Decomposed Soft Prompt Guided Fusion Enhancing for Compositional Zero-Shot Learning

Xiaocheng Lu[1], Song Guo[*,1,2], Ziming Liu[1], Jingcai Guo[*,1,2]
[1]Department of Computing, The Hong Kong Polytechnic University
[2]The Hong Kong Polytechnic University Shenzhen Research Institute
{xiaoclu, song.guo}@polyu.edu.hk, {ziming.liu, jingcai.guo}@connect.polyu.hk

Abstract

*Compositional Zero-Shot Learning (CZSL) aims to recognize novel concepts formed by known states and objects during training. Existing methods either learn the combined state-object representation, challenging the generalization of unseen compositions, or design two classifiers to identify state and object separately from image features, ignoring the intrinsic relationship between them. To jointly eliminate the above issues and construct a more robust CZSL system, we propose a novel framework termed **D**ecomposed **F**usion with **S**oft **P**rompt (DFSP)*[1]*, by involving vision-language models (VLMs) for unseen composition recognition. Specifically, DFSP constructs a vector combination of learnable soft prompts with state and object to establish the joint representation of them. In addition, a cross-modal decomposed fusion module is designed between the language and image branches, which decomposes state and object among language features instead of image features. Notably, being fused with the decomposed features, the image features can be more expressive for learning the relationship with states and objects, respectively, to improve the response of unseen compositions in the pair space, hence narrowing the domain gap between seen and unseen sets. Experimental results on three challenging benchmarks demonstrate that our approach significantly outperforms other state-of-the-art methods by large margins.*

1. Introduction

Given an unseen concept, such as *green tiger*, even though this is a nonexistent stuff humans have never seen, they may associate the known state *green* with an image of *tiger* immediately. Inspired by this, Compositional Zero-Shot Learning (CZSL) is proposed with the purpose of equipping models with the ability to recognize novel concepts generated as humans do. Specifically, CZSL learns on visible primitive composed concepts (state and object) in the training phase, and recognizes unseen compositions in the inference phase.

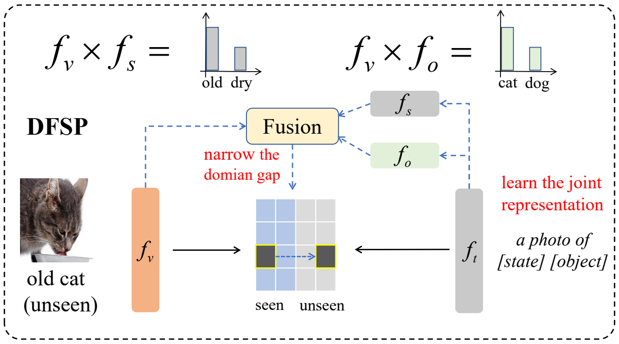

Figure 1. The overview of DFSP. Our method aims to narrow the domain gap between seen and unseen compositions by fusing decomposed features f_o and f_s with image feature f_v, while learn the joint representation between state and object in language branch. Being fused with the state and object features, image feature can learn the response of them respectively and improve the sensitiveness of unseen compositions.

Some prior algorithms [20, 26] design two classifiers to identify state and object separately, while these models overlook the intrinsic relation between them. After the primitive concepts are obtained, the association between state and object could be established again through graph neural network (GNN) [24] or external knowledge compositions [14]. Nevertheless, these are post-processing methods and these classifiers are separated from image features with strong correlation, ignoring entanglement. Some other methods [28, 29] are to directly treat the combination as an entity, converting CZSL into a general zero-shot recognition problem. Generally, the visual features are projected into a shared semantic space and the distance between entities is optimized, such as Euclidean distance [44]. If too much

*Song Guo and Jingcai Guo are the corresponding authors

[1] Code is available at: https://github.com/Forest-art/DFSP.git

attention is paid to the composed concepts in the training stage, the model can not be generalized well to unseen compositions, causing the domain gap between seen and unseen sets. In summary, these methods are all visual recognition models, which are limited by the strong entanglement of states and objects in image features.

In contrast, we focus on designing novel approaches based on vision-language models (VLMs) to cope with CZSL challenges. Since state and object are two separate words in the text, they are less entangled in language features than image features and could be decomposed more easily and precisely. Certainly, state and object are also intrinsically linked in the text, such as *ripe apple* instead of *old apple*. Constructing the combination in the form of text can also establish the joint representation of state and object to pair with images. Meanwhile, the decomposed state and object features can also be independently associated with the image feature, easing the excessive bias of the model towards seen compositions and enhancing the unseen response (shown in Fig. 1). To improve CZSL with VLMs, we design **D**ecomposed **F**usion with **S**oft **P**rompt (DFSP), an efficient framework aimed to both learn about the joint representation of primitive concepts and shrink the domain gap between seen and unseen composition sets, as shown in Fig. 2. To be specific, DFSP is designed as a fully learnable soft prompt including prefix, state and object, which constructs the joint representation between primitive concepts and can be fine-tuned well for new supervised tasks. We then design a decomposed fusion module (DFM) for state and object, which decomposes features extracted from text encoder, such as Bert [6], etc. Meanwhile, the decomposed language features and image features of DFSP interact with information in a cross-modal fusion module, which is crucial for learning high-quality language-aware visual representations. During the phase of fusion, the image can establish separate relationships with the state and object, and then is paired with the composed prompt feature in the pair space, improving its response even for unseen compositions to shrink the domain gap.

