# Open World Compositional Zero-Shot Learning
# -
# Supplementary Material

Massimiliano Mancini[1*], Muhammad Ferjad Naeem[1,2*], Yongqin Xian[3], Zeynep Akata[1,3,4]
[1]University of Tübingen [2]TU München [3]MPI for Informatics [4]MPI for Intelligent Systems

## 1. Expanded Results

### 1.1. Comparison with the State of the Art

In Table 1 of the main paper, we reported the comparison between CompCosand the state of the art, in both closed and open world settings. As highlighted in the methodological section, closed and open world are different problems with different challenges (*i.e*. bias on the seen classes for the first, presence of distractors in the second). For this reason, in the closed world experiments, we reported the results of the closed world version of our model (Section 3.2), while our full model is used for the more complex OW-CZSL (Section 3.3). Here, we expand the table, reporting the results of the closed world (CompCos$^{\text{CW}}$) and full (CompCos) versions of our model for both closed and open world scenarios.

Table 1 shows the complete results for both MIT states and UT Zappos. As we can see, both versions of our model achieve competitive results on the closed world scenario and in both datasets. In this setting, our full model, CompCos, achieves slightly lower performance than CompCos$^{\text{CW}}$, with a 4.1 AUC vs the 4.5 AUC of our closed world counterpart on MIT states, and a 27.1 vs 28.7 of AUC on UT Zappos. This is because our full model focuses less on balancing seen and unseen compositions (the crucial aspect of standard closed world CZSL) but mostly on the margin between feasible and unfeasible compositions. This latter goal is not helpful in the closed world setting, where the subset of feasible compositions seen at test time is known a priori. Nevertheless, the performance of CompCos largely surpasses the previous state of the art on MIT states in AUC, with a 1.1 increase of AUC over SymNet.

On the other hand, if we exclude CompCos, our closed world model (CompCos$^{\text{CW}}$) achieves the highest AUC when applied in the open world scenario, in both datasets. In particular, it is comparable to SymNet on MIT states (0.8 vs 0.9 AUC) while surpassing it by 2.3 AUC on UT Zappos. On MIT states, it achieves a lower performance unseen accuracy with respect to SymNet (*i.e*. 5.5% vs 7.0%). We believe this is because SymNet is already robust to the inclusion of distractors by modeling objects and states separately during inference. Nevertheless, our full approach is the best in all compositional metrics and in both datasets. In particular, on MIT states it improves CompCos$^{\text{CW}}$ by 4.5% on best unseen accuracy, 3.0% on best harmonic mean, and 0.8 of AUC. This confirms the importance of including the feasibility of each composition during training.

### 1.2. Masked Inference

**Values of feasibility scores and thresholds.** The feasibility scores on the unseen compositions range from 0.31 to 0.82 on UT Zappos, and from -0.01 to 0.68 on MIT states. For $f_{\text{HARD}}$, we ablated the threshold values on the validation set of each dataset. We found the best threshold values to be 0.34 and 0.27 respectively.

**Ablating Masked Inference.** In the main paper (Table 3), we tested the impact of thresholding the feasibility scores to explicitly exclude unfeasible compositions from the output space of the model (Section 3.3, Eq. (6)). In particular, Table 3 shows how the binary masks obtained from CompCos can greatly improve the performance of our closed world model, CompCos$^{\text{CW}}$, and other approaches (*i.e*. LabelEmbed+, TMN) while being only slightly beneficial to more robust ones such as our full method CompCos and SymNet.

Here we analyze whether the effect of the mask is linked to limiting the output space of the model or to their ability to excluding the majority of the distractors (*i.e*. less feasible compositions). To test this, we apply to the output space of CompCos and CompCos$^{\text{CW}}$, two additional binary masks. The first is obtained by thresholding the feasibility scores using their median (*median*), keeping as valid unseen compositions all the ones with the score above the median. The second is the reverse, *i.e*. we keep as valid all the seen compositions, and all the unseen compositions whose feasibility scores are below the median (*i.e*. *inv. median*).

