a bidirectional measurement of the distance from one distribution to another, given the cost matrix. Inheritedly,  $T_sCA$  extends the minimization of CT to the alignment of our cross-modal distribution trio.

Specifically,  $T_sCA$  gracefully facilitates the matching of *image-composition-primitive* triplets by meticulously crafting three pairs of CTs, thereby aligning the rich tapestry of cross-modal and intra-modal semantics. First,  $CT(P_1, P_2)$  measures the transport distance between the visual patch set and the composition set. On the one hand, it helps the label be transported to the compositional visuals with higher probabilities, and on the other hand, it regularizes the finetuning for better cross-modal alignments. Next,  $CT(P_1, P_3)$  focuses on discovering the corresponding attribute and object patches from the disentangle perspective. Last, unlike the above two,  $CT(P_2, P_3)$  aims at intra-modal interactions, which are often overlooked by previous studies. By minimizing the transport distance between the composition and its primitives explicitly, the textual outputs are assumed to present high semantic coherence. *E.g.*, the composition vectors are closer to their primitives in the embedding space. Moreover, the learned transport plan in  $CT(P_2, P_3)$  further provides an efficient tool to filter out unfeasible compositions, which benefits the open-world setting.

More importantly, as discussed above,  $T_sCA$  formulates the CZSL as the alignment of three distributions and seeks to minimize their transport costs. An intuitive solution is to run pairwise transportation and optimize the above three CT costs. Unfortunately, this case could not model the complex relations among these sets. To capture deeper interactions, we extend the naive CT to a well-designed consistency-aware CT (CCT) for the CZSL triplet case. Motivated by the fact that these three distributions describe the same semantics, CCT regularizes the product of three transport plans as a diagonal matrix from clockwise direction (shown in Fig. 1). In other

words, the composition label will return to itself after a cycle of transport, which ensures semantic consistency across sets. In summary, our contributions are three-fold:

- We formulate the CZSL task as a transport problem and propose **Trisets Consistency Alignment** model ( $T_sCA$ ), which views an image as three discrete distributions over the patch, composition, and primitive spaces and tries to explore fine-grained alignments between these triplets.
- We develop consistency-aware CT (CCT), considering the semantic consistency across the image-composition-primitive sets, improving the alignment robustness.
- Extensive comparisons and ablations on three benchmarks demonstrate the effectiveness of the proposed  $T_sCA$  with competitive performance on all settings.

## 2 Related Work

### 2.1 Compositional Zero-Shot Learning

CZSL[Mancini *et al.*, 2021] is a specialized case of zero-shot learning (ZSL) [Liu *et al.*, 2023]. Given the same set of objects with associated states, it focuses on recognizing unseen state-object compositions by learning from seen compositions. In general, traditional CZSL can be divided into two streams, *i.e.*, compositional classification and simple primitive classification. The former directly predicts compositions by aligning visual features and composed labels in a shared space, or resorting to a graph network for contextuality modeling [Tanwisuth *et al.*, 2023; Saini *et al.*, 2022; Li *et al.*, 2022]. Conversely, the latter identifies states and objects independently and constructs the joint compositional probability distribution. They either disregard the contextuality between primitives or impose training-specific correlations that are detrimental to generalization.

Recently, equipped with prompt learning, large-scale VLMs like CLIP are empowered to adapt to CZSL by reducing domain shift and leveraging pre-trained knowledge.

Further, researchers attempt to explore merging those two typical paradigms to create an integrated multi-path paradigm. For instance, HPL [Wang *et al.*, 2023] learns three hierarchical prompts by explicitly fixing the unrelated word tokens in the three embedding spaces at different levels. Troika [Huang *et al.*, 2024] effectively aligns the branch-specific prompt representations and decomposed visual features with a cross-modal traction module. Our  $T_sCA$  aligns closely with this paradigm, although with a greater emphasis on direct inquiries into semantic alignment within distributions.

### 2.2 Conditional Transport

Recently, CT has acted as an efficient tool to measure the distance between two discrete distributions [Zheng and Zhou, 2021]. Through learning bidirectional transport plans based on semantic similarity between samples from both source and target distributions, CT achieves fine-grained matching with the mini-batch optimization. More importantly, CT can integrate seamlessly with deep-learning frameworks and its effectiveness has been widely demonstrated in various applications [Li *et al.*, 2023b; Wang *et al.*, 2022; Tanwisuth *et al.*, 2023]. For example, [Tian *et al.*, 2023]

propose a novel CT-based imbalanced transductive few-shot learning model to fully exploit unbiased statistics of imbalanced query samples. Similarly, [Tanwisuth *et al.*, 2021] employs a probabilistic bidirectional transport between target features and class prototypes for unsupervised domain adaptation, showcasing robustness against class imbalance and facilitating domain adaptation without direct access to the source data. For the first time, through empirical demonstration, we establish that CT, boasting attributes like lower complexity and better scalability, is equally viable for CZSL.

### 3 Methodology

The proposed model aims to solve the CZSL from fine-grained alignments under the CT framework, where images are viewed as three discrete distributions over the image, composition, and primitive spaces. A novel consistency-aware CT is further developed to explore deeper interactions of these three domains. In this section, we start with the task of CZSL, and then introduce how to formulate CZSL as a CT problem in detail. The framework of our approach is shown in Fig. 2.

