We thank all reviewers for their valuable and constructive feedback. The consensus appears to be that the paper is well written, well motivated and easy to follow (R1, R2). We also appreciate the assessment of our work as *"innovative and unusual in the field"* (R1) and *"a very valuable contribution"* (R2).

Questions and Comments of Reviewer 1

- **Resolution of the "3Dprinted-realistic" dataset.** Re-rendering and releasing higher resolution images is an important future work we plan to address. It is however beyond the scope of this paper due to the significant engineering challenges it entails. To illustrate the point, the rendering of 1m high resolution (512x512) images for the '3Dprinted-realistic' image dataset, would roughly take 140 days of continuous computation using our existing infrastructure. Moreover, SOTA disentanglement learning algorithms [37] (Locatello et al., ICML 2019) still use 64x64 resolution, such that already the provided 256x256 resolution is a significant improvement and for a resolution of 512x512 we directly provide the real-world dataset.
- **Downsampling of real-world images for transfer learning.** Thanks for pointing this out. We will make clear in the final version that in all experiments we used images with resolution 64x64. This resolution is used in recent large-scale evaluations and by SOTA disentanglement learning algorithms [37](Locatello et al., ICML 2019).
- **Link to code or the dataset.** Unfortunately, the NeurIPS' policy does not allow to provide any external links in rebuttal documents. However, we guarantee to make all datasets and trained models publicly available before publication.

Questions and Comments of Reviwer 2

- ***"Is disentanglement helpful in general? and can authors comment on tasks associated with real-world disentanglement?"*** Thanks for pointing this out. We agree that currently there are multiple research directions for disentangled representation learning. One of them is the suggested investigation of the usefulness of disentangled representations. Our work is encouraging that and allowing to investigate the effectiveness of disentangled representations with access to ground truth labels on real-world data.
  Prior work e.g. [20] (Higgins et al. (ICML 2017) already indicated that disentanglement is useful for reinforcement learning. We believe that applying these ideas on the real-world platform is an interesting direction and we will update the future work section.
  As pointed out by the reviewer, the effort to find a thorough answer to these questions is significantly beyond the scope of one paper. However, we are already looking into replicating the analyses of the provided reference (van Steenkiste et al., arxiv 2019)[1] and a new fairness approach (Locatello et al. arxiv 2019)[2], both illustrating the benefits of disentanglement, on our real-world data set.
- ***"The authors state that the training on low resolution images results in the instability therefore random seed and hyper-parameters being more important than the model. Could the authors clarify this?"*** Our intention was not to suggest that this instability is necessarily due to the low resolution of those datasets. We agree that this is slightly misleading and we will rephrase this sentence in the final version.
- **Used architecture.**
  All the methods use the same convolutional encoder and decoder where the latent dimension is fixed to be 10.
  $Encoder : input(64 \times 64 \times 3) \rightarrow 4 \times (4 \times 4conv, 32ReLU, stride2) \rightarrow FC1(256) \rightarrow FC2(10)$
  $Decoder : 10 \rightarrow FC1(256), ReLU \rightarrow FC2(4 \times 4 \times 64), ReLU \rightarrow 3 \times (4 \times 4upconv, 64ReLU, stride2) \rightarrow 4 \times 4upconv, 3, stride2$
  This can be considered the standard architecture and is used in e.g. [4,11,19,38]. We will make sure to add this information in the final paper.
- ***"Could the authors maybe comment on the difficulty of disentanglement (or reconstruction) for each individual feature? (which features are harder to disentangle?) I would be also interested to see which features cause the most difference in image space."*** Empirically, we observed that some of the best trained models were able to disentangle, though imperfectly, the factors of camera height, background colors, object sizes and the motions along the first and second degrees of freedom. In contrast, they performed poorly in disentangling the factors of some object shapes (for e.g. pyramid and cone) and some colors (for e.g. olive and brown). This may well be because of less pixel variation in the respective factors. The features of camera height and background color cause the most difference (maximum L2 distance) in image space. Similarly the object positions at different joint configurations (not the consecutive frames) also have big L2 distance which may explain why the models focus more on learning these factors. We will add a detailed analysis to the final version.

[1] van Steenkiste, S., Locatello, F., Schmidhuber, J., and Bachem, O. Are disentangled representations helpful for abstract visual reasoning? arXiv preprint arXiv:1905.12506, 2019

[2] Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B. and Bachem, O. On the Fairness of Disentangled Representations. arXiv preprint arXiv:1905.13662. 2019.