What we expect is that, if the feasibility scores are not meaningful, distractors are equally excluded, no matter if we consider the top half or the bottom half of the scores. If this happens, the performance boost would be only linked

| Method | Closed World | | | | | | | | | | | | Open World | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MIT states | | | | | | UT Zappos | | | | | | MIT states | | | | | | UT Zappos | | | | | |
| | Sta. | Obj. | S | U | HM | auc | Sta. | Obj. | S | U | HM | auc | Sta. | Obj. | S | U | HM | auc | Sta. | Obj. | S | U | HM | auc |
| AoP[3] | 21.1 | 23.6 | 14.3 | 17.4 | 9.9 | 1.6 | 38.9 | 69.9 | **59.8** | 54.2 | 40.8 | 25.9 | 15.4 | 20.0 | 16.6 | 5.7 | 4.7 | 0.7 | 25.7 | 61.3 | 50.9 | 34.2 | 29.4 | 13.7 |
| LE+[2] | 23.5 | 26.3 | 15.0 | 20.1 | 10.7 | 2.0 | 41.2 | 69.3 | 53.0 | 61.9 | 41.0 | 25.7 | 10.9 | 21.5 | 14.2 | 2.5 | 2.7 | 0.3 | **38.1** | 68.2 | **60.4** | 36.5 | 30.5 | 16.3 |
| TMN[4] | 23.3 | 26.5 | 20.2 | 20.1 | 13.0 | 2.9 | 40.8 | 69.5 | 58.7 | 60.0 | **45.0** | **29.3** | 6.1 | 15.9 | 12.6 | 0.9 | 1.2 | 0.1 | 14.6 | 61.5 | 55.9 | 18.1 | 21.7 | 8.4 |
| SymNet[1] | 26.3 | 28.3 | 24.2 | **25.2** | 16.1 | 3.0 | 41.3 | 68.6 | 49.8 | 57.4 | 40.4 | 23.4 | 17.0 | 26.3 | 21.4 | 7.0 | 5.8 | 0.8 | 33.2 | 70.0 | 53.3 | 44.6 | 34.5 | 18.5 |
| **CompCos$^{CW}$** | **27.9** | **31.8** | 25.3 | 24.6 | **16.4** | **4.5** | **44.7** | **73.5** | **59.8** | **62.5** | 43.1 | 28.7 | 13.9 | **28.2** | 25.3 | 5.5 | 5.9 | 0.9 | 33.8 | 72.4 | 59.8 | 45.6 | 36.3 | 20.8 |
| **CompCos** | 26.7 | 30.0 | **25.6** | 22.7 | 15.6 | 4.1 | 44.0 | 73.3 | 59.3 | 61.5 | 40.7 | 27.1 | **18.8** | 27.7 | **25.4** | **10.0** | **8.9** | **1.6** | 35.1 | **72.4** | 59.3 | **46.8** | **36.9** | **21.3** |

Table 1. Closed and Open World CZSL results on MIT states and UT Zappos. We measure states (Sta.) and objects (Obj.) accuracy on the primitives, best seen (S) and unseen accuracy (U), best harmonic mean (HM), and area under the curve (auc) on the compositions.

to the fact that we exclude a portion of the output space, and not to the actual unfeasibility of the excluded compositions. Consequently, we would expect CompCos and CompCos$^{CW}$ to achieve the same results when either *median* or *inv. median* are applied as masks on the output space.