#### 3.1 Problem Formulation

Given state set  $\mathcal{S} = \{s_0, s_1, \dots, s_{|\mathcal{S}|}\}$  together with object set  $\mathcal{O} = \{o_0, o_1, \dots, o_{|\mathcal{O}|}\}$ , we can define the label space with their Cartesian product,  $\mathcal{C} = \mathcal{S} \times \mathcal{O}$ . Then  $\mathcal{C}$  can be divided into two disjoint label subsets such that  $\mathcal{C}^{se} \in \mathcal{C}$ ,  $\mathcal{C}^{us} \in \mathcal{C}$ , and  $\mathcal{C}^{se} \cap \mathcal{C}^{us} = \emptyset$  where  $\mathcal{C}^{se}$  and  $\mathcal{C}^{us}$  are the set of the seen and unseen classes respectively. Specifically, during the training phase,  $\mathcal{C}^{se}$  are used to train a discriminative model  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{C}^{se}$  from the input image space to candidate composition label set. At inference time, the model is expected to predict unseen compositions in the test sample space, i.e.,  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{C}^{test}$ . In this paper, we follow the setting of Generalized ZSL [Xian *et al.*, 2017], considering testing samples contain both seen and unseen compositions. In general, in the closed-world evaluation, only the known composition space of test samples is required as  $\mathcal{C}^{test} = \mathcal{C}^{se} \cup \mathcal{C}^{us}$ . For the open-world evaluation [Karthik *et al.*, 2022], the model has to consider all possible permutations of the state-object pairs, i.e.,  $\mathcal{C}^{test} = \mathcal{C}$ .

#### 3.2 Image as Three Sets

Built upon the pre-trained CLIP and with the soft prompt tuning technique, our TS-CA represents an image as three sets: visual patch set, textual composition set, and primitive set. These sets capture the multimodal features of the same content from different semantic domains, acting as a fundamental role in the fine-grained alignments.

**Patch Set.** For an input image  $x$ , CLIP first splits it into  $N$  non-overlapping patches evenly and then feeds them (with a [CLS] token inserted) into the image encoder to extract the embedding of the [CLS] token  $\mathbf{x}_{CLS}^c$  and the patch features  $\mathbf{x} = \{\mathbf{x}_n\}_{n=1}^N \in \mathbb{R}^{d \times N}$ , where  $\mathbf{x}_n \in \mathbb{R}^d$  denotes the embedding of patch  $n$  with the embedding dimension being  $d$ . Naturally, the discrete probability distribution over the patch set can be formulated as:

$$\mathbf{P}_1 = \sum_{n=1}^N \theta_n \delta_{\mathbf{x}_n}, \quad \theta_n = \frac{1}{N}, \quad (1)$$

where  $\delta$  denotes the Dirac delta function. We view all the patches equally and employ the Uniform distribution to model the patch weight  $\theta$ . Note that  $\mathbf{P}_1$  collects the detailed visual features of the local region, thereby bringing benefits to discriminative representation learning, especially when different concepts require emphasis on distinct areas within the input images. Besides, the [CLS] visual token  $\mathbf{x}_{CLS}^c$  is also learned, serving as the global representation, which is often used as the image feature to downstream tasks [Zhou *et al.*, 2022]. Here, we consider it as the visual prior that constructs subsequent textual sets.

**Composition Set.** In the textual domain, we build composition labels into a learnable soft prompt  $[v1][v2][v3][state][object]$ , where  $[v]$  is the prefix vectors, and  $[state][object]$  denote the state and object name, respectively. Then, the prompt will be fed into the text encoder to obtain textual representations  $\mathbf{y} = \{\mathbf{y}_m\}_{m=1}^M \in \mathbb{R}^{d \times M}$ , where  $M$  is the number of training compositions. To embrace the variety of content and minimize the cross-modal disparities, we follow previous work [Huang *et al.*, 2024] and incorporate a residual component obtained through a cross-attention mechanism [Vaswani, 2017] with patch embeddings (we here still use  $\mathbf{y}$  as the output compositions):

$$\mathbf{y} = \text{Cross-Att}(\mathbf{y}^{in}, \mathbf{x}, \mathbf{x}) + \mathbf{y}^{in}, \quad (2)$$

where  $\mathbf{y}^{in}$  denotes the input compositions. Cross-Att is the cross-attention layer with the query  $\mathbf{y}^{in}$ , key  $\mathbf{x}$ , and value  $\mathbf{x}$  as inputs, and outputs the fused features. As a result, the composition set can be viewed as:

$$\mathbf{P}_2 = \sum_{m=1}^M \alpha_m \delta_{\mathbf{y}_m}, \quad \alpha = \sigma(\mathbf{y}^T \mathbf{x}_{CLS}^c), \quad (3)$$

where  $\sigma$  denotes the softmax function. We calculate the composition weights  $\alpha$  via the semantic similarity of the composition label and visual feature. This helps  $\mathbf{P}_2$  to focus on compositions that describe the input image well.

**Primitive Set.** Unlike the composition set that focuses on global textual features, the primitive set aims to explore the disentangled state and object representations. Motivated by the recent multi-path paradigms, we learn the corresponding primitive representations through  $[v1][v2][v3][state]$  and  $[v1][v2][v3][object]$ , with the outputs denoted as  $\mathbf{z}^s \in \mathbb{R}^{d \times |\mathcal{S}|}$  and  $\mathbf{z}^o \in \mathbb{R}^{d \times |\mathcal{O}|}$  respectively. The corresponding probability weights are calculated as:

$$\beta^s = \sigma((\mathbf{z}^s)^T \mathbf{x}_{CLS}^s), \quad \beta^o = \sigma((\mathbf{z}^o)^T \mathbf{x}_{CLS}^o), \quad (4)$$

where  $\mathbf{x}_{CLS}^s$  and  $\mathbf{x}_{CLS}^o$  are two adapted visual features, which can be obtained via a lightweight adapter  $g$ :

$$[\mathbf{x}_{CLS}^s, \mathbf{x}_{CLS}^o] = g(\mathbf{x}_{CLS}^c), \quad (5)$$

where  $g$  takes the image feature  $\mathbf{x}_{CLS}^c$  as input, and outputs the state/object-relevant features, deriving unique visual representations. To complete the primitive set, we concatenate the feature points from both state and object labels as  $\mathbf{z} = \{\mathbf{z}_k\}_{k=1}^K \in \mathbb{R}^{d \times K}$ , where  $K = |\mathcal{S}| + |\mathcal{O}|$  is the total number of state and object labels. Like compositions that

