We thank the reviewers for their constructive feedback and we address their concerns directly.

**Review 1.** We are going to incorporate your suggestions in the next version of the paper. Moreover:

1. Special features are used as an input to the CAE; while they do not have any imposed meaning, they provide information about the image to CAE and allow adapting the ICAE encoder to the needs of CAE. 2. Sparsity is imposed only on the CAE level. 3. Details on the solver are provided in the appendix. 4. Since different aspects of the model considered in the ablation study are not additive, providing a cumulative performance decrease of combined ablations would require rerunning experiments. We will consider adding these results to the next version of the paper.

**Review 2.** Our method is feed-forward without iterations, so it's O(1) in that sense; it's O($n^2$) in the number of capsules n, since the number of model parameters grows quadratically with n, similarly to the previous versions of capsule networks. We are going to release the code in due time.

**Review 3.** General comments:

1. We focus on modeling single objects in order to simplify and better understand the method. In future work, we will extend SCA with deeper hierarchies of transformations, where higher levels handle separate objects.

2. Regarding other metrics, we are going to run few-shot classification and meta-learning studies as future work. This will also validate whether our method learns in a more data-efficient manner compared to other methods.

3. While not included in the paper, our results show that SCA is not sensitive to the number of object or part capsules, and this holds for both constellation and image experiments. We will add this analysis.

4. We think that strong inductive biases are desirable (e.g. CNN). Mammals rely on them and they are a major motivation of this work. Data augmentation helps MI-based methods but it is statistically inefficient.

5. We removed the symbols for camera/part/object; we are unaware of references for coordinate frame usage in human vision that are more relevant. We removed the list of drawbacks of previous capsule architectures from the introduction.

**Section 2.1:**

6. Deformations are currently allowed and modelled as $CPR_{k,n} = CPR^b_{k,n} + f_n^{CPR}(c_k)$; the first term is a bias and the second term is input dependent. We add $\alpha||f_n^{CPR}(c_k)||_2^2$ to the final loss, which discourages input-dependent transforms.

7. Since deformations are penalized, the model tries to not use them. Explaining multiple objects with a single capsule would require severe deformations if the objects appear in different configurations.

8. line 93: the $\arg\max$ in Eq. 5 should be over both $k$ and $n$. Every object capsule predicts $N \neq M$ parts and we need to choose $M$ parts out of all $NK$ predictions. We only require that $M <= NK$.

**Section 2.2**:

9. We do use a spatial mixture model as in Greff et al. 2017, which does model pixels as *conditionally* independent. Many VAEs and all approaches that consider mean-squared error in the pixel space make the same assumption.

10. $f_c$ is linear; it was developed for MNIST and does not generalize to other datasets, which might also cause limited performance on CIFAR10. We will explain it in the paper and investigate better ways of specifying mixing probabilities.

11. We currently do not use discovered objects to refine parts. Instead, we rely on amortized inference, where the encoder can learn to mimic the EM procedure from previous capsules. In principle, we could run CAE iteratively, where at every iteration we encode parts and then reconstruct them. Our initial experiments resulted in unstable training, and we will investigate this in future work.

**Section 2.3**:

12. We do use the true number of classes as a hyperparameter for the sparsity regularizer, but this value need not be known; it can be fitted on the validation set instead. It is a good idea to encourage just one active object capsule per input image and we will try it. The notion of uniformly distributed classes across training data motivated one regularizer, but it is not strictly necessary. We will reformulate this in the paper.

13. The additional cost function in the constellation experiment will be removed as the same regularizers as used in image experiments are sufficient to obtain the reported performance.

**Section 3.1**: 14. we note that our model with simpler decoder is similar to Greff et al. 2017; 2019, but where inference is amortized instead of performing EM as in Greff et al. 2017 or gradient ascent in the latent space as in Greff et al. 2019. What makes our model different is the explicit geometrical structure of the decoder.

**Section 3.2**:

15. LIN-MATCH is used in the ablation study. In retrospect, we feel that only LIN-MATCH and LIN-PRED provide useful information. The other two will be removed from the paper.

16. It is fair to conclude that SCA failed on the Cifar10 experiment. We just developed a mixture of a capsule model with a background model tto deal with clutter and will update the paper.

**Section 4**:

17. We will add the drawbacks of having no routing to this section. We note that iterative refinement of part-capsule assignments is possible with our approach, but has not been explored yet.

18. While it is true that MONet or IODINE would discover parts if applied to single-object data, there is no clear way to impose hierarchy. In our work, it is possible to stack multiple levels of object capsules, in which case we can have more levels of decomposition. This is an area of future work. We will reformulate this paragraph.

19. We are going to make this section more focused. We note that citing technical reports published on arxiv is standard practice.