Generally, this paper makes the following contributions:

- A novel framework named Decomposed Fusion with Soft Prompt (DFSP) is proposed, which is based on vision-language paradigm aiming to cope with CZSL.
- The Decomposed Fusion Module is designed for CZSL specifically, which decomposes the concepts of language features and fuses them with image features to improve the response of unseen compositions.
- We design a learnable soft prompt to construct the joint-representation of state and object, which can be more precisely decomposed than images.
- Extensive experiments demonstrate the effectiveness of DFSP, which greatly outperforms the state-of-the-art CZSL approaches on both closed-world and open-world.

2. Related Work

We describe the compositional zero-shot learning and prompt learning in this section.

Compositional Zero-Shot Learning. CZSL [9, 20, 25, 26, 28] is a task similar to how humans can imagine and discriminate unseen concepts according to the concepts they have learned, which is a significant branch of ZSL [5, 10, 11, 15, 17, 21, 22, 40].

For CZSL, early works learn a classifier for recognition and a transformation module to convert state or object [26, 28]. Some recent works utilize two separate classifiers to recognize state and object respectively [14, 18, 20, 26]. Also, some works combine the encoded attribute/state and object features with late fusion by using multi-layer perceptron [32]. Li *et al.* introduce contrastive learning into CZSL, and design a siamese network to identify state and object in the contrastive space, respectively [18]. Other methods [28, 29] focus on the joint representation of the compositions, which learn an emdedding space to map the compositions like ZSL. Recent works utilize graph networks to represent the state and object relationship and then learn their compositions [24, 35, 42]. Besides, Nihal *et al.* first attempt to use a VLM model for CZSL, replacing the classes in prompt with a learnable combined state and object vector representation [30].

As opposed to the previous closed-world, some work aim at open-world by using external knowledge to filter infeasible compositions [14, 23, 24].

Prompt Learning. Prompt Learning refers to processing the input text through a specific template, and reconstructing the task into a form that can more fully utilize the pre-trained language model [1–3, 36, 39, 46]. Prompting makes the pre-training model and downstream tasks closer, which is different from fine-tuning. Benefiting from having pre-trained on a large-scale data and associating with multi-modal information, prompt learning can achieve great performance in zero-shot and few-shot on a wide range of tasks [33, 34].

Take the CLIP [34] model as an example, discrete prompt has difficulty performing well on downstream tasks even when trained on new data. Some recent works utilize soft prompt to improve downstream tasks and reach fine performance [16, 19, 37]. CoOp [46] convert the prefix part of the prompt to soft prompt like *[v1][v2][v3]object*, fix the parameters of other parts, and only fine-tune the prompt. In contrast, CSP [30] sets the primitive concepts section of the prompt to be soft like *a photo of [state][object]*. While

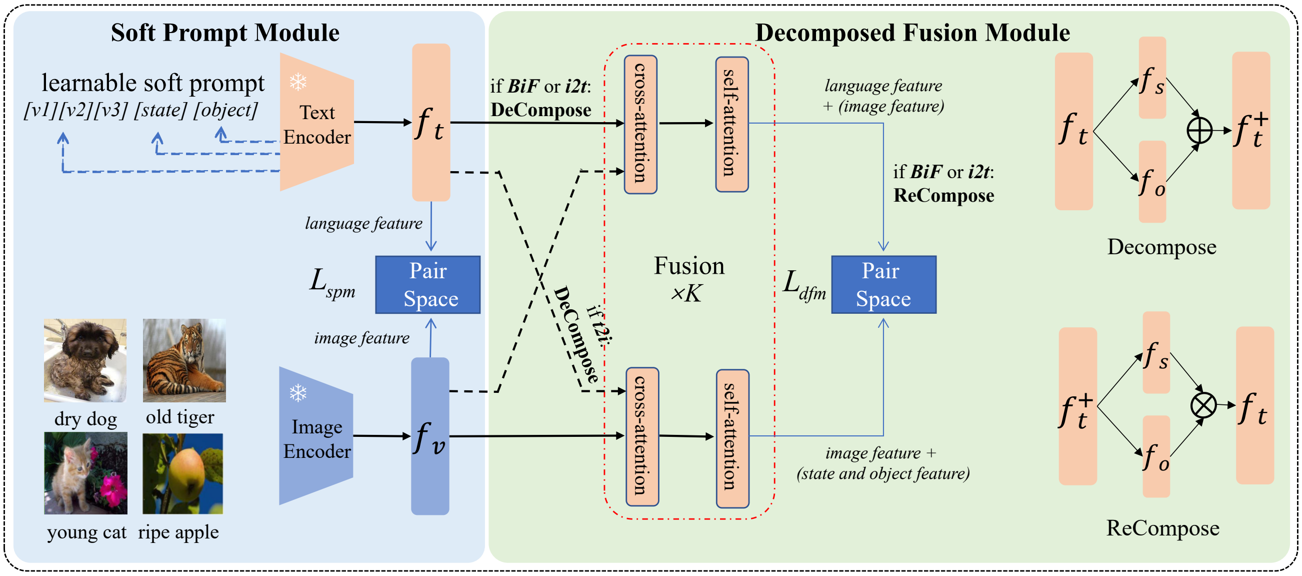

Figure 2. The framework of our proposed DFSP, which consists of Soft Prompt Module (SPM) and Decomposed Fusion Module (DFM). Since DFSP is vision-language model, it could also be divided into two branches, language segment and image segment. SPM aims to construct and preserve the joint representation of state and object, then convert the discrete prompt to learnable soft prompt, causing the extracted language features f_t more discriminative and more suitable for new tasks, especially CZSL. There are three forms of decomposed fusion in SPM, *BiF*, *i2t* and *t2i* respectively, and if the fusion method is *t2i*, only decomposition exists on the language to image branch. Meanwhile, decomposition and recomposition coexist in the fusion method of *BiF* and *i2t*. After decomposing language feature into independent state feature f_s and object feature f_o, DFM fuses them with image feature f_v and calculates similarity in the final pair space. DFSP can not only learn the joint representation of state and object, but also shrink the domain gap of seen and unseen composition sets.

these methods focus only on a certain part of the prompt, we soften the entire prompt to better fine-tune in the new scenario.