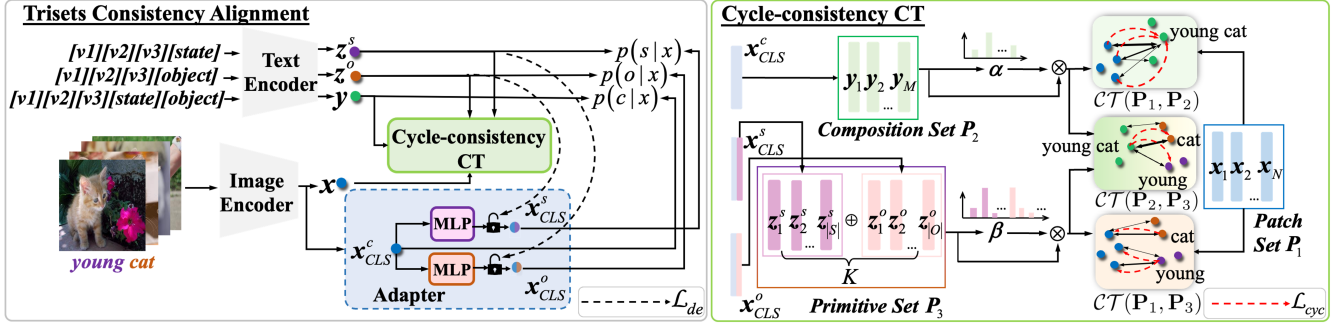


Figure 2: The overall framework of the proposed TSsCA (zoom-in for more details).

fuse the patch features with Eq. 2,  $z$  is fed into the same cross-attention layer:

$$z = \text{Cross-Att}(z^{in}, x, x) + z^{in}. \quad (6)$$

Finally, the primitive set can be expressed as:

$$\mathbf{P}_3 = \sum_{k=1}^K \beta_k \delta_{z_k}, \quad \beta = \sigma(\beta^s \oplus \beta^o), \quad (7)$$

where  $\oplus$  denotes the concatenation operation. Together with  $\mathbf{P}_2$ , these two sets provide primitive-composition textual knowledge for downstream alignment tasks.

### 3.3 Semantic Consistency Alignment

Based on the three carefully constructed sets of the input image, our TSsCA formulates the CZSL task as fine-grained alignments under the consistency-aware CT framework, which consists of pairwise CT distance and cycle consistency regularization.

**Pairwise CT Distance.** Given the source distribution  $\mathbf{P}_1$  and the target distribution  $\mathbf{P}_2$ , CT measures the distance by calculating the total transport costs bidirectionally, leading to the forward CT and backward CT, respectively. Denoting  $c(\mathbf{x}_n, \mathbf{y}_m) \geq 0$  as a cost function to define the difference between points  $\mathbf{x}_n$  and  $\mathbf{y}_m$ , the forward CT is measured as the expected transport cost of transporting all points from  $\mathbf{P}_1$  to the target  $\mathbf{P}_2$ , and the backward CT reverses the transport direction. Mathematically, the CT distance between  $\mathbf{P}_1$  and  $\mathbf{P}_2$  can be expressed as:

$$\begin{aligned} \text{CT}(\mathbf{P}_1, \mathbf{P}_2) &= \min_{\vec{\mathbf{T}}} \left( \sum_{n,m} \vec{t}_{nm} c_{nm} + \sum_{m,n} \overleftarrow{t}_{mn} c_{mn} \right), \quad (8) \\ \text{s.t. } \vec{\mathbf{T}} \mathbf{1}^N &= \boldsymbol{\theta}, \quad \overleftarrow{\mathbf{T}} \mathbf{1}^M = \boldsymbol{\alpha}, \end{aligned}$$

where we employ the cosine distance as the cost function, *i.e.*, the closer the patch  $\mathbf{x}_n$  and composition  $\mathbf{y}_m$  are, the lower the transport cost.  $\mathbf{1}^M$  is the  $M$  dimensional vector of ones.  $\mathbf{T}$  is called the transport plan, which is to be learned to minimize the total distance. The forward transport plan, *e.g.*,  $\vec{t}_{nm}$  in  $\vec{\mathbf{T}}$  describes the transport probability from the source point  $\mathbf{x}_n$  to target point  $\mathbf{y}_m$ :

$$\vec{t}_{nm} = \theta_n \frac{\alpha_m e^{s_\psi(\mathbf{x}_n, \mathbf{y}_m)}}{\sum_{m'=1}^M \alpha_{m'} e^{s_\psi(\mathbf{x}_n, \mathbf{y}_{m'})}}, \quad (9)$$

where  $s_\psi$  denotes the similarity function with learnable parameter  $\psi$ , and we specify it as  $s_\psi(\mathbf{x}_n, \mathbf{y}_m) = \frac{\mathbf{x}_n^T \mathbf{y}_m}{\exp(\psi)}$ . That is, the patch  $\mathbf{x}_n$  will transport with higher probability if the composition  $\mathbf{y}_m$  shares similar semantics with it. Similarly, we have the backward transport expressed as:

$$\overleftarrow{t}_{mn} = \alpha_m \frac{\theta_n e^{s_\psi(\mathbf{y}_m, \mathbf{x}_n)}}{\sum_{n'=1}^N \theta_{n'} e^{s_\psi(\mathbf{y}_m, \mathbf{x}_{n'})}}. \quad (10)$$

More interestingly, the definition in Eq. 9-10 naturally satisfies the constraint in Eq. 8, simplifying the CT optimization.

