Recognizing Unseen States of Unknown Objects by Leveraging Knowledge Graphs -Supplementary Material-

Filippos Gouidis[1,2] Konstantinos Papoutsakis[3] Theodore Patkos[1] Antonis Argyros[1,2]
Dimitris Plexousakis[1,2]
[1] Foundation for Research and Technology-Hellas, Greece
[2] University of Crete, Greece [3] Hellenic Mediterranean University, Greece
{gouidis,patkos,argyros,dp}@ics.forth.gr, kpapoutsakis@hmu.gr

1. Datasets Details

Table 1 presents the following details for each dataset: i) the number of the training, validation and test samples; ii) the number of state and object classes; iii) the valid and iv) the total object-state combinations and v) the average number of states in which an object can be situated.

2. Evaluation of the CW and OW versions

The results for the Open World (OW) and Closed World (CW) versions of the models are shown in Table 2 and Table 3, respectively. For the OW settings our method continues to outperform the competing methods, although the performance gain has predictably been decreased. Moreover, w.r.t OSDD dataset, the 2nd best method is IVR [14], whereas CANET [12] is the 3rd best method. In the case of the CGQA-States dataset, the 2nd and 3rd best method is IVR [14] and CANET [12], respectively. Concerning the MIT-States dataset the 2nd best method is the IVR [14], whereas KG-SP [4] exhibits the 3rd best AUC score and CANET [12] the 3rd best HM score. Finally, in the case of the VAW dataset, the 2nd best performance is achieved by CANET [12], while IVR [14] ranks 3rd.

Regarding the CW settings, our method ranks 1st for the OSDD, VAW and MIT-states datasets and 4th for the CGQA-states dataset. Regarding the OSDD dataset, IVR [14] exhibits the 2nd best performance and KG-SP [4] the 3rd best performance. In the case of MIT-States dataset, CompCos [7] achieves the 2nd best performance and ADE [2] the 3rd best performance. Concerning the CGQA-states dataset, the best performance is achieved by CANET [12], the 2nd best by CompCos [7] and the 3rd best by OADiS [13]. Finally, regarding VAW, the 2nd best method is ADE [2] and the 3rd best method is CANET [12].

3. Additional Results of the Ablation Study

Table 4 outlines the details of the employed KGs, while Table 5 summarizes the performance of all ablated models across the four datasets.
1st Sub-table (GNN Architectures): The Tr-GCN-based model CN+WN_H2_TH_GCN demonstrates the best overall performance.
2nd Sub-table (KGs): The ConceptNet-based model CN_H2_TH_Tr-GCN achieves the highest scores.
3rd Sub-table (Hops): Most models achieve their best performance with two hops.
4th Sub-table (Node Policy): Adopting a node policy slightly improves the performance of most models.

Notably, while CN_H2_TH_Tr-GCN achieves the best scores on two of the three datasets, CN+WN_H2_TH_GCN was selected for comparison with competing methods, as this selection was based on aggregate averages across all four categories.

In seen classes, the model using unrelated embeddings (CN_H3_UN_Tr-GCN) achieves similar accuracy to its counterpart with standard embeddings (CN_H3_Tr-GCN). However, CN_H3_UN_Tr-GCN performs significantly worse in unseen classes, with its HM and AUC scores being three to four times lower than those of CN_H3_Tr-GCN. In contrast, the random model performs poorly across all metrics.

The key distinction between CN_H3_UN_Tr-GCN and the random model lies in their embedding distributions: in the former, the GNN enables a balanced and representative distribution, while in the latter, the distribution is entirely random. This suggests that fine-tuning can yield competitive accuracy for seen classes even when embeddings are unrelated to target labels, provided they are distributed effectively. However, for unseen classes, accuracy depends on a precise mapping between embeddings and target labels.

Dataset	Train	Val	Test	States	Objects	VOSC	TOSC	S\O
OSDD [1]	6,977	1,124	5,275	9	14	35	126	2.36
CGQA-states [7]	244	46	806	5	17	41	75	1.71
MIT-states [3]	170	34	274	5	14	20	70	1.57
VAW [10]	2,752	516	1,584	9	23	51	207	2.61

Table 1. Details about the four image datasets utilized in this work. Train/Val/Test: Number of Training/Validation/Testing Images. States: Number of State classes, Objects: Number of Object classes. VOSC/TOSC: Valid/Total Object-State combinations. S\O: Average number of states than an Object can be situated in.

