Deep Transformation-Invariant Clustering

We thank the reviewers (Rs) for their positive feedback. If accepted, we will incorporate all feedback in the final version.

Lack of details on transformation parameters (R1). We believe this information is already in the paper and supplementary material. As stated L160-162, we follow STN [34] to model the spatial transformations, i.e. we respectively model affine, projective and TPS transformations with 6, 8 and 16 (a 4x4 grid of control points) parameters. We will make this more explicit. As explained L163-168, the color transformation is modeled by an affine transformation with 2, 6 or 12 parameters depending on the scenario. The morphological module has 50 parameters (a, a 7×7 image and a real α) as discussed L182-187. The exact sequence of transformations used for each experiment is specified in supplementary material Table 1. To ensure complete reproducibility, we will release code, data and models.

Initialization (R1, R4). Our results indeed depend on initialization, we will make this more explicit. We provided an analysis over 100 runs for MNIST in the supplementary material and report standard deviations for the experiments of Table 3. As suggested, we will add median and standard deviations for all tables. Note that, as discussed in the supplementary material L19-24, our loss correlates with performances and provides a way to select a good initialization.

Small improvement over SotA (R4). Although we do not report large quantitative improvements, our approach is very different from SotA methods, which we think is highly valuable. It has other significant benefits: (i) it doesn't rely on any hyper-parameter; (ii) the objective loss is simple to formulate and doesn't require any additional ad hoc losses or regularization terms; (iii) results are visually interpretable. Additionally, our method led to strong qualitative results on internet photo collections which, to the best of our knowledge, has not been reported by other approaches.

Effect of the number of clusters K (R2). Similar to many clustering methods, the selection of the number of clusters is indeed a challenge. A purely quantitative analysis could be applied to select K, e.g. in Figure 1 we plot the average loss for DTI K-means as a function of the number of clusters and it is clear an elbow method could be applied to select 10 clusters. We believe our method also has the advantage to provide interpretable prototypes which we used to approximately select the number of clusters on the internet photo collections. We did not find the qualitative results on this data to be very sensitive to this choice, we will add loss plots, qualitative examples and a discussion.

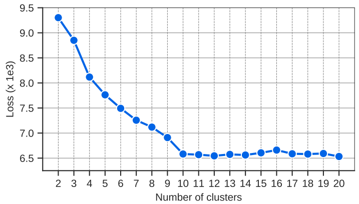

Figure 1: Loss w.r.t nb of clusters for MNIST-test (avg over 5 runs)

Equivariant models (R2). They are indeed relevant, we will add a discussion in related work: our DTI framework learns to predict transformations of the prototypes in an equivariant way. We did not find any specific equivariant clustering method we could compare to, but would be happy to include comparison to any specific work. Note that equivariant models typically aim at learning equivariant representations, whereas our method aims at being invariant to transformations but not at learning a representation.

Applications to other types of data (R3). We will clarify the last sentence of the conclusion. To demonstrate our framework's genericity, we designed a 3D affine module (12 parameters) and applied our framework to cluster the 3D point clouds from ModelNet10 [1] (Table 1, results over 5 runs). Such a simple module already provides a significant boost over standard K-means.

Table 1: Cluster acc. on 3D shapes

Method	avg	median	max	std
K-means	72.1	73.7	74.1	2.4
DTI K-means	**83.2**	**83.4**	**85.1**	1.3

Other comments (R1, R2).

- *Hard GMM optimization (R1).* Our M-step is indeed more difficult than in standard GMM. We mentioned the training details used to make it work in Section 4.2 and in the supplementary material (Section D). We believe our experiments demonstrate the viability of our optimization.
- *Redundant transformations (R1).* Because transformations are all initialized with identity functions, adding redundant transformations in our curriculum learning process doesn't change the results. We experimentally validated it on MNIST using DTI K-means with 3 affine modules and the loss is not impacted when adding the duplicated modules.
- *Sample transformation (R2).* As discussed in the supplementary material (L51-57), a trivial solution if learning sample transformation instead of prototype one is to learn "empty" prototypes and transform samples into "empty" images. For MNIST, we rapidly observed black prototypes and any sample was transformed into a black image.
- *MinLoss criterion (R2).* We experimentally found that our loss is highly informative leading to a criterion (*minLoss*) that enables us to automatically select a high performances run in a fully unsupervised way (Section B of supplementary material). This is not trivial and to the best of our knowledge, such criterion has not been emphasized by other approaches like DEPICT or DSCDAN. Therefore, we argue performances obtained with *minLoss* can be compared with average results from competing methods which don't provide such criterion. We will clarify the claim.
- *Notations (R2).* We intentionally changed prototypes notations for the general framework (c_k) and its application to K-means (m_k) and GMM (μ_k and Σ_k). We will clarify.

[1] Z. Wu, S. Song, A. Khosla, L. Zhang, X. Tang, and J. Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *CVPR*, 2015.