By combining  $\mathbf{P}_1$ ,  $\mathbf{P}_2$ , and  $\mathbf{P}_3$  pairwise, we extend CT to the CZSL scenario and denote the total CT distances as:

$$\text{CT} = \text{CT}(\mathbf{P}_1, \mathbf{P}_2) + \text{CT}(\mathbf{P}_1, \mathbf{P}_3) + \text{CT}(\mathbf{P}_2, \mathbf{P}_3), \quad (11)$$

where the first two terms facilitate cross-domain alignments but focus on different semantic levels.  $\text{CT}(\mathbf{P}_1, \mathbf{P}_2)$  takes the patch set and composition set as inputs and aims to find compositional visuals that match its textual label. While  $\text{CT}(\mathbf{P}_1, \mathbf{P}_3)$  pays more attention to primitive alignment. Intuitively, the primitive set provides decoupled textual guidance, and it helps the model extract patches that contain the state or object-relevant visuals, which will improve the fine-grained predictions. Moving beyond the vision-language alignments, the last term attempts to explore the intra-modal interactions.  $\text{CT}(\mathbf{P}_2, \mathbf{P}_3)$  optimizes the composition and its primitives to be semantically close in the embedding space, showing linguist coherence.

**Cycle Consistency Regularization.** Due to the independent operation in Eq. 11, pairwise CTs may lead to potential inconsistencies and inefficiencies, as they fail to fully capture the global relationships among all sets involved. Cycle consistency, which has been extensively explored in domains such as image matching [Bernard *et al.*, 2019], multi-graph alignment [Wang *et al.*, 2021], and 3D pose estimation [Dong *et al.*, 2019], offers a natural solution to this problem. By enforcing a closed-loop structure, cycle consistency can enhance the coherence of semantic relationships across all sets, ensuring more consistent and efficient alignment. Here, we denote the probability of the composition set being self-mapping back to its initial state as:

$$\mathbf{T}_{22} = \overleftarrow{\mathbf{T}}_{21} \cdot \vec{\mathbf{T}}_{13} \cdot \overleftarrow{\mathbf{T}}_{32}, \quad (12)$$

where  $\overleftarrow{\mathbf{T}}_{21}$  from  $\text{CT}(\mathbf{P}_1, \mathbf{P}_2)$  denotes the transport plan from  $\mathbf{P}_2$  to  $\mathbf{P}_1$ . This guarantees each label is transferred with a

probability of 1 during transportation among three sets. For a given input, the corresponding point from the composition set first interacts with key patches in the patch set, then moves to the relevant state and object points within the primitive set, and finally returns to the composition set. Accordingly, we derive the cycle-consistency constraint by:

$$\mathcal{L}_{cyc} = \sum_{m=1}^M \mathbf{y}_m^c (\mathbf{T}_{22} - \mathbf{I}), \quad (13)$$

where  $\mathbf{y}_m^c$  denotes the one-hot binary label vector of the input image.

For one thing, such a closed-loop transport constraint establishes a close link between independent pairwise CTs, which supports maintaining semantic consistency throughout the CT transport process. For another, it facilitates more robust composition representation learning, enhancing the model's accuracy.

**Primitive Decoupler.** In light of the intricate entanglement between attributes and objects within an image, we devise a decoupling loss with the idea that visual representation  $\mathbf{x}_{CLS}^s$  and  $\mathbf{x}_{CLS}^o$  can be seen as state-expert and object-expert if their paired sub-concepts fail to be inferred:

$$\mathcal{L}_{de} = \|\cos(\mathbf{x}_{CLS}^s, \mathbf{z}_{gt}^o)\| + \|\cos(\mathbf{x}_{CLS}^o, \mathbf{z}_{gt}^s)\|, \quad (14)$$

where  $\mathbf{z}_{gt}^o$  and  $\mathbf{z}_{gt}^s$  are textual representations for the ground-truth object class and attribute class, respectively. By mitigating the entanglement among visual representations,  $\text{TsCA}$  enhances its capacity to pinpoint correlative image representations that align with specific knowledge, thereby offering guidance on the textual sets prior and playing a complementary role in deriving a more accurate composition during inference.

### 3.4 Training and Inference

**Training Objectives.** Recalling that the probability weights  $\alpha$ ,  $\beta^s$ , and  $\beta^o$ , calculated during the construction of the three sets, capture the semantic similarity between the visual image and the corresponding textual features, this naturally defines the probability for predicting the labels of state  $s$ , object  $o$ , and composition  $c$  for the image  $x$ :

$$p(s|x) = \beta^s, \quad p(o|x) = \beta^o, \quad p(c|x) = \alpha \quad (15)$$

Then the classification losses are given by:

$$\begin{cases} \mathcal{L}_s = -\frac{1}{|\mathcal{X}|} \sum_{\hat{x} \in \mathcal{X}} \log p(s|\hat{x}) \\ \mathcal{L}_o = -\frac{1}{|\mathcal{X}|} \sum_{\hat{x} \in \mathcal{X}} \log p(o|\hat{x}) \\ \mathcal{L}_c = -\frac{1}{|\mathcal{X}|} \sum_{\hat{x} \in \mathcal{X}} \log p(c|\hat{x}) \end{cases} \quad (16)$$

Let  $\mathcal{L}_{base} = \mathcal{L}_c + \mathcal{L}_s + \mathcal{L}_o$ , the overall training loss is defined as follows:

$$\mathcal{L} = \lambda_0 \mathcal{L}_{base} + \lambda_1 \mathcal{CT} + \lambda_2 \mathcal{L}_{cyc} + \lambda_3 \mathcal{L}_{de}, \quad (17)$$

where  $\lambda$  are hyper-parameters to balance the losses. Like the previous works, the first term is our base classification loss. It will be used to predict the final label via a combined strategy during the inference. In addition, it helps to construct increasingly coherent textual sets as the loss decreases. The last three terms act as semantic regularization, which guides the learning process from various domain experts.