3. Approach

For CZSL, entities exist in the form of a combination of state and object, and the model needs to be trained on the seen composition set while tested on the unseen set. To address this challenge, we propose a novel formulation termed Decomposed Fusion with Soft Prompt (DFSP), which constructs a vision-language paradigm with soft prompt and decomposed fusion module. DFSP first construct soft prompt with state and object to establish the joint representation of them. Meanwhile, DFSP decomposes language features to separated state and object features, and utilizes cross-modal fusion to transfer knowledge between decomposed language and images. The framework of our proposed method is shown in Fig. 2.

3.1. Problem Formulation

Given state set $\mathcal{A} = \{s_0, s_1, \ldots, s_n\}$ and object set $\mathcal{O} = \{o_0, o_1, \ldots, o_m\}$ as the primitive concepts of CZSL, we can compose them as a composition set $\mathcal{C} = \mathcal{A} \times \mathcal{O}$, where the size of $\mathcal{C}$ is $n \times m$. Besides, we denote two disjoint sets $\mathcal{C}^s$ and $\mathcal{C}^u$, where $\mathcal{C}^s$, $\mathcal{C}^u$ are subsets of the composition set $\mathcal{C}$ and $\mathcal{C}^s \cap \mathcal{C}^u = \phi$. Specifically, $\mathcal{C}^s$, $\mathcal{C}^u$ represent the seen and unseen sets, respectively, where $\mathcal{C}^s$ is used for training and $\mathcal{C}^u$ is used for testing. $\mathcal{T} = \{(x_i, c_i | x \subset \mathcal{X}, c \subset \mathcal{C}^s)\}$ is the training set where $\mathcal{X}$ is the input image space and c belongs to the seen composition label set.

The CZSL task aims to train a model $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{C}^t$ to predict compositions in the test samples space $\mathcal{C}^t$. If $\mathcal{C}^t \cap \mathcal{C}^s \equiv \phi$, where the model only predicts unseen compositions. Follow the setting of Generalized ZSL [43], testing samples contain seen and unseen compositions, i.e., $\mathcal{C}^s \cup \mathcal{C}^u$ in this paper. Generally, when testing, only the known composition space of test samples is required, which is called closed-world. For Open-World [14], the composition space for testing is all possible combinations, i.e., $\mathcal{C}^t = \mathcal{C}$.

3.2. Decomposed Fusion with Soft Prompt Network

Given an image such as *dry dog*, *dry* and *dog* have strong joint representation in image features, which is entanglement of state and object. Directly identifying the state and object of an image feature separately will lose its joint representation. For a short sentence of *dry dog*, although its embedding is also a combination of *dry* and *dog*, this combination is not strong entangled and could be precisely decomposed. Inspired by this, we propose a vision-language paradigm for CZSL, called Decomposed Fusion with Soft

Prompt (DFSP), which includes two modules, Soft Prompt Module (SPM) and Decomposed Fusion Module (DFM). SPM is responsible for the construction of joint representation between state and object, and DFM is to decompose and fuse language features with image features to improve the sensitiveness of unseen compositions. The overall architecture of DFSP is shown in Fig. 2.

Soft Prompt Module. DFSP is a vision-language model, in which encoders utilize Contrastive Language-Image Pre-Training (CLIP) [34], which has pretrained on nearly 400M text-image pairs. The feature extracted section of DFSP consists of the image encoder and the text encoder. For the image encoder, it can be a vision transformer (ViT) [7] or a convolutional neural network [12], while the text encoder includes several transformer encoder layers. It is worth mentioning that the parameters of image encoder and text encoder are frozen, because the CLIP model has done enough pre-training, we only need to finetune for downstream tasks.

To be specific, the entities consist of state and object in CZSL are transformed into natural language prompts like *a photo of [state][object]*, compared with *a photo of [class]* in CLIP. Before extracting the text representations, the prompts need to be converted to tokens for each word by tokenizer and the embedding function maps the tokens to the vocabulary. Due to the discrete vocabulary *[state][object]*, the model cannot be adapted to CZSL well. CSP [30] compares the CLIP model and soft *[state][object]* in prompt, which achieves some progress. Nevertheless, while CSP works well for CZSL, prefix is still fixed, which is too dependent on state and object learning for the model. And a prefix like *a photo of* is not necessarily the best prompt form; if it can also be updated, the model can be generalized better.

In DFSP, the prompt is fully learnable soft prompt *[v1][v2][v3][state][object]*. First, we build a prompt set with a prefix context, state and object, which is formulated as follows:

$$P(s,o) = \{x_0, x_1, \dots, x_p, x_s, x_o\}\,, \tag{1}$$

where $\{x_0, \dots, x_p\}$ is the prefix context and the x_s and x_o represents the state and object vocabulary for the composition set $P(s,o)$. Then, the prompt will be converted to learnable embeddings as follows:

$$P^{soft} = \Gamma(P(s,o)) = \{\theta_0, \theta_1, \dots, \theta_p, \theta_s, \theta_o\}\,, \tag{2}$$

where Γ is the embedding function, $\{\theta_0, \dots, \theta_p\}$ is the learnable prefix context and θ_s and θ_o denotes the learnable state and object embeddings. For the training samples $x \in \mathcal{X}$ and the soft prompt P^{soft}, we can extract the image and language features:

$$f_v = \frac{E_v(x)}{\|E_v(x)\|}, f_t = \frac{E_t(P^{soft})}{\|E_t(P^{soft})\|}, \tag{3}$$

where E_v and E_t represent the image encoder and text encoder, and $\|\cdot\|$ means the norm calculation. Next, we can compute the class probability $p_{spm}(\frac{y=(s,o)}{x:\theta})$ as follows:

$$p_{spm}(\frac{y=(s,o)}{x;\theta}) = \frac{exp(f_v \cdot f_t)}{\sum_{(\bar{s},\bar{o})\in\mathcal{C}^s} exp(f_v \cdot f_t)}. \tag{4}$$