The results of this analysis are reported in Table 2, where we report also the results of not excluding any composition (-) and of the best threshold value (*best*). As the Table shows, the performance gaps are very large if we take as valid the compositions having the top or the bottom half of the scores. In particular, in CompCos$^{CW}$ performance go from 1.3 to 0.03 in AUC, 7.5% to 0.3% in harmonic mean, and from 6.9% to 0.1% in best unseen accuracy. CompCos shows a similar behavior, with the AUC going from 2.2 to 0.06, the harmonic mean from 10.9% to 0.6%, and the best unseen accuracy from 11.1% to 0.4%. These results clearly demonstrate that i) the boost brought by masking the output space is linked to the exclusion of unfeasible compositions rather than a simple reduction of the search space; ii) feasibility scores are meaningful, with the feasible compositions tending to receive the top-50% of the feasibility scores.

Finally, in Figure 1 we analyze the impact of that the hard masking threshold on CompCoson the validation set of MIT states. As the figure shows, low threshold values remove a few percentage (green) of the compositions and the AUC is comparable to the base model with no hard masking (red). By increasing the threshold, the AUC increases up to the point where the output space is overly restricted and also (feasible) compositions of the dataset are discarded. Indeed, hard masking can work only if the similarity scores (and the ranking of compositions they produce) are meaningful, otherwise even low values would mask out feasible compositions from the dataset, harming the model's performance.

| | Mask | Seen | Unseen | HM | AUC |
|---|---|---|---|---|---|
| CompCos$^{CW}$ | - | 28.0 | 6.0 | 7.0 | 1.2 |
| | median | | 6.9 | 7.5 | 1.3 |
| | inv. median | | 0.1 | 0.3 | .03 |
| | best | | 8.1 | 8.7 | 1.6 |
| CompCos | - | 27.1 | 11.0 | 10.8 | 2.1 |
| | median | | 11.1 | 10.9 | 2.2 |
| | inv. median | | 0.4 | 0.6 | .06 |
| | best | | 11.2 | 11.0 | 2.2 |

Table 2. Results on MIT states validation set for applying our feasibility-based binary masks ($f_{\text{HARD}}$) on CompCos$^{CW}$ and CompCos with different strategies.

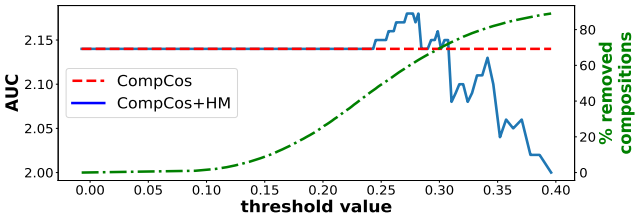


Figure 1. CompCos: AUC vs hard masking threshold on MIT states' validation set. The green line denotes the percentage of removed compositions at a given threshold value.

## 2. Additional Qualitative Results

### 2.1. Feasibility Scores

In this section, we focus on MIT states and we report additional qualitative analyses on the most and least feasible compositions, as for the feasibility scores computed by our model. In particular, in Table 3 we show the top-3 and bottom-3 states associated to 25 randomly selected objects, while in Table 4 we show the top-3 and bottom-3 objects associated to 25 randomly selected states.

Similarly to the analysis of the main paper, in Table 3 we can see how the highest feasibility scores are generally linked to related sub-categories of objects/states. For instance, *gate* is related to conservation-oriented status (*i.e. cracked*, *dented*) while cooking states (*i.e. cooked*, *raw*, *diced*) are considered its most unfeasible. A similar observation applies to *necklace*, associated to conservation status (*i.e. pierced*, *scratched*) while states related to atmospheric conditions (*i.e. cloudy*, *open* related to *sky*) are considered unfeasible. Cooking states are the most feasible for *chicken*

