An Integrated Approach to Open-World Compositional Zero-Shot Learning

Hirunima Jayasekara[*1] **Khoi Pham**[1] **Nirat Saini**[1]

Abhinav Shrivastava[1]
[1]University of Maryland, College Park
[*]hirunima@umd.edu

Abstract

Open-World Compositional Zero-Shot Learning (OW-CZSL) addresses the challenge of recognizing novel compositions of known primitives and entities. Even though prior works utilize language knowledge for recognition, such approaches exhibit limited interactions between language-image modalities. Our approach primarily focuses on enhancing the inter-modality interactions through fostering richer interactions between image and textual data. Additionally, we introduce a novel module aimed at alleviating the computational burden associated with exhaustive exploration of all possible compositions during the inference stage. While previous methods exclusively learn compositions jointly or independently, we introduce an advanced hybrid procedure that leverages both learning mechanisms to generate final predictions. Our proposed model, achieves state-of-the-art in OW-CZSL in three datasets.

1 Introduction

Compositional Zero-Shot Learning (CZSL) involves generating new compositions from established primitives and entities. Prior works are generally classified into closed-world and open-world settings. Conventional CZSL Saini u. a. (2022); Kim u. a. (2023); Xu u. a. (2021); Wang u. a. (2023); Yang u. a. (2020) approaches use a closed-world setting, requiring prior knowledge of unseen pairs. Mancini et al. proposed an open-world setting, encompassing all possible combinations of attributes and objects $|A| \cdot |O|$ (with $|A|$ attributes and $|O|$ objects), which is more effective for real-world deployments. To manage the large label space in OW-CZSL, KG-SP Karthik u. a. (2022) classifies primitives (objects and attributes) separately, reducing computational burden, though it does not explicitly learns pair compositions. In contrast, direct composition recognition, as used by Nayak *et al.*Nayak u. a. (2022), becomes challenging with an increasing number of attributes and objects. Our proposed approach integrates the strengths of both paradigms, enabling the model to learn primitives both independently and compositionally via a novel sparse linear layer.

Textual features serve a crucial role in incorporating semantic knowledge into the learning framework Saini u. a. (2022). Therefore, fusion of visual and textual modalities essential for disentangling and generating compositions. However, most works in CZSL Mancini u. a. (2021); Naeem u. a. (2021) adopt a simple fusion approach, projecting both modalities into a shared embedding space and generating a similarity function between them, which requires computing similarity scores for each potential composition during inference. To mitigate this, we propose a Top-K selection module that automatically identifies a subset of candidate text embeddings at an early stage. We hypothesize that simple fusion, even from high-performing unimodal embedders, may be insufficient for learning complex vision-and-language compositions Kim u. a. (2021); Nguyen u. a. (2020), motivating the need for a more effective inter-modal interaction method in compositional learning. Our key contributions are as follows:

NeurIPS 2024 Workshop on Compositional Learning: Perspectives, Methods, and Paths Forward.

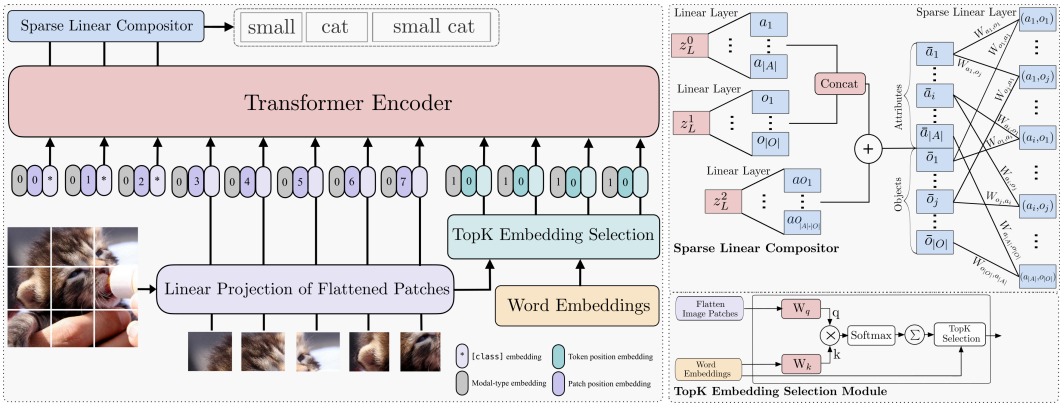

Figure 1: The overall architecture of the proposed method. Input embeddings to the transformer encoder are formed by concatenating image patch embeddings and text embeddings. **TopK Embedding Selection Module** effectively selects relevant text embeddings that align with the provided image through cross attention. **Sparse Linear Compositor** computes attribute and object predictions alongside a final prediction vector, utilizing a sparse linear layer.