Finally, we can minimize the cross entropy loss in the soft prompt module:

$$\mathcal{L}_{spm} = -\frac{1}{|\mathcal{C}^s|} \sum_{(x,y)\in\mathcal{C}^s} log\left(p_{spm}(\frac{y=(s,o)}{x;\theta})\right). \tag{5}$$

Decomposed Fusion Module. The language and image representations of SPM are directly calculated their similarity in the pair space, causing the model tend to the seen compositions of the training set, lacking the ability to perceive unseen samples. To narrow the domain gap between seen and unseen sets, we propose Decomposed Fusion Module (DFM), which decomposes language features into state and object features and fuses them with image feature.

The design of DFM has three key points for DFSP: i) DFM only decomposes the language features, in which state and object has less entanglement. Since state and object are separated in text, they can be decomposed easily and precisely, which preserves the joint representation of state and object; ii) The decomposed state and object features are fused with image features to realize information interaction between the two modalities. Meanwhile, DFM establishes respective associations of the image with the state and object, improving the responsiveness in the pair space, especially for the unseen compositions; iii) If there are many categories of states and objects, the composition set will be particularly large, limiting the performance of the model. DFM can reduce the complexity of the composition, from $O(n\times m)$ to $O(n+m)$. Since training is on the seen set and testing is on the unseen set, the number of compositions is inconsistent. For cross-modal fusion, decomposition is essential to ensure that the model can maintain a fixed amount of parameters when the composition set changes.

With SPM, we can extract the image feature f_v and language feature f_t, and then the state feature f_s and object feature f_o can be decomposed as follows:

$$f_s, f_o = \mathcal{D}_e(f_t), \tag{6}$$

where $De(\cdot)$ denotes the decomposition, and it's formulation is:

$$\mathcal{D}_e = \left\{ \sum_i \frac{f_t(i,j)}{|j|}, \sum_j \frac{f_t(i,j)}{|i|} | i \in \mathcal{A}, j \in \mathcal{O}, (i,j) \in \mathcal{C}^s \right\}. \tag{7}$$

The language feature f_t is the combined seen feature set, and $\mathcal{D}_e$ calculates its average state feature relative to each object and the average object feature of each state. The class scale of them is n and m respectively, and then f_s, f_o will be concatenated to f_t^+, which is consistent during training and testing.

The decomposed state and object features can also be supervised to provide some guidance for subsequent training. We can compute the state probability $p(\frac{y=s}{x;\theta})$ and object probability $p(\frac{y=o}{x;\theta})$ as follows:

$$p(\frac{y=s}{x;\theta}) = \frac{exp(f_v \cdot f_s)}{\sum_{(\bar{s})\in\mathcal{A}} exp(f_v \cdot f_t)}, \tag{8}$$

$$p(\frac{y=o}{x;\theta}) = \frac{exp(f_v \cdot f_o)}{\sum_{(\bar{o})\in\mathcal{O}} exp(f_v \cdot f_o)}. \tag{9}$$

And the cross entropy loss can be minimized by:

$$\begin{aligned}\mathcal{L}_{st+obj} = &-\frac{1}{|\mathcal{A}|}\sum_{(x,y)\in\mathcal{C}^s} log\left(p(\frac{y=s}{x;\theta})\right)\\ &-\frac{1}{|\mathcal{O}|}\sum_{(x,y)\in\mathcal{C}^s} log\left(p(\frac{y=o}{x;\theta})\right)\end{aligned} \tag{10}$$

Due to the mismatching of feature dimension, decomposed language features and image features need to be converted to consistent, the formulation is as follows:

$$f^+_{t\to i} = \mathcal{T}_{txt2img}(f_t^+), f_{i\to t} = \mathcal{T}_{img2txt}(f_v), \tag{11}$$

where $\mathcal{T}_{txt2img}$ is the transformation from language feature to image feature and $\mathcal{T}_{img2txt}$ is opposite.

The key to the image-text matching task is how to accurately calculate the visual-semantic similarity between images and texts. However, most of the existing algorithms only focus on the association between elements within a single modality, and do not combine the image and text features. Since the image and text encoders are all based on transformer layers, we utilize cross-attention and self-attention mechanism to fuse the decomposed feature with image feature [4, 38, 41]. Let S_1 and S_2 be the two modalities to be fused, and the fusion from S_1 to S_2 can be described as follows:

$$\mathcal{F}(S_1 \to S_2) = softmax((W_Q S_2)(W_K S_1)^T) W_V S_1, \tag{12}$$

where W_Q, W_K, W_V are trainable parameters and denote the query, key and value similar to Multi-Head Self-Attention [38]. So the specific fusion of cross-attention is as follows:

$$\begin{aligned} f_v^{fused} &= \mathcal{F}(f^+_{t\to i} \to f_v),\\ f_t^{fused} &= \mathcal{F}(f_{i\to t} \to f_t^+).\end{aligned} \tag{13}$$

