CaSPR: Learning Canonical Spatiotemporal Point Cloud Representations

We thank the reviewers for their comments. We appreciate that they find our work *solid* (**R2**) and *completely finished* (**R3**), our formulation *elegant* (**R1**), the CaSPR method *novel* (**All**) and *well-designed* (**R1**), the demonstrated applications *promising* (**R4**), and the paper *well-written* (**All**). In this rebuttal, we address the limitations and specific questions raised by reviewers.

Known Category Assumption (R2, R4): *Training needs to be performed for the individual category which is not practical in the real world. (R2)* Indeed, for best performance CaSPR should be trained on individual categories. However, Section 2.4 of the supplement demonstrates that a single model trained on all three shape categories still gives superior results to baselines in many cases. Additionally, category-specific models can have important real world utility, e.g., in self-driving vehicles where high accuracy is critical but only for a handful of categories like cars, bicycles, and pedestrians. *How does CaSPR generalize to unseen categories? (R4)* We agree that generalization to unseen categories is interesting. Yet, this is a formidable open problem in computer vision and ML beyond the scope of our work. In CaSPR, we focus on many other problems of importance by leveraging a category-level prior on object shape.

Novelty of Individual Components (R4): *Correspondences with NOCS and spacetime neural ODEs are studied previously, the idea of combining them is novel...The idea is not surprisingly new.* We emphasize that the novelty of our method is in the design and execution of its individual components – CaSPR is **not** simply a trivial combination of previous work. NOCS has been previously used to establish correspondences, but not in a temporal setting or by explicitly accounting for and normalizing time. Neural ODEs have been used to model point cloud time series in physical space (Occupancy Flow), but CaSPR does so in a learned latent space which enables more accurate temporally continuous reconstruction (see Table 4) and interesting properties like approximate latent disentanglement of shape and motion (see supplement Section 2.6). Finally, another novel contribution is TPointNet++, which facilitates operating on dynamic point cloud sequences through spatial *and* temporal canonicalization within the overall processing pipeline. Prior spatiotemporal (Occupancy Flow) and point cloud reconstruction (PointFlow) methods lack this step. We are a little confused why **R4** says the combination of these novel components is “not surprisingly new" since we are not aware of prior work that learns a spatiotemporally continuous latent representation in this way.

Canonicalization Supervision (R1, R2): *The method does require supervised NOCS which can be difficult to acquire. (R1) Training requires the ground-truth for supervision, however, in practice it is difficult to obtain...is it possible to use some unsupervised criterion? (R2)* As aptly noted by **R1**, moving away from supervision in the context of NOCS was out of the scope of the current work and will be a substantial undertaking in its own right. As hinted by **R2**, there are several promising avenues that may allow the use of NOCS in real-world data. Recent work on auto-labeling NOCS in autonomous vehicle data[1] could enable direct supervised training. Alternatively, weakly-supervised approaches[2] allow learning a canonicalization with easier annotations like 2D keypoints. We will explore these in future work.

Object vs. Scene-Level (R1, R2): *The method is object centric. (R1) This proposal is based on object-level sequences, it will be beneficial to generalize to scene-level. (R2)* We agree that supporting scene-level processing is a great addition and very desirable in practice; it will be a main focus of future work. Note that with recent successes in scene-level 3D segmentation and object detection, CaSPR can be applied in its current form by first localizing objects of interest.

Move Related Work to Main Paper (R3): We fully agree and will integrate the extended literature overview from the supplement into the main paper for the camera-ready version.

“Adopting” CaSPR to Applications (R3): *Give more details on how CaSPR is adopted to different applications.* CaSPR is a very flexible and general framework that does not require adoption to different applications: a single trained model on a category can be used for every presented application. We will make this point clearer in the paper.

Specific Questions (R2): **(i)** *Why is the latent splitting reasonable? There is no mechanism that can guarantee such splitting.* We agree that CaSPR cannot guarantee a theoretical disentanglement; instead, our design encourages the canonicalization network to respect this split by only advecting the dynamic feature with the ODE. Note that this is not a CaSPR-specific drawback and many SoTA disentanglement networks rely upon the same intuition. We demonstrate experimentally that disentanglement is achieved to a large extent (see supplementary Section 2.6 for non-rigid case and video for rigid). **(ii)** *How can CaSPR guarantee it generates scene rather than deformation flow?* Since temporal correspondences naturally emerge from the CNF by using the same Gaussian noise at each timestep, there is no theoretical guarantee of scene flow. Despite this, CaSPR maintains accurate correspondences for non-rigid deformation (Table 4). If a discrepency between scene and deformation flow were to appear, a simple post-processing step such as optimal transport between generated deformations could provide temporal correspondences/scene flow. **(iii)** *Does Gaussian noise need to be the same to generate temporal correspondences?* Yes, see above.

[1] *Autolabeling 3D Objects with Differentiable Rendering of SDF Shape Priors*, Zakharov *et al.*, CVPR, 2020

[2] *C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion*, Novotny *et al.*, ICCV, 2019