| Objects | States | |
|---|---|---|
| | Most Feasible (Top-3) | Least Feasible (Bottom-3) |
| aluminum | unpainted, thin, coiled | full, closed, young |
| apple | peeled, caramelized, diced | full, standing, short |
| bathroom | grimy, cluttered, steaming | fallen, unripe, cooked |
| beef | browned, sliced, steaming | standing, cluttered, fallen |
| blade | broken, straight, shiny | young, sunny, cooked |
| bronze | melted, crushed, pressed | full, open, winding |
| cave | tiny, verdant, damp | blunt, whipped, diced |
| chicken | diced, thawed, cooked | standing, open, closed |
| dress | wrinkled, ripped, folded | cloudy, open, closed |
| fence | crinkled, weathered, thick | short, full, cooked |
| garlic | browned, sliced, squished | standing, full, closed |
| gate | closed, cracked, dented | cooked, raw, diced |
| glasses | broken, dented, crushed | cloudy, smooth, sunny |
| island | small, foggy, huge | blunt, short, open |
| jacket | crumpled, wrinkled, torn | cloudy, young, full |
| library | huge, modern, heavy | blunt, cooked, viscous |
| necklace | thick, pierced, scratched | cloudy, full, open |
| potato | caramelized, sliced, mashed | full, short, standing |
| ribbon | creased, frayed, thick | cloudy, sunny, cooked |
| rope | thick, curved, frayed | modern, ripe, cluttered |
| shower | dirty, empty, tiny | standing, unripe, young |
| steps | small, large, dented | blunt, fresh, raw |
| stream | foggy, verdant, dry | young, standing, closed |
| sword | shattered, blunt, rusty | ripe, full, cooked |
| wool | thick, crumpled, ruffled | full, fallen, closed |

Table 3. Unseen compositions wrt their feasibility scores: Top-3 highest and Bottom-3 lowest feasible state per object.

| States | Objects | |
|---|---|---|
| | Most Feasible (Top-3) | Least Feasible (Bottom-3) |
| ancient | town, house, road | sauce, well, foam |
| barren | canyon, river, jungle | penny, handle, paint |
| bright | coast, cloud, island | handle, key, drum |
| closed | gate, window, garage | persimmon, berry, copper |
| cloudy | coast, shore, beach | gemstone, penny, shoes |
| cooked | meat, salmon, chicken | field, gate, library |
| creased | shirt, newspaper, shorts | animal, fire, lightning |
| fresh | vegetable, pasta, meat | handle, key, steps |
| loose | shorts, jacket, clothes | butter, seafood, salmon |
| mashed | tomato, potato, fruit | book, stream, deck |
| moldy | apple, pear, sauce | handle, drum, key |
| molten | candy, butter, milk | shoes, cat, animal |
| painted | wood, granite, metal | fig, well, book |
| peeled | tomato, apple, pear | cat, cave, ocean |
| pressed | steel, cotton, silk | field, well, beach |
| ripped | jacket, hat, dress | cat, seafood, cave |
| shiny | blade, stone, sword | city, well, animal |
| squished | vegetable, garlic, bean | road, shore, cat |
| sunny | beach, ocean, sea | card, penny, wire |
| thawed | meat, seafood, chicken | wave, handle, cat |
| unpainted | aluminum, roof, metal | animal, book, lightning |
| unripe | persimmon, vegetable, potato | gear, shoes, phone |
| verdant | valley, pond, coast | penny, keyboard, book |
| winding | highway, tube, wire | bear, armor, beef |
| worn | pants, clothes, shorts | seafood, animal, fire |

Table 4. Unseen compositions wrt their feasibility scores: Top-3 highest and Bottom-3 lowest feasible object per state.

(*i.e*. *diced*, *thawed*, *cooked*) while cloth states are related to *jacket* (*i.e*. *crumpled*, *wrinkled*, *torn*), as expected.

In Table 4, we show a different analysis, *i.e*. we check what are the most/least feasible objects given a state. Even in this case, we see a similar trend, with food (*i.e*. *potato*, *tomato*, *sauce*) associated as feasible to food-related states (*e.g*. *cooked*, *mashed*, *moldy*, *unripe*) while clothing items (*e.g*. *shirt*, *jacket*, *dress*) associated as feasible to clothing-related states (*i.e*. *draped*, *loose*, *ripped*). On the other hand, we can see how environments (*e.g*. *ocean*, *beach*) are associated as feasible to their meteorological state (*e.g*. *sunny*, *cloudy*) but not to manipulation ones (*e.g*. *bent*, *pressed*).