Method	OSDD				CGQA-States				MIT-States				VAW			
	S	Un	HM	AUC	S	Un	HM	AUC	S	Un	HM	AUC	S	Un	HM	AUC
AoP [9]	69.9	33.3	31.6	13.3	14.5	4.3	4.4	0.3	36.4	4.8	8.4	1.3	59.6	5.4	6.1	1.3
LE+ [8]	71.6	14.3	20.8	6.5	29.1	4.0	7.0	0.6	45.5	14.9	15.1	4.3	23.7	12.3	13.7	0.4
TMN [11]	73.4	43.6	33.7	19.0	45.5	29.7	19.3	6.1	69.7	18.4	22.4	6.3	77.6	35.5	26.8	14.3
SymNet [6]	77.7	14.0	21.1	7.5	94.0	7.1	13.7	6.1	97.0	1.9	2.1	0.9	82.2	3.1	3.5	1.2
CompCos [7]	78.7	31.5	42.0	22.1	95.5	4.0	7.7	3.4	75.8	2.5	4.9	1.2	75.8	2.5	4.9	1.2
KG-SP [4]	77.0	29.8	35.4	17.9	94.0	16.9	26.1	12.7	97.0	15.5	22.6	12.0	74.3	12.3	17.6	8.6
SCEN-NET [5]	75.8	25.5	26.3	10.7	83.6	7.4	13.6	5.9	36.4	8.5	13.0	1.6	22.0	12.0	11.1	2.5
IVR [14]	78.8	61.6	**44.2**	**30.8**	94.0	40.3	**37.4**	**26.4**	96.9	22.5	**24.5**	**14.9**	87.2	37.4	29.7	18.2
OADiS [13]	76.5	20.5	27.1	10.7	94.8	26.3	20.3	12.0	93.9	29.1	23.4	12.5	82.8	8.9	11.0	4.2
CANET [12]	79.2	43.9	43.7	27.2	95.5	51.3	41.9	26.1	96.9	19.3	22.7	11.4	90.1	53.9	**40.4**	**29.7**
ADE [2]	80.2	27.6	32.3	12.3	95.5	16.3	25.7	12.8	78.8	4.5	4.7	0.8	80.8	22.3	14.3	8.4
OaSC (Ours)	87.7	69.9	**48.6**	**39.8**	97.1	73.4	**43.6**	**36.5**	85.7	69.9	**51.1**	**41.2**	83.7	58.6	**42.9**	**32.8**

Table 2. Aggregate results for Open World Versions. S: Best Accuracy on seen classes. UN: Best accuracy on unseen classes. HM: Best harmonic mean. AUC: Area under curve for the pairs of accuracy for seen and unseen classes. Red/Bold/Underlined text indicates best/2nd best/3rd best performance.

Method	OSDD				CGQA-States				MIT-States				VAW			
	S	UN	HM	AUC	S	UN	HM	AUC	S	UN	HM	AUC	S	UN	HM	AUC
AoP [9]	75.9	53.5	32.2	19.5	95.5	50.0	35.9	27.8	48.5	20.9	15.1	4.1	55.1	44.7	24.1	11.6
LE+ [8]	68.6	31.7	34.5	16.9	93.5	16.1	16.1	8.1	63.6	14.6	20.3	7.1	41.6	2.3	2.6	1.2
TMN [11]	71.5	49.8	35.0	20.8	97.0	76.0	39.9	32.2	84.9	30.7	27.4	16.1	82.6	55.5	37.3	25.6
SymNet [6]	77.7	59.4	44.2	31.0	95.5	27.4	39.4	24.4	96.9	27.5	26.8	15.7	89.2	46.6	40.0	27.4
Compcos [7]	76.3	45.3	38.7	23.8	92.5	73.9	**48.1**	**41.5**	100.0	44.9	**32.3**	**23.8**	88.4	51.4	39.3	29.1
KG-SP [4]	78.0	55.0	**47.6**	29.7	95.5	17.7	27.2	13.5	97.1	15.5	22.6	12.0	89.4	37.3	39.3	23.4
SCEN-NET [5]	75.1	45.6	39.4	22.7	94.1	53.4	41.1	31.0	84.9	23.1	22.1	11.5	90.5	44.2	37.7	23.5
IVR [14]	78.4	60.5	46.0	**31.8**	94.0	43.4	35.2	25.2	87.9	28.8	27.1	14.0	86.7	38.2	30.5	18.5
OADiS [13]	78.7	59.7	38.3	26.2	95.5	78.6	43.5	36.7	93.9	29.4	28.3	17.2	89.9	61.8	39.8	30.5
CANET [12]	80.3	43.6	45.1	27.9	95.5	64.9	**50.0**	**43.3**	96.9	23.0	28.2	15.9	90.3	54.6	40.8	30.5
ADE [2]	82.0	42.5	35.9	20.6	94.8	58.3	45.5	34.9	93.9	27.5	30.4	19.2	90.7	45.0	**40.9**	**30.6**
OaSC (Ours)	87.7	69.9	**48.6**	**39.8**	97.1	73.4	43.6	36.5	85.7	69.9	**51.1**	**41.2**	83.7	58.6	**42.9**	**32.8**