- Introducing a unified framework for OW-CZSL by utilizing a single transformer while preserving model performance to attain superior results compared to LLVMs showing state-of-the-art performance.
- Proposed Top-K selection module mitigate the challenge of exhaustively exploring all potential pairings during the inference process.
- A novel sparse linear layer is proposed to facilitate the integration of multiple tokens, aiding in the disentanglement of attributes and objects and enabling the generation of compositions with computational complexity reduced to $\mathcal{O}(|A| \cdot |O|)$.

2 Method

2.1 Problem Formulation

We follow the open world setting proposed by Mancini *et al.*Mancini u. a. (2021) for OW-CZSL. Each training sample consists of two elements, Image $x \in X$ and corresponding text pair $t = (t_{attr}, t_{obj}) \in T$. Followed by attribute and object labels, $y = (y_{attr}, y_{obj}) \in Y$. Y consists of two subsets: seen pairs Y^s and unseen pairs Y^u which, contain all the attribute and object compositions that were present during training and compositions that are not available during training respectively. Resulting, $Y^s \cup Y^u = Y$ and $Y^s \cap Y^u = \emptyset$.

2.1.1 Linear Patch Projection and Language Embeddings

We utilize the linear flattened patch projection schema from ViT Dosovitskiy u. a. (2020) where, the image $x \in \mathbb{R}^{H \times W \times C}$ flattened into to $v \in \mathbb{R}^{N \times (P^2 \cdot C)}$ with $N = HW/P^2$ number of $P \times P$ patch projections. Followed by linear projection $V \in \mathbb{R}^{(P^2 \cdot C) \times H}$ of v, producing $x_v \in \mathbb{R}^{N \times (P^2 \cdot C)}$ as visual embeddings for the network. x_v, combined with three $[\texttt{class}]$ tokens for attributes, objects and pairs produces patch embeddings $\tilde{v}$. Lastly, positional embeddings $V^{pos} \in \mathbb{R}^{(N+1) \times H}$ are embedded to the patch embeddings.

We create a fixed auxiliary vocabulary input for each dataset by collecting BERT embeddings Kenton und Toutanova (2019) for each attribute or object in T. For each text, the last embedding output $u_i^{attr} \in \mathbb{R}^{(|A|) \times d}, u_i^{obj} \in \mathbb{R}^{(|O|) \times d}$ is extracted and concatenated to form a vocabulary embedding matrix $u_{vocab} \in \mathbb{R}^{(|A|+|O|) \times d}$ where d is the dimension of the vocabulary embedding and $d = P^2 \cdot C$.

$$\tilde{v} = [v_{attr}; v_{obj}; v_{pair}; v_1 V; \cdots; v_N V] + V^{pos} \quad \text{and} \quad u_{voc} = [u_1^{attr}; \cdots; u_{|A|}^{attr}; u_1^{obj}; \cdots; u_{|O|}^{obj}] \tag{1}$$

2.2 TopK embedding Selection

The objective of this module is to perform visual-assisted vocabulary mapping to create the text input for the primary multi-modality transformer. To represent a class, we combine two text embeddings from the vocabulary: one pertaining to attribute and the other to object. For brevity, we explain TopK attribute selection, while TopK object selection follows the same procedure. To find relevance scores

for the attribute entry, we use respective image embeddings $\tilde{v}$ as the Query vector and the attribute portion of text embeddings $u_{vocab}[0:|A|]$ as the Key vector. We impose cross attention between Query and Key to produce an attention map, $\mathcal{A}_a \in \mathbb{R}^{(P^2 \cdot C) \times |A|}$. As illustrated in Figure 1, final score is calculated by summing $\mathcal{A}_a$ along the image axis to produce a 1-D vector consisting of attention scores, $\tilde{\mathcal{A}}_a \in \mathbb{R}^{|A|}$ for each word. .