**Inference.** With multi-path union, the prediction results of states and objects can be incorporated to assist the composition branch. Formally, the integrated composition probability can be denoted as:

$$\tilde{p}(c|x) = \gamma p(c|x) + (1 - \gamma)(p(s|x) \times p(o|x)), \quad (18)$$

where  $\gamma$  balances the contributions of primitive prediction and composition prediction in multi-path learning.

Additionally, we apply a single CT computation to filter out infeasible compositions that might be present in the open-world setting. Concretely, we calculate the bidirectional transport plans between primitives and compositions as their feasible scores, and discarded less relevant pairs by a threshold  $T$  in :

$$\mathcal{C}^{test'} = \left\{ c = (s, o) : \overrightarrow{t}_{cs} \overrightarrow{t}_{co} + \overleftarrow{t}_{sc} \overleftarrow{t}_{oc} < T \right\}. \quad (19)$$

This strategy reduces the search space and increases performance simultaneously.

## 4 Experiment

### 4.1 Experimental Settings

**Datasets.** The proposed  $\text{TsCA}$  is evaluated on three real-world CZSL benchmark datasets: MIT-states [Isola *et al.*, 2015], UT-Zappos [Yu and Grauman, 2014] and C-GQA [Naeem *et al.*, 2021]. MIT-States comprises images of naturally occurring objects, with each object characterized by an accompanying adjective description. It comprises 53,753 images depicting 115 states and 245 objects. UT-Zappos is a fine-grained dataset consisting of different kinds of shoes with texture attributes, totaling 16 states and 12 objects. C-GQA, derived from the Stanford GQA dataset [Hudson and Manning, 2019], features 453 states and 870 objects with over 9,500 compositions, making it the most pairs dataset for CZSL. We follow the split suggested by the previous work [Purushwalkam *et al.*, 2019] to ensure fair comparisons.

**Metrics.** Following the common practice of prior works [Lu *et al.*, 2023], we utilize the standard evaluation protocols and assessed all results using four metrics in both closed-world and open-world scenarios. Concretely, **S** measures the best seen accuracy when calibration bias is  $+\infty$  and **U** denotes the accuracy specifically for unseen compositions when the bias is  $-\infty$ . To provide an overall performance measure on both seen and unseen pairs, we also report the area under the seen-unseen accuracy curve (**AUC**) by varying the calibration bias from  $+\infty$  to  $-\infty$  and identify the point that achieves the best harmonic mean (**H**) between the seen and unseen accuracy. Among these, **H** and **AUC** are the core metrics for comprehensively evaluating the model.

**Implementation Details.** The proposed  $\text{TsCA}$  and all baselines are implemented with a per-trained CLIP ViT-L/14 model in PyTorch [Paszke *et al.*, 2019]. For the UT-Zappos, the hyper-parameters  $\lambda_0, \lambda_1, \lambda_2, \lambda_3$  in losses are set as 1, 0.1, 10, and 0.1. For the MIT-States, the hyper-parameters are set as 1, 0.01, 0.1, and 0.01. For the C-GQA, the hyper-parameters are set as 1, 0.01, 0.3, and 0.01. During inference,  $\gamma$  is set to 0.8 for UT-Zappos and 0.4 for MIT-States and C-GQA in the closed-world scenario and is adjusted to 0.4, 0.3, and 0.2 for



Method	MIT-States				UT-Zappos				C-GQA			
	S	U	H	AUC	S	U	H	AUC	S	U	H	AUC
CLIP [Radford <i>et al.</i> , 2021]	30.2	46.0	26.1	11.0	15.8	49.1	15.6	5.0	7.5	25.0	8.6	1.4
CoOp [Zhou <i>et al.</i> , 2022]	34.4	47.6	29.8	13.5	52.1	49.3	34.6	18.8	20.5	26.8	17.1	4.4
PromptCompVL [Xu <i>et al.</i> , 2022]	48.5	47.2	35.3	18.3	64.4	64.0	46.1	32.2	—	—	—	—
CSP [Nayak <i>et al.</i> , 2023]	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 [Wang <i>et al.</i> , 2023]	47.5	50.6	37.3	20.2	63.0	68.8	48.2	35.0	30.8	28.4	22.4	7.2
GIPCOL [Xu <i>et al.</i> , 2024]	48.5	49.6	36.6	19.9	65.0	68.5	48.8	36.2	31.9	28.4	22.5	7.1
DFSP (i2t) [Lu <i>et al.</i> , 2023]	47.4	52.4	37.2	20.7	64.2	66.4	45.1	32.1	35.6	29.3	24.3	8.7
DFSP (BiF) [Lu <i>et al.</i> , 2023]	47.1	52.8	37.7	20.8	63.3	69.2	47.1	33.5	36.5	32.0	26.2	9.9
DFSP (t2i) [Lu <i>et al.</i> , 2023]	46.9	52.0	37.3	20.6	66.7	71.7	47.2	36.0	38.2	32.0	27.1	10.5
PLID [Bao <i>et al.</i> , 2023]	49.7	52.4	39.0	22.1	67.3	68.8	52.4	38.7	38.8	33.0	27.9	11.0
Troika [Huang <i>et al.</i> , 2024]	49.0	<b>53.0</b>	<b>39.3</b>	22.1	66.8	73.8	54.6	41.7	41.0	35.7	29.4	12.4
CDS-CZSL [Li <i>et al.</i> , 2024]	<b>50.3</b>	52.9	39.2	22.4	63.9	<b>74.8</b>	52.7	39.5	38.3	34.2	28.1	11.1
Retrieval-Augmented [Jing <i>et al.</i> , 2024]	50.0	<b>53.3</b>	39.2	<b>22.5</b>	<b>69.4</b>	72.8	<b>56.5</b>	<b>44.5</b>	<b>45.6</b>	<b>36.0</b>	<b>32.0</b>	<b>14.4</b>
TsCA	<b>51.2</b>	52.9	<b>39.9</b>	<b>23.0</b>	<b>68.7</b>	<b>75.8</b>	<b>58.5</b>	<b>46.1</b>	<b>43.8</b>	<b>38.9</b>	<b>33.1</b>	<b>15.2</b>

 Table 1: CZSL comparisons in the closed-world setting. The best results are in **bold**. The second best results are in **blue**.