To establish a deeper relationship among the fused features, they can to be fine-tuned by self-attention and the formulation is as follows:

$$\begin{aligned} f_v^{fused} &= \mathcal{F}(f_v^{fused} \to f_v^{fused}),\\ f_t^{fused} &= \mathcal{F}(f_t^{fused} \to f_t^{fused}).\end{aligned} \tag{14}$$

Besides, the fusion module consists of cross-attention and self-attention can be repeated K times to better adapted to complex tasks. Since the language features are decomposed, they can be recomposed before comparing with image features, which is a reverse phase compared to the decomposition. The formula is as follows:

$$f_s, f_o = S(f_t^{fused}), f_t = \{MLP(f_s \cdot f_o) | (s,o) \in \mathcal{C}^s\}, \tag{15}$$

where $S(\cdot)$ is the split function and $MLP(\cdot)$ is the multilayer perception to fine-tune. Finally, the DFM class probability $p_{dfm}(\frac{y=(s,o)}{x;\theta})$ as follows:

$$p_{dfm}(\frac{y=(s,o)}{x;\theta}) = \frac{exp(f_v \cdot f_t)}{\sum_{(\bar{s},\bar{o})\in\mathcal{C}^s} exp(f_v \cdot f_t)}. \tag{16}$$

And the cross entropy loss in DFM can be minimized:

$$\mathcal{L}_{dfm} = -\frac{1}{|\mathcal{C}^s|}\sum_{(x,y)\in\mathcal{C}^s} log\left(p_{dfm}(\frac{y=(s,o)}{x;\theta})\right). \tag{17}$$

The overall loss of the framework DFSP can be summarized as follows:

$$\mathcal{L} = \mathcal{L}_{dfm} + \alpha\mathcal{L}_{st+obj} + \beta\mathcal{L}_{spm}, \tag{18}$$

in which α and β are the weighting coefficients to balance the influence of each loss.

DFSP can be divided into three categories according to the fusion methods, *BiF*, *i2t* and *t2i*. While *i2t* and *t2i* respectively mean the fusion of image feature into text feature and text feature into image feature, while *BiF* means the fusion in both directions. Fusion of image feature with text feature requires decomposition of the text feature, and to maintain joint representation in the pair space, the fused features need to be recomposed, which is reverse of decomposition.

3.3. Inference

We utilize the final fused probability to infer on test set, and the test set includes seen and unseen compositions, which can be denoted as $\mathcal{C}^s \cup \mathcal{C}^u$. For both closed-world and open-world settings in testing phase, the most likely predicted result can by calculated as follows:

$$\hat{y} = argmax(p_{dfm}(\frac{y=(s,o)}{x:\theta})), y \in \mathcal{C}^s \cup \mathcal{C}^u. \tag{19}$$

To filter out infeasible compositions in the open-world setting, we follow the post-training calibration method [24, 30]. First, we calculate the similarities between objects:

$$q_o(s,o) = max\frac{\phi(o)\cdot\phi(\hat{o})}{\|\phi(o)\|\,\|\phi(\hat{o})\|}, (o,\hat{o}) \in \mathcal{O}, \tag{20}$$

where $q_o(s, o)$ denotes the similarity between object o and $\hat{o}$, and $\phi(\cdot)$ is an embedding function. Also, the similarities of states can be obtained in the same way. Next, the feasibility score can be calculated by mean pooling μ:

$$q(s, o) = \mu(q_s(s, o), q_o(s, o)). \tag{21}$$

Finally, the infeasible compositions can be filter out by a threshold T:

$$\hat{y} = argmax(p(\frac{y = (s, o)}{x : \theta})), y \in \mathcal{C}^s \cup \mathcal{C}^u, q_(s, o) > T. \tag{22}$$

4. Experiment

In this section, we describe all datasets and our experiments. And the comparisons with other state-of-the-art methods are presented in detail. Finally, the ablation experiments prove the efficiency of our algorithm.

4.1. Experiment Setup

Datasets. We experiment with three real-world challenging benchmark datasets: MIT -States [13], UT-Zappos [45] and C-GQA [27] respectively. Specifically, MIT-States contains 53753 natural images, with 115 states and 245 objects. In the closed-world settings, the search space contains 1262 seen compositions and 300 unseen for validation and 400 unseen for test. UT-Zappos consists of 50025 images of shoes, with 16 states and 12 objects. For the closed-world experiments, it is constrained to the 83 seen and 15/18 (validation/test) unseen compositions. And for UT-Zappos, we follow the split in [32]. For about C-GQA, the most pairs dataset for CZSL, contains 453 states and 870 objects, with 39298 images in total, which contains over 9500 compositions. Finally, in the open-world settings, these datasets contain 28175, 192 and 278362 compositions respectively.

Metrics. Following the setting of prior work [23], we compute the prediction accuracy based on the seen and unseen compositions both in the closed-world and open-world scenarios. Specifically, *Seen (S)* denotes the accuracy tested only on seen compositions and *Unseen (U)* represents the accuracy evaluated only on unseen compositions. Also, we can calculate *Harmonic Mean (H)* of the *S* and *U* metrics. Since zero-shot models have inherent bias for seen compositions, we can draw a seen-unseen accuracy curve at different operating points with the bias from $-\infty$ to $+\infty$ to compute the *Area Under the Curve (AUC)*. To sum up, the metrics consist of *S*, *U*, *H* and *AUC*.

Implementation Details. We implement DFSP with PyToch 1.12.1 [31] and optimized by Adam optimizer over the three challenging datasets for 20 epochs. The image encoder and text encoder are both based on the pretrained CLIP Vit-L/14 model, and the entire model are trained and evaluated on 1×NVIDIA RTX 3090 GPU. Besides, we set the number of fusion blocks K and self-attention section as 1, and the evaluation metrics are tested on the model which computes lowest loss during the validation phase.