$$\mathcal{A}_a = \text{softmax}\left(\frac{\tilde{v} \cdot u_{voc}[0:|A|]}{\sqrt{P^2 \cdot C}}\right) \quad \text{and} \quad \tilde{\mathcal{A}}_a = \sum_{i=0}^{P^2 C} \mathcal{A}_a[i] \tag{2}$$

From sorted $\tilde{\mathcal{A}}$, we select text embeddings of words that have top K highest attention scores. The same procedure is repeated for selecting top K object embeddings by utilizing the sorted attention scores corresponding to objects $\tilde{\mathcal{A}}_o \in \mathbb{R}^{|O|}$. Text input $\tilde{u} \in \mathbb{R}^{2K}$ to multi modality transformer is created by

$$\tilde{u} = [u^{attr}_{\tilde{\mathcal{A}}_a[1]}; \cdots; u^{attr}_{\tilde{\mathcal{A}}_a[K]}; u^{obj}_{\tilde{\mathcal{A}}_o[1]}; \cdots; u^{obj}_{\tilde{\mathcal{A}}_o[k]}] \tag{3}$$

2.3 Language based Visual Modality Transformer

Similar to Kim *et al.*Kim u. a. (2021), we initialize the transformer weights from pre-trained ViT Dosovitskiy u. a. (2020) weights rather than that of BERT. This expects to bolster the model's ability to process visual features effectively, thereby addressing the challenge of lacking a separate uni-model visual embedder.

In order to separate the two modalities, the text and image embeddings are combined with their respective modal-type learnable embedding vectors, denoted as v^{type} and u^{type}, where $u_{type}, v_{type} \in \mathbb{R}^{P^2 \cdot C}$ then concatenated along the embedding axis to form input sequence $z^0 \in \mathbb{R}^{M \times P^2 \cdot C}$ where, $M = N + 2K + 3$ is the total number of input tokens.

$$z^0 = [\tilde{v} + v^{type}; \tilde{u} + u^{type}] \tag{4}$$

Following Dosovitskiy *et al.*Dosovitskiy u. a. (2020), we iteratively update the contextualized vector through 12 transformer layers with multi-head self-attention, using ViT-B/16 pre-trained on ImageNet-21K with a hidden size of 768, patch size of 16, and 12 attention heads.

2.3.1 Sparse Linear Compositor

Given three [`class`] tokens, Sparse Linear Compositor(SLC) computes an attribute prediction vector, an object prediction vector as auxiliary outputs and a composition vector between $|A|$ number of attributes and $|O|$ number of objects. To derive the auxiliary outputs, the first two [`class`] tokens are processed through two learnable MLP heads.

$$\tilde{y}_{\text{attr}} = MLP_{\text{attr}}(z_0^D) \quad \text{and} \quad \tilde{y}_{\text{obj}} = MLP_{\text{obj}}(z_1^D) \tag{5}$$

Where, $\tilde{y}_{\text{attr}} \in \mathbb{R}^{|A|}$ and $\tilde{y}_{\text{obj}} \in \mathbb{R}^{|O|}$. In order to compute the decomposition pair predictions, we normalize the linear head outputs and multiply to create $\tilde{y}_{\text{decompose}} \in \mathbb{R}^{(|A| \cdot |O|)}$

$$\tilde{y}_{\text{decompose}} = Norm(\tilde{y}_{\text{attr}}) \times Norm(\tilde{y}_{\text{obj}}) \tag{6}$$

Third [`class`] token is combined with the row wise concatenation of attribute and object predictions to produce $\tilde{z}_{\text{pair}} \in \mathbb{R}^{|A|+|O|}$.

$$z_{\text{pair}} = MLP_{\text{pair}}(z_2^D) \quad \text{and} \quad \tilde{z}_{\text{pair}} = z_{\text{pair}} + \text{concat}(\tilde{y}_{\text{attr}}, \tilde{y}_{\text{obj}}) \tag{7}$$