Table 3. Aggregate results for Closed World Versions. S: Best Accuracy on seen classes. UN: Best accuracy on unseen classes. HM: Best harmonic mean. AUC: Area under curve for the pairs of accuracy for seen and unseen classes. Red/Bold/Underlined text indicates best/2nd best/3rd best performance.

KG	N	E	RT	RC
WN_H2	70 / 54 / 49 / 79	321 / 223 / 105 / 365	5	LX
WN_H3	429 / 311 / 295 / 465	873 / 680 / 655 / 912	5	LX
CN_H2	715 / 552 / 504 / 743 /	2,132 / 1,981 / 1,864 / 2,342	13	CS
CN_H3	2,139 / 1,872 / 1,788 /2,349 /	2,542 / 2,194 / 2,103 / 2,874	24	CS
CN_H2_TH	611 / 505 / 485 / 785	1,710 / 1,521 / 1,415 / 1,956	12	CS
CN_H3_TH	12,733 / 9,839 / 9,212 / 13,045	29,794 / 25,105 / 24,292 / 32,456	29	CS
CN+WN_H2	667 / 581 / 506 / 845	1,906 / 1,682 / 1,602 / 2,136	13	CS
CN+WN_H2_TH	590 / 492 / 431 / 705	1,442 / 1,167 / 1,089 / 1,673	12	CS/LX
CN+WN_H3_TH	10,165 / 8,842 / 7,948 / 12,116	26,735 / 23,176 / 22,602 / 29,672	29	CS/LX

Table 4. KGs Details. N: Number of Nodes. E: Number of Edges. RT: Number of Different Relation Types between nodes. RC: Category of Relation Types. CS: Common-Sense. LX: Lexicographic. First/Second/Third/Fourth number in the N and E columns refers to the KG for OSDD/CGQA-States/MIT-States/ VAW dataset, respectively.

