We would like to thank reviewers for recognizing our original contribution (R7), convincing experiments (R6, 7, 8) and clarity/reproducibility of the paper (R6, 8). We appreciate suggestions from R6, 7, 8 and will include these in the paper.

Response to Reviewer 5

– C1: "*lack of novelty: pretty similar to DAZLE*" **Answer:** We respectfully disagree. As mentioned by R7, our contribution is to generate attribute-based features and compose them for recognition of unseen and seen classes, which is original and has not been done before, including DAZLE.

– C2: "*how to classify unseen classes.*" **Answer:** We use the discriminative model $p(y|\boldsymbol{H}, \boldsymbol{z})$ to compute probabilities of unseen classes given their semantic vectors $\boldsymbol{z}$ (line 187) and classify a sample as its most probable unseen class.

– C3: "*... compared with more state-of-the-art methods.*" **Answer:** As mentioned by R6, 7, 8, our experiments are extensive. We have included most competitive methods with comparable settings to ours at the submission time.

– C4: "*What if the attributes of one unseen class are not shared with seen classes?*" **Answer:** Please notice that ALL zero-shot learning works rely on sharing attributes between seen and unseen classes. Without sharing common attributes between seen and unseen classes, there is no way to transfer knowledge to unseen classes for recognition.

Response to Reviewer 6

– C1: "*The improvements ... on CUB and DeepFashion are limited.*" **Answer:** Please notice that we significantly improve zero-shot accuracy by at least 3.2% and 2.2% on DeepFashion and CUB, respectively, compared to other methods (lines 252-253) while having no extra learnable parameter w.r.t. DAZLE.

– C2: "*performance of ... seen categories is not so good.*" **Answer:** Please notice that high performance on seen classes is not the main goal of zero-shot learning as it can cause seen class bias (low unseen class performance). Our method achieves competitive seen class accuracies while obtaining the best unseen accuracies, hence, high harmonic means.

Response to Reviewer 7

– C1: "*Providing e.g. pseudo code describing the whole process ... training process*" **Answer:** Thanks for the suggestion. We will include the shown Algorithm 1.

Algorithm 1 Self-Composition

Input: Training set D, DAZLE model, ResNet101 backbone
Initialize discriminative model $p(y|\boldsymbol{H}, \boldsymbol{z})$ from DAZLE
for $t = 1, \ldots, N_{iteration}$ **do**
    Sample a set $S \subset D$ with an equal number of samples per class (lines 153 and 242)
    Extract region features $\{\boldsymbol{f}_i^r\}_{r=0}^R$ for $i \in S$ via ResNet101
    Extract dense features $\boldsymbol{H}_i$ for $i \in S$ via DAZLE model
    Construct $\mathrm{Q}(\boldsymbol{z}^u)$ for $u \in \mathcal{C}_u$ via solving Eq (8) with OMP
    Compose features for $u \in \mathcal{C}_u$ based on $\mathrm{Q}(\boldsymbol{z}^u)$ via Eq (10)
    Update $p(y|\boldsymbol{H}, \boldsymbol{z})$ on real/composed features via Eq (11)
end for
Output: Optimal discriminative model $p(y|\boldsymbol{H}, \boldsymbol{z})$

– C2: "*miss ... work relying on class similarity graphs*" **Answer:** Thanks. We will discuss these works, if accepted. Compared to the reported numbers from the most comparable work (Ding et al., CVPR 2019), our method outperforms this work in harmonic mean by a significant margin of 25% on CUB, AWA2 and 5% on SUN.

– C3: "*Is a feature vector constructed per attribute? ... Is S sampled such that all attributes are present?*" **Answer:** As mentioned in Remark 1, the composed dense features consist of a feature vector per attribute which is required for $p(y|\boldsymbol{H}, \boldsymbol{z})$ in DAZLE. Although we try to include samples from various classes having different attributes (line 153, 242), we cannot guarantee S contains all present attributes without extra annotations of present attributes in images. However, our framework can compose features from any set S by solving Eq (10) even with missing attributes in S.

– C4: "*How ... no Comp relate to the DAZLE model?*" **Answer:** Please notice that they are different. The `No Comp` variant is trained via cross-entropy loss while DAZLE [10] is trained with an extra self-calibration loss (lines 67-68).

– C5: "*Could the approach been extended to ... learned attributes?*" **Answer:** For learned attributes, we can factorize attribute representations into common components (via PCA) which can be used as the building blocks for composition.

Response to Reviewer 8

– C1: "*semantic vectors z^c are available at training time for both seen *and* unseen classes ... [15, 16, 17], ... goes against the philosophy of zero-shot learning.*" **Answer:** To be comparable with [15 - 17], we follow their setting which enables our model to compose unseen class features during training, thus the model can learn the testing distributions with both seen and unseen classes (lines 70-72). Please notice that without the availability of unseen class semantics at training time, we cannot use self-composition to alternate between training a classifier and composing features.

– C2: "*how they obtained the results of Table 1 and Table 2 (left) for prior works?*" **Answer:** Thanks. We will include the following clarification: On DeepFashion, we run each baseline using their released codes with their default settings. On the remaining datasets, we use the performances reported in their papers to ensure their best performances.

– C3: "*Do all these works use a ResNet 101 backbone*" **Answer:** Except for `SMA` using VGG19 and `LFGAA` combining VGG19, GoogleNet, and ResNet101, all remaining baselines use ResNet101. We will clarify this in the paper.

– C4: "*does one really need to restrict the set of possible features to semantically-related samples?*" **Answer:** Please notice that the related sample set $\mathrm{Q}(\boldsymbol{z})$ is required for the sampling probability $p(\boldsymbol{H}|\boldsymbol{z})$ and is computed via nonnegative OMP (lines 171-173 and Algorithm 1). Without related samples, we cannot solve Eq (10) efficiently through sampling.

– C5: "*Lines 274-275 ... Can you please clarify?*" **Answer:** Please notice that we set the attribute values of each sample $\boldsymbol{z}^i$ to its class semantic $\boldsymbol{z}^{y_i}$ (lines 239-240). However, due to occlusion, a sample has many missing attributes compared to its $\boldsymbol{z}^{y_i}$, thus relying only on $\boldsymbol{z}^{y_i}$ without $\boldsymbol{H}_i$ would compose features lacking many discriminative attributes.

We hope our responses answer the questions and kindly ask the reviewers to raise their scores in light of our responses.