As illustrated in Figure 1, pair output is calculated by learnable weighted ($W_a \in \mathbb{R}^{|A| \cdot |O|}, W_o \in \mathbb{R}^{|O| \cdot |A|}$) addition of corresponding attribute and object of each composition of $\tilde{z}_{\text{pair}}$. Resulting compositional pair prediction $\tilde{y}_{\text{compose}} \in \mathbb{R}^{|A| \cdot |O|}$.

$$\tilde{y}_{\text{compose}}[A_i, O_j] = \tilde{z}_{\text{pair}}[A_i] \odot W_a^i + \tilde{z}_{\text{pair}}[O_j] \odot W_o^j \tag{8}$$

Where, $W_a \in \mathbb{R}^{|A| \cdot |O|}, W_o \in \mathbb{R}^{|O| \cdot |A|}$, $i \in (1, ..., |A|)$ and $j \in (1, ..., |O|)$. Proposed Sparse Linear Layer requires only $2(|A| \cdot |O|)$ number of learnable parameters while a standard linear layer would require $(|A| + |O|)(|A| \cdot |O|)$ number of parameters. Lastly, we combine both decomposition pair prediction and compositional pair prediction to produce final pair prediction with scale factor η.

$$\tilde{y} = \tilde{y}_{\text{decompose}} + \eta \tilde{y}_{\text{compose}} \tag{9}$$

2.4 Classification Head

For classification of attributes, objects and pairs, we extract the three [`class`] tokens from the last layer output z^D of the transformer network and process each token through SLC to get $\tilde{y}$.

$$\tilde{y} = SLC(z_0^D, z_1^D, z_2^D) \tag{10}$$

Table 1: Open world performance on MIT-States, C-GQA and VAW-CZSL. As evaluation matrices we refer to AUC with seen and unseen accuracy with different bias terms along with HM.

Method	Backbone	MIT				C-GQA				VAW-CZSL			
		S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
TMN Purushwalkam u. a. (2019)	R18	12.6	0.9	1.2	0.1	-	-	-	-	-	-	-	-
VisProd Misra u. a. (2017)	R18	20.9	5.8	5.6	0.7	24.8	1.7	2.8	0.33	-	-	-	-
SymNet Li u. a. (2020)	R18	21.4	7.0	5.8	0.8	26.7	2.2	3.3	0.43	-	-	-	-
ComCos Mancini u. a. (2021)	R18	25.4	10.0	8.9	1.6	28.4	1.8	2.8	0.39	4.3	1.0	1.1	0.03
CGE Naeem u. a. (2021)	R18	32.4	5.1	6.0	1.0	32.7	1.8	2.9	0.47	8.6	2.8	2.2	0.13
KG-SP Karthik u. a. (2022)	R18	28.4	7.5	6.7	1.3	31.5	2.9	4.7	0.78	6.4	2.4	1.8	0.08
SAD-SP Li u. a. (2023)	R18	29.1	7.6	7.8	1.4	31.0	3.9	5.9	1.0	-	-	-	-
DRANet Liu u. a. (2023)	R18	29.8	7.8	7.9	1.5	31.3	3.9	6.0	1.05	-	-	-	-
ADE Hao u. a. (2023)	ViT-B	-	-	-	-	35.1	4.8	7.6	1.42	-	-	-	-
KG-SP$_{vit}$	ViT-B	28.6	11.8	10.3	2.1	31.3	3.4	5.1	0.87	15.0	4.5	3.9	0.38
Ours	ViT-B	**36.3**	**12.5**	**12.4**	**3.1**	**35.8**	**5.6**	**7.8**	**1.6**	**16.5**	**6.7**	**6.7**	**0.82**

In OW-CZSL, as the label space expands proportionally to $|A| \cdot |O|$, we deploy a filtering schema similar to previous OW-CZSL works Karthik u. a. (2022); Nayak u. a. (2022) to exclude unfeasible compositions. Aggregated text embeddings from ConceptNet Speer u. a. (2017) and GloVe Pennington u. a. (2014) are utilized to calculate feasibility scores by using cosine similarity between attributes and objects. A threshold-based binary mask is then applied to generate the final compositional output.