Overall, Tables 3 and 4 show how the feasibility scores capture the subgroups to which objects/states belong. This suggests that when the feasibility scores are introduced as margins within the model, related subgroups are enforced to be closer in the output space than unrelated ones, improving the discrimination capabilities of the model.

## 2.2. Qualitative examples

In this subsection, we report additional qualitative examples, comparing the predictions of our full model CompCos with its closed world counterpart, CompCos$^{CW}$, on MIT states. Similarly to Figure 3 of the main paper, in Figure 2, we show examples of images misclassified by CompCos$^{CW}$ but correctly classified by CompCos. The figure confirms that CompCos$^{CW}$ is less capable than CompCos to deal with the presence of distractors. In fact, there are cases where CompCos$^{CW}$ either misclassifies the object (*e.g*. *cave* vs *canyon*, *bread* vs *brass*), the state (*e.g*. *steaming* vs *thawed*, *moldy* vs *frayed*) or both terms of the composition (*i.e*. *broken well* vs *rusty gear*, *curved light-bulb* vs *coiled hose*). While in some cases the answer is close to the correct one (*e.g*. *unripe tomato* vs *unripe lemon*, *crushed coal* vs *crushed rock*) in others the error is mainly caused by the presence of less feasible compositions in the output space (*e.g*. *deflated chicken*, *melted soup*). These compositions are not correctly isolated by CompCos$^{CW}$, thus they hamper the discriminative capability of the model itself. This does not happen with our full model CompCos where unfeasible compositions are better modeled and isolated in the compositional space.

As a second analysis, in Figure 3, we show some examples where both CompCos and CompCos$^{CW}$ are incor-

rect. Even in this case, it is possible to highlight the differences among the answers given by the two models. CompCos$^{CW}$ being less capable of dealing with the presence of distractors, tends to give implausible answers in some cases (*e.g*. *inflated apple*, *coiled car*, *young copper*, *wilted tiger*). On the other hand, our full model still gives plausible answers, despite those being different from the ground-truth. For instance, while CompCos$^{CW}$ misclassifies the *caramelized chicken* as *caramelized pizza*, CompCos classifies it as *caramelized beef*, with the actual object (*i.e*. *beef* vs *chicken*) being hardly distinguishable from the picture, even for a human. There are other examples in which CompCos recognizes a state close to the one of the ground-truth (*e.g*. *inflated* vs *filled*, *shattered* vs *broken*, *weathered* vs *rusty*, *eroded* vs *muddy*) or a plausible composition given the content of the image e.g. *crinkled fabric, spilled cheese*. We also reported one example where the prediction of our model is correct while the annotation being incorrect (*i.e*. *sliced potato* vs *squished bread*) and some where the prediction of the model is compatible with the content of the image, as well as the ground-truth (*e.g*. *young bear*, *dented car*, *thick pot*). We found the last observations to be particularly interesting, highlighting another problem that future works should tackle in CZSL: the presence of multiple states in a single image.

## References


[1] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *CVPR*, 2020. 2

[2] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *CVPR*, 2017. 2

[3] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *ECCV*, 2018. 2

[4] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *ICCV*, 2019. 2


Figure 2. Examples correct predictions of CompCos in the OW-CZSL scenario when the CompCos$^{CW}$ fails. The first row shows the predictions of the closed world model, the bottom row shows the results of CompCos. The images are randomly selected.


Figure 3. Examples of wrong predictions of CompCos and CompCos$^{CW}$ in the OW-CZSL scenario. The first row shows the predictions of the closed world model, the second row shows the results of CompCos, the third row the ground-truth (GT). Images are randomly selected.