Method	OSDD				CGQA-States				MIT-States				VAW			
	S	Un	HM	AUC	S	UN	HM	AUC	S	UN	HM	AUC	S	UN	HM	AUC
CN_H3_LSTM	85.1	38.0	38.0	24.3	96.4	57.1	37.3	27.0	92.9	65.4	50.9	36.9	55.7	43.9	22.1	12.5
CN_H3_GCN	86.7	58.5	**44.1**	**34.0**	95.7	62.5	40.0	28.7	88.1	66.7	47.1	32.2	70.3	49.5	30.2	20.8
CN_H3_R-GCN	87.7	49.0	42.7	30.4	95.7	71.4	**40.9**	**34.0**	78.6	73.4	47.4	32.9	79.5	57.5	38.9	28.8
CN_H3_Tr-GCN	87.4	42.2	40.2	27.7	93.6	56.3	39.2	28.8	88.1	67.0	**53.6**	**43.7**	80.2	56.8	**40.7**	**29.9**
WN_H3_LSTM	86.0	60.0	**43.3**	**33.9**	96.4	13.4	16.6	8.7	90.5	24.4	24.2	13.2	37.4	55.6	18.1	10.2
WN_H3_GCN	86.8	39.5	36.7	21.2	86.4	49.0	34.2	24.1	88.1	54.8	**50.1**	**37.9**	64.2	38.3	24.4	19.4
WN_H3_R-GCN	85.5	36.0	36.5	22.1	93.6	52.9	**40.5**	**28.9**	78.6	47.4	42.9	21.4	69.7	56.0	38.9	28.8
WN_H3_Tr-GCN	89.2	48.4	36.6	23.9	86.4	56.6	37.6	26.6	88.1	44.2	37.3	25.9	65.0	54.5	**31.8**	**21.3**
CN_H2_TH_LSTM	86.5	50.0	43.0	28.8	97.1	71.7	38.8	31.9	78.6	60.3	47.8	26.0	61.0	52.6	27.9	17.9
CN_H2_TH_GCN	84.6	52.8	43.7	30.7	95.7	67.5	40.5	32.0	85.7	73.1	46.6	29.4	74.3	48.3	36.4	27.4
CN_H2_TH_R-GCN	85.9	48.0	41.2	28.5	95.0	63.6	41.6	31.6	81.0	69.2	51.8	30.0	82.4	57.6	40.5	31.5
CN_H2_TH_Tr-GCN	85.7	63.7	**45.6**	**34.5**	97.1	70.0	**43.5**	**35.6**	85.7	70.2	**51.6**	**40.5**	82.4	59.4	**38.0**	**32.6**
WN_H2_Tr-GCN	87.9	23.0	28.6	13.0	92.9	53.8	**38.2**	**28.1**	83.3	45.8	**39.7**	**27.3**	69.7	45.8	30.5	18.3
WN_H3_Tr-GCN	89.2	48.4	**36.6**	**23.9**	86.4	56.6	37.6	26.6	88.1	44.2	37.3	25.9	65.0	54.5	**31.8**	**21.3**
CN_H2_Tr-GCN	86.4	60.6	**45.1**	**34.3**	97.1	73.4	**46.3**	**39.5**	88.1	69.6	**56.2**	**43.5**	82.4	58.9	**37.3**	**32.0**
CN_H3_Tr-GCN	87.4	42.2	40.2	27.7	93.6	56.3	39.2	28.8	88.1	67.0	53.6	43.7	81.1	48.3	36.9	26.3
CN_H3_UN_Tr-GCN	85.7	14.8	**17.0**	**7.6**	93.6	13.2	**15.1**	**7.4**	83.3	26.6	**20.6**	**7.6**	83.1	10.2	**14.8**	**5.3**
RN_Tr-GCN	12.9	11.3	3.2	1.6	15.7	9.7	5.1	2.5	26.7	24.2	12.5	4.6	12.0	9.8	3.0	1.3
CN+WN_H2_Tr-GCN	85.7	60.9	45.2	33.9	97.1	72.0	**46.0**	**38.9**	88.1	68.9	55.3	43.3	82.0	58.9	39.8	32.6
CN+WN_H2_TH_Tr-GCN	87.7	69.9	**48.6**	**39.8**	97.1	73.4	43.6	36.5	85.7	69.9	**51.1**	**41.2**	83.7	58.6	**42.9**	**32.8**
WN_H2_Tr-GCN	87.9	23.0	28.6	13.0	92.9	53.8	**38.2**	**28.1**	83.3	45.8	**39.7**	**27.3**	69.7	45.8	30.5	18.3
WN_H3_Tr-GCN	89.2	48.4	**36.6**	**23.9**	86.4	56.6	37.6	26.6	88.1	44.2	37.3	25.9	65.0	54.5	**31.8**	**21.3**
CN_H2_Tr-GCN	86.4	60.6	45.1	34.3	97.1	73.4	**46.3**	**39.5**	88.1	69.6	**56.2**	**43.5**	82.4	58.9	**37.3**	**32.0**
CN_H3_Tr-GCN	87.4	42.2	**40.2**	**27.7**	93.6	56.3	39.2	28.8	88.1	67.0	53.6	43.7	80.2	56.8	40.7	29.9
CN+WN_H2_TH_Tr-GCN	87.7	69.9	**48.6**	**39.8**	97.1	73.4	43.6	36.5	85.7	69.9	51.1	41.2	83.7	58.6	**42.9**	**32.8**
CN+WN_H3_TH_Tr-GCN	87.1	56.3	44.6	31.9	97.1	60.5	41.0	32.5	83.3	68.6	**55.9**	**41.0**	80.6	59.2	38.8	30.6
WN_H3_Tr-GCN	87.3	46.4	**35.7**	23.0	85.5	53.6	35.3	25.2	87.2	44.3	**37.4**	25.7	65.0	54.5	31.8	21.3
WN_H3_TH_Tr-GCN	89.2	48.4	36.6	**23.9**	86.4	56.6	**37.6**	**26.6**	88.1	44.2	37.3	**25.9**	68.1	56.0	**32.7**	**23.4**
CN_H2_Tr-GCN	86.4	60.6	45.1	34.3	97.1	73.4	**46.3**	**39.5**	88.1	69.6	**56.2**	**43.5**	82.4	58.9	37.3	32.0
CN_H2_TH_Tr-GCN	85.7	63.7	**45.6**	**34.5**	97.1	70.0	43.5	35.6	85.7	70.2	51.6	40.5	82.4	59.4	**38.0**	**32.6**

