# NeurIPS Rebuttal for "Explicit Disentanglement of Appearance and Perspective in Generative Models"

We thank the reviewers for their constructive and fair reviews. We here address their key concerns. Our main contribution to NeurIPS is the first practical realization of the theory from Higgins et al. (2017), which argue that disentanglement need to appear from *group actions*. Our work is, thus, an important first-step towards bridging theory and practice. All comments regarding text+figure updates are appreciated and have been added to the paper.

**Experiments(R1+R2+R5):** As requested by multiple reviewers, we empirically evaluated our proposed VITAE model on more datasets. Specifically, the paper now include performance on the dSprites (Higgins et al., 2017) dataset (quantitative evaluation, see Table 1) and qualitative disentanglement of facial attributes on CelebA dataset. Figure 1 illustrates how our model is able to capture geometrical face information such as the shape of the face. We found similarly that we could also capture head pose and viewing angle.

| | $D_{score}$ |
|---|---|
| VAE | 0.05 |
| $\beta$-VAE | 0.18 |
| $\beta$-TCVAE | 0.30 |
| DIP-VAE-II | 0.12 |
| C-VITAE | **0.38** |

Table 1: dSprites

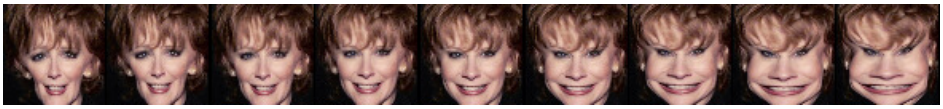

Figure 1: CelebA

**Encoder network (R2+R5):** There is concern about the importance of making the appearance encoder dependent on the perspective through the inverse transformation. We stress that this choice is crucial, and constitutes a notable part of our contribution. Empirically we find that C-VITAE (with encoder-transformer) outperforms U-VITAE (without encoder-transformer) in all tests, but did only include the results in the case of MNIST. As R5 notes, this conditioning minimizes the mutual information between $z_A$ and $z_P$, leading to better disentanglement of the two latent spaces. Thus, there is no need to directly discourage the perspective factors from predicting the appearance factors by for example including an auxiliary loss term. This is rather elegant. Lastly, by forcing the encoder-transformer to be the inverse of the decoder-transformer we mimic the behaviour of normalizing flow models, which have better inference properties.

**R1:** We agree that measuring "disentanglement" is an unsolved problem, but Locatello et al. (2019, ICML best paper) find that most proposed disentanglement metrics, including the disentanglement score from Eastwood et al. (2018) that we are using, are highly correlated. Hence, our results should still hold under other established metrics. We focus on the disentanglement score as it is the only one that seems to be of general interest in the field.

While the underlying generative factors are not known for MNIST in the sense as for synthetic data, we do not agree that the label of an image cannot be interpreted as a generative factor. If we ever are to move away from only quantitatively evaluating disentanglement on synthetic data, we need to consider which generative factors real life data could have been generated from. We agree that the generative factors for SMPL where made discrete on purpose, and we have adjusted this in the revised paper. This discretization does not invalidate our results, since the general framework from Eastwood et al. (2018) still holds. Secondly, when Eastwood et al. (2018) state that "discretization is unnecessary" they comment on the discretization of latent factors and not on discretization of generative factors, as in our case. Regarding significance, we re-ran all algorithms three times, and can conclude the results indeed are significant.

**R2:** Indeed, it should be possible to exchange the VAE framework with a WAE, by essentially changing the loss function. We do not expect better disentanglement, but rather improved reconstructions as in the original WAE article.

We agree that the work of Lorenz et al. (2019) and Xing et al. (2019) are related to ours in that they rely on spatial transformations to disentangle data (note: both articles were published at CVPR 2019 after the NeurIPS deadline). That said, the theory of Higgins et al., strongly relies on disentanglement through *group actions*. This places hard constraints on which spatial transformations are allowed: *they have to form a smooth group*. Both TPS-transformation (Lorenz et al.) and displacement fields (Xing et al.) are not invertible and hence do not correspond to proper group actions. Our work, thus, remains unique as the first realization of the theory from Higgins et al. (2017). As a practical benefit, we have also shown that the invertibility of the group action lead to efficient inference through the C-VITAE architecture.

**R4:** We thank you for the positive review. We agree that even if our approach is novel, it can always be improved upon. We want to stress that the results on the SMPL dataset actually take advantage of the CPAB transformations. We made this more clear in the final version.

**R5:** We do not agree that our SMPL dataset is toy-like, which should be evident from the fact that even state-of-the-art models seem to struggle with disentanglement of the generative factors. Additionally, they are in the same level of complexity as the dSprites, Cars3D, SmallNORB and Shapes3D datasets that are currently used for disentanglement evaluation, see Locatello et al. (2019). We choose explicitly to work on datasets where we have access to generative factors, such that our results could be quantitatively evaluated. This however heavily limits the datasets to be either synthetic or very simple in complexity.