$$\bar{y} = \tilde{y} \cdot f_{pair} \tag{11}$$

2.5 Training Objectives

Compared to conventional OW-CZSL setting, rather than computing cosine similarity in an embedding space, the proposed method utilizes cross-entropy loss over predictions. Namely, **Pair Loss:** Cross entropy loss between pair predictions $\bar{y}$ with ground truth y. **TopK embedding Loss:** Cross-entropy loss between TopK selection module outputs with ground truth y_{attr} and y_{obj}. **Disentangling Loss:** Cross-entropy loss between final attribute and object predictions $\tilde{y}_{attr}$ and $\tilde{y}_{obj}$ with the ground truth y_{attr} and y_{obj}. A combined loss function $\mathcal{L}$ is minimized over all the training images, to train the proposed method end-to-end manner. The weights for each loss ($\alpha_i, i = 1, 2, 3$) are empirically computed.

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha_1 \mathcal{L}_{topk} + \alpha_2 \mathcal{L}_{attr} + \alpha_3 \mathcal{L}_{obj} \tag{12}$$

3 Experiments

3.1 Datasets and Metrics

When evaluating the proposed model, we refer to three datasets, MIT-states Isola u. a. (2015) (28175 pairs), C-GQA Mancini u. a. (2021)(278362 pairs) and VAW-CZSL Saini u. a. (2022)(238040 pairs). We use general OW-CZSL setup suggested by Purushwalkam u. a. (2019) combine with the evaluation statistics. Namely, best Seen accuracy (S), best Unseen accuracy (U), Area Under the Curve (AUC) and Harmonic Mean (HM).

3.2 Results

Table 1 presents a comprehensive summary of the primary experiments and compares the performance of the proposed model against that of various baseline models. Even though, MIT-states contains label noise Atzmon u. a. (2020), our model shows better performance over seen accuracy with a 7.7% margin along with a 2.1% increment in HM and a 1% increment in AUC demonstrating higher performance compared to KG-SP$_{vit}$. Amidst C-GQA containing significantly higher number of compositions, proposed model was able to achieve state-of-the-art performance. With 4.1% seen accuracy improvement in addition to 2.2% increment in HM over KG-SP$_{vit}$. Resulting gain in AUC of 0.6%. We evaluate our model on the VAW-CZSL dataset as a refined alternative to C-GQA, achieving a 9.5% improvement in seen accuracy, a 0.6% increase in AUC, and a 3.5% gain in HM, while maintaining comparable unseen accuracy with KG-SP$_{vit}$.

4 Conclusion

In this work, we introduce a unified framework for OW-CZSL to enhance inter-modality interactions beyond prior approaches with shallow interactions. Top-K selection module reduces inference ambiguity, while the sparse linear compositor improves generalization by decomposing and composing attributes and objects. Extensive experiments show our model outperforms previous methods across three benchmarks.

References

[Atzmon u. a. 2020] ATZMON, Yuval ; KREUK, Felix ; SHALIT, Uri ; CHECHIK, Gal: A causal view of compositional zero-shot recognition. In: *Advances in Neural Information Processing Systems* 33 (2020), S. 1462–1473

[Dosovitskiy u. a. 2020] DOSOVITSKIY, Alexey ; BEYER, Lucas ; KOLESNIKOV, Alexander ; WEISSENBORN, Dirk ; ZHAI, Xiaohua ; UNTERTHINER, Thomas ; DEHGHANI, Mostafa ; MINDERER, Matthias ; HEIGOLD, Georg ; GELLY, Sylvain u. a.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *arXiv preprint arXiv:2010.11929* (2020)

[Hao u. a. 2023] HAO, Shaozhe ; HAN, Kai ; WONG, Kwan-Yee K.: Learning attention as disentangler for compositional zero-shot learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, S. 15315–15324

[Isola u. a. 2015] ISOLA, Phillip ; LIM, Joseph J. ; ADELSON, Edward H.: Discovering states and transformations in image collections. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, S. 1383–1391

[Karthik u. a. 2022] KARTHIK, Shyamgopal ; MANCINI, Massimiliano ; AKATA, Zeynep: 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*, 2022, S. 9336–9345

[Kenton und Toutanova 2019] KENTON, Jacob Devlin Ming-Wei C. ; TOUTANOVA, Lee K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of naacL-HLT* Bd. 1, 2019, S. 2