4.2. Comparision with State-of-the-Arts

Experimental comparisons with the prior compositional zero-shot learning methods are reported, including AoP [28], LE+ [27], TMN [32], SymNet [20], CompCos [23], CGE [27], Co-CGE [24], SCEN [18], KG-SP [14] and recently proposed CSP [30]. For our proposed method DFSP, we test a variety of different fusion methods: *BiF*, *i2t* and *t2i*, which denotes the fusion direction as bidirectional fusion, fusion on the text branch and fusion on the image branch. The experiment is conducted on both closed-world and open-world, and the results are shown in Tab. 1 and Tab. 2.

For the closed-world setting, DFSP in Tab. 1 shows that our method achieves the new state-of-the-art on MIT-States, UT-Zappos and CGQA datasets. DFSP reaches the highest *AUC* of 20.8% on MIT-States, 36.0% on UT-Zappos and 10.5% on CGQA, which outperforms CSP by 4.3%. Besides, we improve the harmonic mean by 6.6% on CGQA relative to other existing methods. And the seen and unseen accuracies on these datasets are also the best results.

Tab. 2 shows the DFSP results on open-world setting and we also get the best results on all metrics. During the inference stage, the infeasible filter threshold T is fixed on 0.4, and DFSP outperforms CSP by 1.1% on MIT-States and KG-SP by 3.8% on UT-Zappos for the *AUC* metric. It can be clearly seen that the unseen accuracy in open-world has improved a lot like 15.9% on UT-Zappos, which proves that the decomposes state and object features fused with the image feature can really enhance the sensitiveness for unseen compositions. This improvement can also drive the *H* metric to the best.

Combining the results of closed-world and open-world, it can be seen that the fusion method of *t2i* is the best, showing that fused with decomposed features on the image branch can achieve better performance than that on the language branch due to the destruction of joint representation even they can be recomposed. Experimental results on three challenging datasets demonstrate that our proposed Decomposed Fusion with Soft Prompt framework (DFSP) can effectively improve the performance of the model for compositional zero-shot learning.

4.3. Ablation Study

To evaluate the effectiveness of DFSP, we also establish an ablation study on MIT-States and UT-Zappos. The soft prompt module in Fig. 2 is the base recognition model and the DFSP version is *t2i*. Meanwhile, we evaluate models with only self-attention (*SA*), with only fusion (*Fusion*),

Table 1. Closed-world results on MIT-States, UT-Zappos and C-GQA. *S* and *U* are the predict accuracies evaluated on seen and unseen compositions. *H* is the harmonic mean of *U* and *S* and *AUC* is the area under the curve. The best results are in bold.

Method	MIT-States				UT-Zappos				CGQA			
	S	U	H	AUC	S	U	H	AUC	S	U	H	AUC
AoP [28]	14.3	17.4	9.9	1.6	59.8	54.2	40.8	25.9	17.0	5.6	5.9	0.7
LE+ [27]	15.0	20.1	10.7	2.0	53.0	61.9	41.0	25.7	18.1	5.6	6.1	0.8
TMN [32]	20.2	20.1	13.0	2.9	58.7	60.0	45.0	29.3	23.1	6.5	7.5	1.1
SymNet [20]	24.2	25.2	16.1	3.0	49.8	57.4	40.4	23.4	26.8	10.3	11.0	2.1
CompCos [23]	25.3	24.6	16.4	4.5	59.8	62.5	43.1	28.1	28.1	11.2	12.4	2.6
CGE [27]	28.7	25.3	17.2	5.1	56.8	63.6	41.2	26.4	28.7	25.3	17.2	5.1
Co-CGE [24]	31.1	5.8	6.4	1.1	62.0	44.3	40.3	23.1	32.1	2.0	3.4	0.5
SCEN [18]	29.9	25.2	18.4	5.3	63.5	63.1	**47.8**	32.0	28.9	25.4	17.5	5.5
CSP [30]	46.6	49.9	36.3	19.4	64.2	66.2	46.6	33.0	28.8	26.8	20.5	6.2
DFSP(*i2t*)	**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*)	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*)	46.9	52.0	37.3	20.6	**66.7**	**71.7**	47.2	**36.0**	**38.2**	**32.0**	**27.1**	**10.5**

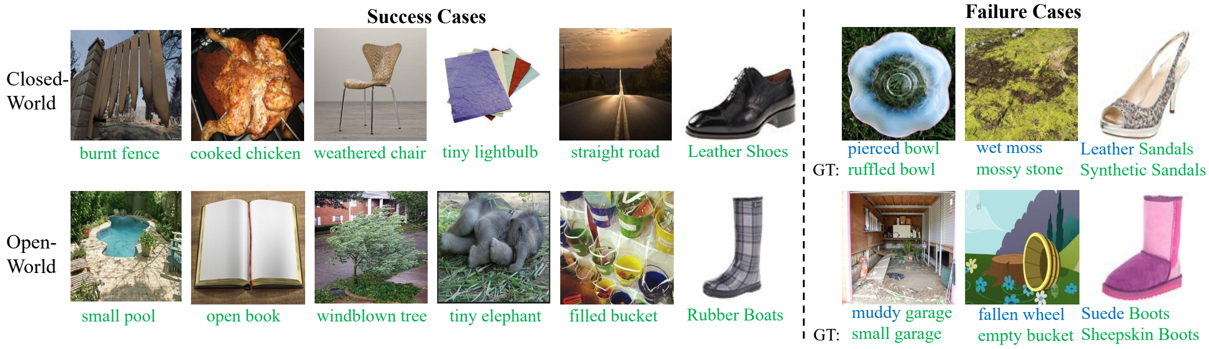

Figure 3. Qualitative results. We evaluate top-1 predictions for some cases on MIT-States and UT-Zappos. The first row shows the results of the closed-world and the bottom row is the open-world. Six cols on the left are examples of successful predictions, and three on the right are examples of failures. For the failure cases, blue denotes the wrong prediction and all images are randomly selected.

with only decomposition section (*DeC*) and with decomposed fusion module (*DFM*). The closed-world and open-world experimental results can be seen in Tab. 3.