Method	MIT-States				UT-Zappos				C-GQA			
	S	U	H	AUC	S	U	H	AUC	S	U	H	AUC
CLIP [Radford <i>et al.</i> , 2021]	30.1	14.3	12.8	3.0	15.7	20.6	11.2	2.2	7.5	4.6	4.0	0.3
CoOp [Zhou <i>et al.</i> , 2022]	34.6	9.3	12.3	2.8	52.1	31.5	28.9	13.2	21.0	4.6	5.5	0.7
PromptCompVL [Xu <i>et al.</i> , 2022]	48.5	16.0	17.7	6.1	64.6	44.0	37.1	21.6	—	—	—	—
CSP [Nayak <i>et al.</i> , 2023]	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.2
HPL [Wang <i>et al.</i> , 2023]	46.4	18.9	19.8	6.9	63.4	48.1	40.2	24.6	30.1	5.8	7.5	1.4
GIPCOL [Xu <i>et al.</i> , 2024]	48.5	16.0	17.9	6.3	65.0	45.0	40.1	23.5	31.6	5.5	7.3	1.3
DFSP (i2t) [Lu <i>et al.</i> , 2023]	47.2	18.2	19.1	6.7	64.3	53.8	41.2	26.4	35.6	6.5	9.0	2.0
DFSP (BiF) [Lu <i>et al.</i> , 2023]	47.1	18.1	19.2	6.7	63.5	57.2	42.7	27.6	36.4	7.6	10.6	2.4
DFSP (t2i) [Lu <i>et al.</i> , 2023]	47.5	18.5	19.3	6.8	66.8	60.0	44.0	30.3	38.3	7.2	10.4	2.4
PLID [Bao <i>et al.</i> , 2023]	49.1	18.7	20.0	7.3	67.6	55.5	46.6	30.8	39.1	7.5	10.6	2.5
Troika [Huang <i>et al.</i> , 2024]	48.8	18.7	20.1	7.2	66.4	61.2	47.8	33.0	40.8	7.9	10.9	2.7
CDS-CZSL [Li <i>et al.</i> , 2024]	49.4	<b>21.8</b>	<b>22.1</b>	8.5	64.7	61.3	48.2	32.3	37.6	8.2	11.6	2.7
Retrieval-Augmented [Jing <i>et al.</i> , 2024]	49.9	20.1	21.8	8.2	69.4	59.4	47.9	33.3	<b>45.5</b>	11.2	<b>14.6</b>	<b>4.4</b>
TsCA (w/o filter)	<b>50.7</b>	21.5	22.0	<b>8.6</b>	<b>69.7</b>	<b>63.3</b>	<b>52.0</b>	<b>37.0</b>	43.7	<b>11.3</b>	<b>14.6</b>	4.3
TsCA (w filter)	<b>50.8</b>	<b>21.7</b>	<b>22.3</b>	<b>8.7</b>	<b>69.8</b>	<b>63.4</b>	<b>52.2</b>	<b>37.1</b>	<b>44.3</b>	<b>11.4</b>	<b>14.7</b>	<b>4.5</b>

 Table 2: CZSL comparisons in the open-world setting. We use ‘w’ and ‘w/o’ to distinguish models adopting filtering strategy to filter unfeasible compositions. The best results are in **bold**. The second best results are in **blue**.

each of the datasets in the open-world setting. The primitive adapters consist of two individual single-layer MLPs. All experiments are performed on a NVIDIA RTX A6000 GPU.

## 4.2 Comparison with State-of-the-Arts

We compare our TsCA with the most recent CZSL methods. The results are shown in Tab. 1 and Tab. 2. On the closed-world setting, TsCA exceeds the previous SOTA methods on all datasets in core metrics. Specifically, it yields improvements of 0.6% on MIT-States, 2% on UT-Zappos, and 1.1% on C-GQA in **H** over the second-best methods. It also attains considerable gains in **AUC** with increases of 0.5%, 1.6%, and 0.8%, respectively. Notably, the advantage of UT-Zappos is more significant, suggesting that the fine-grained nature of UT-Zappos can better align with TsCA’s ability to achieve precise alignment across local visual features and textual representations. Similarly, the results in the open-world setting are also promising. Our TsCA attains the highest **AUC** across all datasets, with the scores of 8.7%, 37.1% and 4.5%. It should be noted that the CT-based filtering strategy contributes to the improvements in this challenging setting. All numerical results substantiate our motivation to empower the model to capture semantic consistency among the image-composition-

Components	closed-world				open-world			
	S	U	H	AUC	S	U	H	AUC
$\mathcal{L}_{base}$	65.7	73.5	54.4	40.8	64.3	62.5	48.0	33.3
+ $\mathcal{L}_{de}$	68.2	73.9	56.8	44.3	66.9	62.2	49.2	34.6
+ $\mathcal{CT}$	67.8	74.3	57.5	44.6	67.0	64.6	50.1	35.9
+ $\mathcal{CT} + \mathcal{L}_{cyc}$	68.9	74.5	58.0	45.0	67.6	63.6	51.9	36.2
+ $\mathcal{L}_{de} + \mathcal{CT}$	69.0	75.4	58.2	45.3	67.9	63.7	52.0	36.3
+ $\mathcal{L}_{de} + \mathcal{CT} + \mathcal{L}_{cyc}$	68.7	75.8	58.5	46.1	69.8	63.4	52.2	37.1

Table 3: Ablation results for each component on UT-Zappos.

primitive interactions.

