We thank all the reviewers for their valuable feedback.

Response to Reviewer 1 We thank the reviewer for the helpful feedback. We agree with the reviewer about the possible confusion regarding the term “saliency”, and will clarify this in the text. Thanks for the suggestion on communicating the results in Fig. 6; we will improve the visualization here. We will also correct all of the typos.

We agree with the reviewer that a good concept-based explanation method should be closely aligned with human common-sense. ACE is a step toward this goal, and it opens the door for many interesting follow up works. Regarding the choice of importance score, we chose TCAV score due to its simplicity and its compatibility with our goal. Given a few examples of a concept, TCAV returns its scalar importance value for the prediction of a target class. ACE can be combined with other global interpretation methods such as (Yeh et al. NeurIPS 2018) to further select 'good examples' to summarize the discovered concept, and this is a good direction of additional work.

We have added more details and discussions of the experiments based on the reviewer's feedback. First, the reviewer asked why we used 50 images for each class. As shown in the original TCAV paper, this importance score performs well given a few number of examples for each concept (10 to 20). In our experiments on ImageNet classes, 50 images was sufficient, possibly because the concepts are frequently present in these images. Second, the reviewer refers to the discussion of human experiment results. For each question (i.e. set of images) in the experiment, participants were asked to provide a one-word explanation of the shown concept. As a result, for each question, a set of words (e.g. `bike`, `wheel`, `motorbike`) are provided and we tally how many individuals use the same word to describe each set of image. We then find the most frequent word for each image set. Averaged across the 15 questions in the experiment, 56% of the participants used the same word to describe the image set, and 77% of the participants used one of the two most frequent words. Because many participants independently came up with the same word to describe a set of images, this strongly suggests that each set of images is capturing one coherent concept that's meaningful to the participants. It is also interesting to compute the similarity of individuals in answering the questionnaire. For every two participant in the experiment, on average they give the same answer to 5.8 out of the 15 questions. We will make this clear in the revised paper. We find the suggestion of replacing humans with a second step of ACE (or in general testing humans in the context of concept-based explanations) very interesting and we believe this suggestion could be a very informative project of its own.

Response to Reviewer 2 We thank reviewer for the thoughtful feedback and we are happy that you liked the paper! Our ACE framework is natural and easy to implement, and it opens the door for several interesting directions for future work in concept based explanations. In the revision, we have improved the discussion and provided more details on our experiments and validations. We will also release a practitioner-friendly implementation of ACE.

Response to Reviewer 3 Thank you for your helpful feedback. We agree with you that the goal of our work is in augmenting (and not replacing) human experts in explanation of the models. We will further emphasize this point in the text and will delete the sentence on line 47. We will also add discussions of connections to dictionary learning; thanks for this suggestion. A very interesting direction of future work here is to apply more sophisticated dictionary learning approaches, beyond clustering, on the representation space (e.g. sparse coding). This could potentially reduce the need of image segmentation, and learn more complex concepts. As mentioned in the text, we agree with you that rescaling the image patches could introduce noise in some settings. Previous work (Dabkowski et al, NeurIPS 2017) has found that ImageNet classifiers is relatively robust to changes in aspect ratio of segments; our experiments also verified that ACE is robust to aspect scaling. This is a good point to investigate further in follow up works.

We are glad that you found Figure 6 interesting; we will emphasize it more in the text. In this paper, we focused on visual concepts in image data as a good starting point, because it is relatively easy to illustrate the concepts in this setting. It'd be very interesting to expand the method to other modalities such as text; e.g. we could investigate having a cluster made of relevant words and phrases as a unit of explanation in NLP applications. This is an interesting direction of future work. We will make the discussion of the desiderata more concise and clearer (we will also move it to the Discussion section later in the paper to improve the flow).