Table 5. Ablation Study. 1st section of the table: comparison for the GNN architecture. 2nd section: comparison for the KG source. 3rd section: comparison for max number of hops. 4th section: comparison for the node inclusion policy. Bold font indicates top performance across ablation category. Blue colour indicates top performance across ablation subcategory. S: Best Accuracy on seen classes. UN: Best accuracy on unseen classes. HM: Best harmonic mean. AUC: Area under curve for the pairs of accuracy for seen and unseen classes. CN: ConceptNet-based model. WN: WordNet-based model. UN: Embeddings corresponding to concepts unrelated to the target classes. RN: Random embeddings. H2(3): Maximum number of hops equal to 2(3). TH: Thresholding policy for the nodes of the KG.

References

[1] Gouidis, F., Patkos, T., Argyros, A., Plexousakis, D.: Detecting object states vs detecting objects: A new dataset and a quantitative experimental study. In: Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP). vol. 5, pp. 590–600 (2022)

[2] Hao, S., Han, K., Wong, K.Y.K.: Learning attention as disentangler for compositional zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15315–15324 (2023)

[3] Isola, P., Lim, J.J., Adelson, E.H.: Discovering states and transformations in image collections. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **07-12-June**, 1383–1391 (2015). https://doi.org/10.1109/CVPR.2015.7298744

[4] Karthik, S., Mancini, M., Akata, Z.: Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In: 35th IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2022)

[5] Li, X., Yang, X., Wei, K., Deng, C., Yang, M.: Siamese contrastive embedding network for compositional zero-shot learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9326–9335 (2022)

[6] Li, Y.L., Xu, Y., Mao, X., Lu, C.: Symmetry and group in attribute-object compositions pp. 11316–11325 (2020)

[7] Mancini, M., Naeem, M.F., Xian, Y., Akata, Z.: Learning Graph Embeddings for Open World Compositional Zero-Shot Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **8828**(c), 1–15 (2022). https://doi.org/10.1109/TPAMI.2022.3163667

[8] Misra, I., Gupta, A., Hebert, M.: From red wine to red tomato: Composition with context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1792–1801 (2017)

[9] Nagarajan, T., Grauman, K.: Attributes as operators: factorizing unseen attribute-object compositions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 169–185 (2018)

[10] Pham, K., Kafle, K., Lin, Z., Ding, Z., Cohen, S., Tran, Q., Shrivastava, A.: Learning to predict visual attributes in the wild. In: Proceedings of the IEEE/CVF CVPR. pp. 13018–13028 (June 2021)

[11] Purushwalkam, S., Nickel, M., Gupta, A., Ranzato, M.: Task-driven modular networks for zero-shot compositional learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3593–3602 (2019)

[12] Qingsheng Wang, Lingqiao Liu, C.J.H.C.G.L.P.W.C.S.: Learning conditional attributes for compositional zero-shot learning. In: CVPR (2023)

[13] Saini, N., Pham, K., Shrivastava, A.: Disentangling visual embeddings for attributes and objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13658–13667 (June 2022)

[14] Zhang, T., Liang, K., Du, R., Sun, X., Ma, Z., Guo, J.: Learning invariant visual representations for compositional zero-shot learning. In: ECCV (2022)