## 4.3 Ablation Study

We empirically verify the effectiveness of each component in TsCA on UT-Zappos by comparing it against five variants. The results, seen in Tab. 3, illustrate several observations: 1) The baseline model, i.e., removes the consistency-aware CT module and decoupler loss, shows the lowest performance. 2) Introducing either pairwise CT loss or decoupler loss on top of the baseline model significantly boosts all metrics. Notably, both primitive decoupler and cycle-consistency constraint positively contribute to pairwise CT. The former improves the







Success Cases				Failure Cases			
	small bathroom tiny bathroom wide bathroom small bathroom S → C	small bathroom wide bathroom tiny bathroom clean room O → C	small bathroom tiny shower tiny shower tiny shower C → S		cracked wall cracked concrete cracked paint cracked wall S → C	cracked wall cracked concrete cracked paint concrete O → C	cracked wall broken concrete dented room concrete C → S
	wrinkled velvet crushed velvet crumpled velvet wrinkled velvet S → C	wrinkled velvet crushed velvet crumpled velvet crinkled silk O → C	wrinkled velvet crushed fabric crinkled silk crinkled silk C → S		accient building accient library accient house accient house S → C	accient building accient house old city grimy column O → C	accient building concrete old city column C → S
	empty field verdant field sunny sky small tree S → C	empty field verdant field sunny sky small tree O → C	empty field verdant field sunny sky small tree C → S		inflated toy inflated balloon filled balloon inflated toy S → C	inflated toy inflated balloon filled balloon filled balloon O → C	inflated toy inflated balloon filled balloon grimy rubber C → S

Figure 3: Qualitative results of the intra-modal transport plans on the MIT-States. For each sample, we show an image with the ground-truth composition, with the state indicated in red and the object in blue. The top-3 predictions are presented in two formats: from the primitive class to the composition set in the first two columns, and from the composition label to the primitive set in the third and fourth columns. The annotations ‘s’, ‘o’, and ‘c’ correspond to state, object, and composition, respectively.

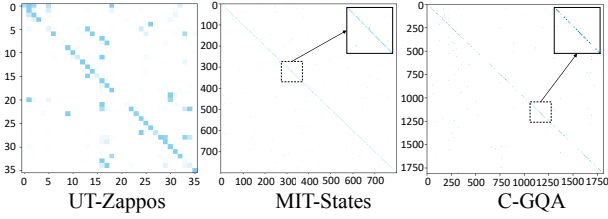


Figure 4: Cycle-consistency (zoom-in for more details).

quality of the textual sets, while the latter ensures semantic consistency during the transport chain, each leading to further enhancements in the model’s effectiveness. 3) Combining the strengths of all components, our complete model achieves the best performance in terms of **H** and **AUC**.

To demonstrate the effectiveness of cycle-consistency, we draw the self-mapping matrix of the composition set for our full model on test data in Fig. 4. Though degradation occurs in the unseen pair positions, the results still exhibit a close similarity to the identity matrix.

#### 4.4 Qualitative Results

We further provide some visualization examples of transport plans learned in our  $T_{SCA}$ . Recalling that the columns  $\vec{T}_{21}$  and  $\vec{T}_{31}$  depict how likely the corresponding textual semantics are transported to each visual patch. We convert each transport plan into heatmaps and resize them to combine with the raw image at Fig. 5. We observe that different prompts from multi-path tend to align different patch regions, each of which contributes to the final prediction. For example, in the first row, the state branch emphasizes the wet fur, the object branch highlights the nose and mouth as key features for recognizing a cat, while the composition branch provides a more comprehensive focus on both aspects.

Moving beyond the visualization of cross-modal alignment,  $\vec{T}_{32}$  and  $\vec{T}_{23}$  also grant us access to intra-modal components, prompting an exploration of the interplay between compositions and primitives in an interpretable form. Fig. 3 presents the top-3 transport retrieval results for bidirectional transfers between the composition and primitive sets. Note that our model can both retrieve the correct composition from the primitive set, and also adeptly performs the inverse mapping of conceptual pairs. Besides, all top-3 retrieval results not only

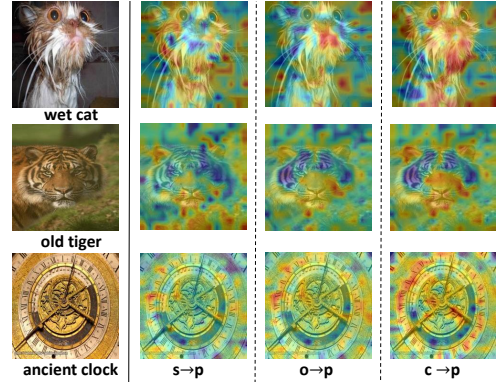


Figure 5: Visualization of the cross-modal transport plans on the Mit-States. Columns 1-3 represent the transport from the ground-truth state and object points in the primitive set, as well as the composition points in the composition set, to the patch set. The annotations ‘p’ correspond to patch.

ensure the rationality of the state-object combination but also conform to the description of the images. We also present some failure cases. Interestingly, even when incorrect, the retrieved labels still capture the content of the given images, demonstrating the effectiveness of our  $T_{SCA}$ .

## 5 Conclusion

In this paper, we revisit compositional generalization in CZSL through a conditional transport perspective. We explore pairwise CTs between the local visual features of images and two textual label sets within a multi-path paradigm. We then introduce cycle-consistency as a link to bond all sets, promoting robust learning. The primitive decoupler further improved accuracy and decoupled global primitive visual representations during prediction. Benefited from transport plan between primitive set and composition set, this approach successfully narrowed the composition search space by excluding unfeasible pairs during inference. Extensive experiments across three benchmarks consistently demonstrate the superiority of  $T_{SCA}$ . Comprehensive ablation studies and visualizations confirm our motivation and the essential role of each component. With its inherent flexibility and simplicity, we hope our work inspires innovative ideas for future research.