Effectiveness of SPM. The experimental results show that the model with only *SPM* can improve a little compared with CSP, demonstrating its fully learnable soft prompts can be better adapted to downstream supervised tasks. If only the language feature is decomposed and not integrated, the experimental results can be improved a lot in multiple metrics, including closed-world and open-world. Meanwhile, the results of *+SA* are also significantly improved compared to *SPM*, which proves the fine-tuning effect of adding self-attention to the model, making it better to transfer VLMs to new tasks.

Effectiveness of DFM. Eventually, the results of *+DFM* show a very large improvement, proving the effectiveness of DFSP. However, the results of *+SA* and *+DeC* are not much different, and *+DFM* both decomposes language features and fuses with image features, which will lead to a particularly large improvement. Also, only *+Fusion* will make the effect worse, mainly causing increasing the bias of seen compositions, which does not meet the definition of CZSL. To a certain extent, this shows that decomposition and fusion complement each other, and only decomposition in the form of combination is the same as the essence of *+SA*. From the results of metric *U*, DFSP can also demonstrate the high response of DFM to unseen compositions. Besides, this is not limited to closed-world, open-world has also seen significant improvements.

4.4. Qualitative Results

We report qualitative results for seen and unseen compositions with top-1 predictions both on closed-world and open-world in Fig. 3. Our model can really be generalized well to unseen compositions, alleviating the domain gap between seen and unseen sets. Meanwhile, benefited from the joint prompt consists of state and object, DFSP can predict the compositions with high accuracy. For the failure cases, the most prone to error is the prediction of state, but even if the prediction is not correct, it still conforms to the combi-

Table 2. Open-world results on MIT-States, UT-Zappos and C-GQA. *S* and *U* are the predict accuracies evaluated on seen and unseen compositions. *H* is the harmonic mean of *U* and *S* and *AUC* is the area under the curve. The best results are in bold.

Method	MIT-States				UT-Zappos				CGQA			
	S	U	H	AUC	S	U	H	AUC	S	U	H	AUC
AoP [28]	16.6	5.7	4.7	0.7	50.9	34.2	29.4	13.7	-	-	-	-
LE+ [27]	14.2	2.5	2.7	0.3	60.4	36.5	30.5	16.3	19.2	0.7	1.0	0.08
TMN [32]	12.6	0.9	1.2	0.1	55.9	18.1	21.7	8.4	-	-	-	-
SymNet [20]	21.4	7.0	5.8	0.8	53.3	44.6	34.5	18.5	26.7	2.2	3.3	0.43
CompCos [23]	25.4	10.0	8.9	1.6	59.3	46.8	36.9	21.3	-	-	-	-
CGE [27]	32.4	5.1	6.0	1.0	61.7	47.7	39.0	23.1	32.7	1.8	2.9	0.47
Co-CGEˆClosed [24]	31.1	5.8	6.4	1.1	62.0	44.3	40.3	23.1	32.1	2.0	3.4	0.53
Co-CGEˆOpen [24]	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 [14]	28.4	7.5	7.4	1.3	61.8	52.1	42.3	26.5	31.5	2.9	4.7	0.78
CSP [30]	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.20
DFSP(*i2t*)	47.2	18.2	19.1	6.7	64.3	53.8	41.2	26.4	35.6	6.5	9.0	1.95
DFSP(*BiF*)	47.1	18.1	19.2	6.7	63.5	57.2	42.7	27.6	36.4	**7.6**	**10.6**	2.39
DFSP(*t2i*)	**47.5**	**18.5**	**19.3**	**6.8**	**66.8**	**60.0**	**44.0**	**30.3**	**38.3**	7.2	10.4	**2.40**

Table 3. Ablation study experiments on MIT-States and UT-Zappos with the setting of closed-world (CW) and open-world (OW). The best results are in bold.

Method		MIT-States				UT-Zappos			
		S	U	H	AUC	S	U	H	AUC
CW	*SPM*	45.8	50.2	35.8	19.1	64.6	63.9	47.0	33.1
	+*SA*	47.6	51.9	**37.4**	20.6	65.3	66.2	46.6	32.6
	+*Fusion*	46.9	47.5	34.9	18.3	62.9	41.4	38.0	21.4
	+*DeC*	**48.3**	51.3	37.3	**20.7**	62.3	70.7	46.8	33.5
	+*DFM*	46.9	**52.0**	37.3	20.6	**66.7**	**71.7**	**47.2**	**36.0**
OW	*SPM*	45.8	16.7	18.3	6.1	64.6	44.6	40.8	23.5
	+*SA*	**47.6**	17.6	19.0	6.6	65.4	54.6	43.0	26.9
	+*DeC*	47.4	17.7	19.3	6.6	61.7	57.1	42.3	26.4
	+*DFM*	47.5	**18.5**	**19.3**	**6.8**	**66.8**	**60.0**	**44.0**	**30.3**

nation logic of state and object. Great results can be seen on both closed-world and open-world setting, indicating that the model is not restricted by open-world scenario.

