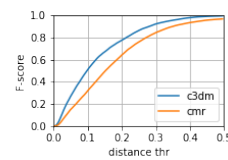
Why C3DM is more than CMR/CSM without a mesh. Our C3DM representation is *not* a mere drop-in replacement for the meshes in CMR/CSM. C3DM has major advantages: beyond removing the complexity of differentiable rendering and re-projecting to a mesh, a key one is that C3DM losses leverage appearance cues (RBG values) to learn the 3D geometry, while CMR/CSM do not. This may look surprising given that CMR does extract a texture model from the RGB values, but only silhouette and keypoint supervision affect the geometry (see top of page 9 in [29]). Attempting to jointly learn generative models for 3D shape and texture is a recipe for failure because such combination has too many DoFs. Because C3DM generalizes unsupervised monocular depth estimation, we can instead borrow re-projection losses (e.g. min-k) and use correspondences to constrain the geometry *regardless* of the texture model's quality. Note that without those appearance cues (in addition to keypoints), CMR fails to reconstruct faces, which C3DM masters.

Reviewer 1. **Texture transfer not evaluated. Evaluate on keypoint transfer.** Our main contribution is improving *3D reconstruction* of object categories via a new canonical representation of shape. We use texture transfer as means to demonstrate the consistency of this canonical map across instances. As suggested, we will also report keypoint transfer; on CUB, we improve PCK@0.1 drastically: 0.85 vs. 0.48 (CSM) and 0.47 (CMR). Note though that we use keypoint annotations during training, so the canonical map quality is expected to be better at keypoint locations than between them. **Cite Kulkarni et al.** OK. **Train and test with automatically detected keypoints.** We *do* use automatically detected masks/keypoints for training/testing in *all* cases where possible: Freiburg Cars and FlorenceFace (Appendix E). For P3D and CUB birds, there is no other dataset to train the keypoint detector, so we use GT annotations for training. **Try sinusoidal embedding for B, C.** Thank you; we are planning to experiment with spherical harmonics in the future. **Report F-score.** We will add the plots to the final version. Results on FreiCars are in the figure to the right. Consistent with Chamfer distance, C3DM outperforms CMR on all thresholds.



Reviewer 2. **How does [your] method improve over meshes?** Our representation is not a mere drop-in replacement for CMR's meshes. Specifically, C3DM innovatively bypasses the complexities of CMR/CSM. It provides a better performing alternative to the widespread mesh rendering paradigm. Said that, we can indeed convert our representation to a mesh by warping an icosphere vertices with eq. (1). When done after training, on FreiCars, it increases d_{pcl} from 0.13 to 0.18 due to finite mesh resolution. If used as a representation during training, swapping $\mathcal{L}_{\text{repro}}$ and $\mathcal{L}^{\text{min-k}}_{\text{percep}}$ with CSM's cycle consistency loss through the mesh further increases d_{pcl} to 0.31, even worse than C3DM without $\mathcal{L}_{\text{repro}}$! We conclude that enforcing cycle concistency through mesh is not adequate for our setting. **CMR fails on faces. How was it initialized?** For fairness, we did not apply any dataset-specific initialization to any of the benchmarked methods. CMR fails on faces because it relies on silhouette loss, which is insufficient for learning detailed facial geometry. **Evaluate mask IOU and PCK metric?** Please refer to the answer to R1 for PCK on CUB. Note that the IOU/PCK metrics are 2D and do not evaluate 3D reconstruction, e.g. flat 3D shapes with matching deformation/viewpoint can satisfy them. CMR has to use IOU/PCK because CUB lacks 3D annotations. Our evaluation on the datasets with 3D ground truth (Freiburg Cars, Florence Face) is thus an improvement over CMR's evaluation on CUB. **Similar to Atlasnet-sphere.** Will cite; indeed, C3DM canonical map is similar to Atlasnet-sphere, but, crucially, the rest of the pipeline, including handling 3D deformations, focus on real image data and weak supervision, are *significantly different*. **The explicit basis is not clearly motivated.** We believe that our continuous extension of the sparse NRSfM basis is novel and appropriately motivates the explicit basis. Other works, including CMR, only re-use the camera parameters from [NR]SfM, while we also exploit the deformation basis. **Adhoc losses: min-k, $L_{\text{emb-align}}$ not used in CMR.** As empirically proven in Tab. 1, those losses are crucial for achieving SoTA. We disagree that they are ad-hoc: as noted above, our representation is very different from CMR's meshes, motivating the different losses: The min-k loss densifies the supervisory signal in landmark-less areas, while $L_{\text{emb-align}}$ fixes the coordinate distribution on the sphere.

Reviewer 3. **Novelty: building on CMR.** We solve a similar problem, but everything else is rather different from CMR, including representation and loss functions. **Why is the model non-rigid [but] … rigid objects?** See lines 22–23: Since we model a class of objects, even if its instances are rigid, we still need to account for the *deformations between instances* (e.g. birds deform to starling or seagull). Prior work [2,29,43,12,34,41,62] also tests the algorithms by modelling deformations between different instances. **Combination of too many losses. Not a major concern if authors apply to non-rigid objects.** CMR uses 8 loss terms in total, more than C3DM. We outperform CMR on all datasets they use, and additionally on Freiburg Cars and FlorenceFace (all of them have non-rigid deformations). We also demonstrate that all loss terms are crucial for good reconstruction in Tab. 1. **Handling view-dependent effects with perceptual loss is unsatisfying. Use viewpoints explicitly?** In fact, C3DM *explicitly models* view-dependent effects in the top-k loss by comparing the reference image with a *warp* of the K target images produced with predicted viewpoints. The top-k selection is instead needed to mitigate effects of *self-occlusion* (l. 182). **Limitation: relies on successful NRSfM initialization.** NRSfM is used as initialization in most related methods [12,29,62]; CMR [12], in particular, uses old rigid SfM. We don't see it as a limitation given that NRSfM from keypoints is a much easier problem: when it fails, dense reconstruction is probably impossible. Furthermore, NRSfM supervision is injected in a *soft* manner in (1), so can be corrected. **Why is depth prediction of a CNN considered non-parametric?** We define non-parametric depth estimation in l. 36 and on. This is in contrast to CMR and others predicting the whole shape.