## Acknowledgments

This research was supported by funding from the Hong Kong RGC General Research Fund (No. 152211/23E, 15216424/24E, and 152115/25E), the PolyU Internal Fund (No. P0056171), and the Huawei Gifted Fund.

## References

- [Bao *et al.*, 2023] Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. Prompting language-informed distribution for compositional zero-shot learning. *arXiv preprint arXiv:2305.14428*, 2023.
- [Bernard *et al.*, 2019] Florian Bernard, Johan Thunberg, Paul Swoboda, and Christian Theobalt. Hippo: Higher-order projected power iterations for scalable multi-matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10284–10293, 2019.
- [Dong *et al.*, 2019] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7792–7801, 2019.
- [Huang *et al.*, 2024] Siteng Huang, Biao Gong, Yutong Feng, Min Zhang, Yiliang Lv, and Donglin Wang. Troika: Multi-path cross-modal traction for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24005–24014, 2024.
- [Hudson and Manning, 2019] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.
- [Huynh and Elhamifar, 2020] Dat Huynh and Ehsan Elhamifar. Compositional zero-shot learning via fine-grained dense feature composition. *Advances in Neural Information Processing Systems*, 33:19849–19860, 2020.
- [Isola *et al.*, 2015] 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.
- [Jing *et al.*, 2024] Chenchen Jing, Yukun Li, Hao Chen, and Chunhua Shen. Retrieval-augmented primitive representations for compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2652–2660, 2024.
- [Karthik *et al.*, 2022] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9326–9335, 2022.
- [Li *et al.*, 2020] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. 2020 IEEE. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11313–11322, 2020.
- [Li *et al.*, 2022] 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.
- [Li *et al.*, 2023a] Lin Li, Guikun Chen, Jun Xiao, and Long Chen. Compositional zero-shot learning via progressive language-based observations. *arXiv preprint arXiv:2311.14749*, 2023.
- [Li *et al.*, 2023b] Miaoge Li, Dongsheng Wang, Xinyang Liu, Zequn Zeng, Ruiying Lu, Bo Chen, and Mingyuan Zhou. Patchct: Aligning patch set and label set with conditional transport for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15348–15358, 2023.
- [Li *et al.*, 2024] Yun Li, Zhe Liu, Hang Chen, and Lina Yao. Context-based and diversity-driven specificity in compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17037–17046, 2024.
- [Liu *et al.*, 2023] 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*, 2023.
- [Lu *et al.*, 2023] 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.
- [Mancini *et al.*, 2021] 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.
- [Mancini *et al.*, 2022] 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*, 46(3):1545–1560, 2022.
- [Misra *et al.*, 2017] 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.
- [Naeem *et al.*, 2021] 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.
- [Nagarajan and Grauman, 2018] 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.



- [Nayak *et al.*, 2023] Nihal V. Nayak, Peilin Yu, and Stephen H. Bach. Learning to compose soft prompts for compositional zero-shot learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- [Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
- [Purushwalkam *et al.*, 2019] 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.
- [Radford *et al.*, 2021] 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.
- [Saini *et al.*, 2022] 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.
- [Tanwisuth *et al.*, 2021] Korawat Tanwisuth, Xinjie Fan, Huangjie Zheng, Shujian Zhang, Hao Zhang, Bo Chen, and Mingyuan Zhou. A prototype-oriented framework for unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 34:17194–17208, 2021.
- [Tanwisuth *et al.*, 2023] Korawat Tanwisuth, Shujian Zhang, Huangjie Zheng, Pengcheng He, and Mingyuan Zhou. Pouf: Prompt-oriented unsupervised fine-tuning for large pre-trained models. In *International Conference on Machine Learning*, pages 33816–33832. PMLR, 2023.
- [Tian *et al.*, 2023] Long Tian, Jingyi Feng, Xiaoqiang Chai, Wenchao Chen, Liming Wang, Xiyang Liu, and Bo Chen. Prototypes-oriented transductive few-shot learning with conditional transport. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16317–16326, 2023.
- [Vaswani, 2017] Ashish Vaswani. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
- [Wang *et al.*, 2021] Runzhong Wang, Junchi Yan, and Xiaokang Yang. Neural graph matching network: Learning lawler’s quadratic assignment problem with extension to hypergraph and multiple-graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5261–5279, 2021.
- [Wang *et al.*, 2022] Dongsheng Wang, Dandan Guo, He Zhao, Huangjie Zheng, Korawat Tanwisuth, Bo Chen, and Mingyuan Zhou. Representing mixtures of word embeddings with mixtures of topic embeddings. *arXiv preprint arXiv:2203.01570*, 2022.
- [Wang *et al.*, 2023] Henan Wang, Muli Yang, Kun Wei, and Cheng Deng. Hierarchical prompt learning for compositional zero-shot recognition. In *IJCAI*, volume 1, page 3, 2023.
- [Xian *et al.*, 2017] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:2251–2265, 2017.
- [Xu *et al.*, 2021] Guangyue Xu, Parisa Kordjamshidi, and Joyce Y Chai. Zero-shot compositional concept learning. *arXiv preprint arXiv:2107.05176*, 2021.
- [Xu *et al.*, 2024] Guangyue Xu, Joyce Chai, and Parisa Kordjamshidi. Gipcol: Graph-injected soft prompting for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5774–5783, 2024.
- [Xu *et al.*, 2022] Guangyue. Xu et al. Prompting large pre-trained vision-language models for compositional concept learning. *arXiv preprint arXiv:2211.05077*, 2022.
- [Yu and Grauman, 2014] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 192–199, 2014.
- [Zheng and Zhou, 2021] Huangjie Zheng and Mingyuan Zhou. Exploiting chain rule and bayes’ theorem to compare probability distributions. *Advances in Neural Information Processing Systems*, 34:14993–15006, 2021.
- [Zhou *et al.*, 2022] 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.