4.5. Why DFSP can work well?

Extensive experiments show the efficiency of DFSP and we analyze the reasons of this. Firstly, since the encoders of DFSP have been pretrained on a large-scale image-text pairs dataset, the model can be really fine-tuned well with a targeted prompt like CSP. Compared with CSP, there are more parameters can be fine-tuned in DFSP to be adapted to new supervised tasks [8]. With the Decomposed Fusion Module (DFM), the pair space is transformed from the original pairing of ”language” and ”image” to the pairing of ”language + (image)” and ”(state and object) + image” (as shown in Fig. 2), and both decomposed state feature and object feature can respond to image feature in the fusion stage. Being fused with the decomposed features, image feature can establish its relation to state and object, then pair with another branch (like language feature in DFSP (*t2i*)) to guide the results beyond the seen compositions during training. Benefited from this, the overall model can be more sensitive to the unseen compositions in the pair space.

5. Conclusion

In this work, we propose a novel framework termed Decomposed Fusion with Soft Prompt (DFSP) to effectively recognize the unknown compositions of state and object during training. Based on the vision-language paradigm, we firstly establish a learnable soft prompt consists of prefix, state and object to construct the joint representation of state and object. Besides, we design a Decomposed Fusion Module (DFM) to fuse the language features with image features, which can enable cross-modal interactions between them. Meanwhile, DFM decomposes the language feature to unattached state and object features, and then they will be fused with image feature to guide enhancing fusion. Benefited from DFM, image feature could learn relations with state and object features, which improves the response of unseen compositions in the pair space. Extensive experiments on three challenging datasets demonstrate the efficiency of our proposed method DFSP.

Acknowledgement

This research was supported by fundings from the Key-Area Research and Development Program of Guangdong Province (No. 2021B0101400003), Hong Kong RGC Research Impact Fund (No. R5060-19), Areas of Excellence Scheme (AoE/E-601/22-R), General Research Fund (No. 152203/20E, 152244/21E, 152169/22E, 152211/23E), the National Natural Science Foundation of China (No. 62102327), PolyU Internal Fund (No. P0043932), and Shenzhen Science and Technology Innovation Commission (JCYJ20200109142008673).

References

[1] Stephen H Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, et al. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv preprint arXiv:2202.01279*, 2022. 3

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 3

[3] 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. 3

[4] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 6

[5] Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Wen Zhang, Yin Fang, Jeff Z Pan, Wenting Song, and Huajun Chen. Duet: Cross-modal semantic grounding for contrastive zero-shot learning. *arXiv preprint arXiv:2207.01328*, 2022. 3

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[8] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 9

[9] Yanan Gu, Cheng Deng, and Kun Wei. Class-incremental instance segmentation via multi-teacher networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1478–1486, 2021. 3

[10] Jingcai Guo and Song Guo. A novel perspective to zero-shot learning: Towards an alignment of manifold structures via semantic feature expansion. *IEEE Transactions on Multimedia*, 23:524–537, 2020. 3

[11] Jingcai Guo, Song Guo, Qihua Zhou, Ziming Liu, Xiaocheng Lu, and Fushuo Huo. Graph knows unknowns: Reformulate zero-shot learning as sample-level graph recognition. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI-23*, 2023. 3

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[13] 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. 7

[14] 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. 2, 3, 4, 7, 9

[15] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465, 2013. 3

[16] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 3

[17] Xiangyu Li, Zhe Xu, Kun Wei, and Cheng Deng. Generalized zero-shot learning via disentangled representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1966–1974, 2021. 3

[18] 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. 3, 7, 8

[19] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 3

[20] 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. 2, 3, 7, 8, 9

[21] Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and Tatsuya Harada. Goal-oriented gaze estimation for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3794–3803, 2021. 3

[22] Ziming Liu, Song Guo, Jingcai Guo, Yuanyuan Xu, and Fushuo Huo. Towards unbiased multi-label zero-shot learning with pyramid and semantic attention. *IEEE Transactions on Multimedia*, 2022. 3

[23] 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. 3, 7, 8, 9

[24] 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, 3, 6, 7, 8, 9

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 3

[26] 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. 2, 3

[27] 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. 7, 8, 9

[28] 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, 7, 8, 9

[29] Zhixiong Nan, Yang Liu, Nanning Zheng, and Song-Chun Zhu. Recognizing unseen attribute-object pair with generative model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8811–8818, 2019. 2, 3

[30] Nihal V Nayak, Peilin Yu, and Stephen H Bach. Learning to compose soft prompts for compositional zero-shot learning. *arXiv preprint arXiv:2204.03574*, 2022. 3, 5, 6, 7, 8, 9

[31] 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. 7

[32] 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. 3, 7, 8, 9

[33] Guanghui Qin and Jason Eisner. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*, 2021. 3

[34] 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. 3, 5

[35] Frank Ruis, Gertjan Burghouts, and Doina Bucur. Independent prototype propagation for zero-shot compositionality. *Advances in Neural Information Processing Systems*, 34:10641–10653, 2021. 3

[36] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021. 3

[37] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020. 3

[38] 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. 6

[39] Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. Spot: Better frozen model adaptation through soft prompt transfer. *arXiv preprint arXiv:2110.07904*, 2021. 3

[40] Kun Wei, Cheng Deng, Xu Yang, et al. Lifelong zero-shot learning. In *IJCAI*, pages 551–557, 2020. 3

[41] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020. 6

[42] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *International conference on machine learning*, pages 6861–6871. PMLR, 2019. 3

[43] 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(9):2251–2265, 2018. 4

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

[45] 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. 7

[46] 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. 3