[Kim u. a. 2023] KIM, Hanjae ; LEE, Jiyoung ; PARK, Seongheon ; SOHN, Kwanghoon: Hierarchical visual primitive experts for compositional zero-shot learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, S. 5675–5685

[Kim u. a. 2021] KIM, Wonjae ; SON, Bokyung ; KIM, Ildoo: Vilt: Vision-and-language transformer without convolution or region supervision. In: *International Conference on Machine Learning* PMLR (Veranst.), 2021, S. 5583–5594

[Li u. a. 2020] LI, Yong-Lu ; XU, Yue ; MAO, Xiaohan ; LU, Cewu: Symmetry and group in attribute-object compositions. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, S. 11316–11325

[Li u. a. 2023] LI, Yun ; LIU, Zhe ; JHA, Saurav ; YAO, Lina: Distilled reverse attention network for open-world compositional zero-shot learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, S. 1782–1791

[Liu u. a. 2023] LIU, Zhe ; LI, Yun ; YAO, Lina ; CHANG, Xiaojun ; FANG, Wei ; WU, Xiaojun ; EL SADDIK, Abdulmotaleb: Simple Primitives with Feasibility-and Contextuality-Dependence for Open-World Compositional Zero-shot Learning. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023)

[Mancini u. a. 2021] MANCINI, Massimiliano ; NAEEM, Muhammad F. ; XIAN, Yongqin ; AKATA, Zeynep: Open world compositional zero-shot learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, S. 5222–5230

[Misra u. a. 2017] MISRA, Ishan ; GUPTA, Abhinav ; HEBERT, Martial: From red wine to red tomato: Composition with context. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, S. 1792–1801

[Naeem u. a. 2021] NAEEM, Muhammad F. ; XIAN, Yongqin ; TOMBARI, Federico ; AKATA, Zeynep: Learning graph embeddings for compositional zero-shot learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, S. 953–962

[Nayak u. a. 2022] NAYAK, Nihal V. ; YU, Peilin ; BACH, Stephen H.: Learning to compose soft prompts for compositional zero-shot learning. In: *arXiv preprint arXiv:2204.03574* (2022)

[Nguyen u. a. 2020] NGUYEN, Duy-Kien ; GOSWAMI, Vedanuj ; CHEN, Xinlei: Movie: Revisiting modulated convolutions for visual counting and beyond. In: *arXiv preprint arXiv:2004.11883* (2020)

[Pennington u. a. 2014] PENNINGTON, Jeffrey ; SOCHER, Richard ; MANNING, Christopher D.: Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, S. 1532–1543

[Purushwalkam u. a. 2019] PURUSHWALKAM, Senthil ; NICKEL, Maximilian ; GUPTA, Abhinav ; RANZATO, Marc'Aurelio: Task-driven modular networks for zero-shot compositional learning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, S. 3593–3602

[Saini u. a. 2022] SAINI, Nirat ; PHAM, Khoi ; SHRIVASTAVA, Abhinav: Disentangling Visual Embeddings for Attributes and Objects. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, S. 13658–13667

[Speer u. a. 2017] SPEER, Robyn ; CHIN, Joshua ; HAVASI, Catherine: Conceptnet 5.5: An open multilingual graph of general knowledge. In: *Proceedings of the AAAI conference on artificial intelligence* Bd. 31, 2017

[Wang u. a. 2023] WANG, Qingsheng ; LIU, Lingqiao ; JING, Chenchen ; CHEN, Hao ; LIANG, Guoqiang ; WANG, Peng ; SHEN, Chunhua: Learning Conditional Attributes for Compositional Zero-Shot Learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, S. 11197–11206

[Xu u. a. 2021] XU, Ziwei ; WANG, Guangzhi ; WONG, Yongkang ; KANKANHALLI, Mohan S.: Relation-aware compositional zero-shot learning for attribute-object pair recognition. In: *IEEE Transactions on Multimedia* 24 (2021), S. 3652–3664

[Yang u. a. 2020] YANG, Muli ; DENG, Cheng ; YAN, Junchi ; LIU, Xianglong ; TAO, Dacheng: Learning unseen concepts via hierarchical decomposition and composition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, S. 10248–10256