Scene Representation Networks (SRNs): Continuous 3D-Structure-Aware Neural Scene Representations

We are glad that the reviewers found SRNs to "represent an important step forward" (R3), be "very interpretable" (R2), and that they will "open the door to more complex models combining 3D processing and deep-learning" (R1). We thank the reviewers for their detailed, constructive feedback, which we incorporate as follows.

Probabilistic formulation & uncertainty (R1, R2) Our intention was to state that it could *in principle* be possible to embed SRNs in a probabilistic framework. *We will clarify & soften the claim.* To encourage future work, we will formalize an instantiation similar to "Consistent Jumpy Predictions for Videos and Scenes" (Kumar et al. 2018, follow-up work to Eslami et al. 2018) in the supplement, *stating that this has not been experimentally verified.* High-level idea: images & pose observations are encoded into a code vector $\mathbf{r}$, $\mathbf{r}$ is used to parameterize a prior distribution over latent variables $\mathbf{z}$, and sampled latent variables $\mathbf{z}$ are decoded into a scene representation Φ by a hypernetwork.

Complex scenes & compositionality (R1, R2, R3) We'd like to disentangle SRNs from the notion of generalizing SRNs. **(1)** In Sections 3.1 to 3.2.2, we formalize an SRN as a *single* function Φ, representing a *single* scene. The minecraft room at the end of the video demonstrates that this may represent challenging scenes. We are happy to add more such examples. **(2)** For generalization, we demonstrate that the space spanned by SRNs allows learning strong shape & appearance priors. We demonstrate this using hypernetworks, *assuming that scenes lie in a low-dimensional subspace (Sec. 3.3.).* It is an open research question if this assumption holds for complex 3D scenes. We thus focus generalized SRNs on single-object scenes. Other approaches to generalization, such as Model-Agnostic Meta-Learning (Finn et al. 2018) may relax this assumption. Intuitively, it may be possible to copy and compose learned representations of objects & primitives in later layers of an SRN by "re-wiring" earlier layers. This is an interesting avenue of future work, outside the scope of this manuscript. We will clarify that generalization via hypernetworks is only valid if the assumption in Sec. 3.3. holds, which we only demonstrate for single-object scenes, and discuss alternatives.

dGQN solves harder task (R2) We will clarify that the dGQN solves a more difficult problem, and that the auto-decoder framework requires optimization to infer a scene representation.

Required camera poses (R1, R2) We will clarify that camera poses and intrinsic parameters are required, and will clarify abstract, lines 11 and 71 to point out that poses are a form of geometric supervision.

Absence of camera poses (R1) Sparse bundle-adjustment provides fast pose & intrinsics estimation. Recent work also formulates pose estimation in a learning framework (Ba-net, Tang et al., 2018). As SRNs are differentiable w.r.t. to camera poses, they may be integrated with any such algorithm. We will add a discussion and references to Sec. 5.

Metrics (R1, R3) We will add forward pass duration ($\approx$ 120ms), training memory requirements ($\approx$ 3GB per batch item), and training time ($\approx$ 6 days for chairs, cars) in Sec. 4 (numbers for resolution 128×128, 10 raymarching steps).

View-dependent effects, transparency (R2) SRNs may be extended as follows: specular highlights can be addressed by supplying view direction to the renderer, transparencies by accumulating features along each ray, reflections by introducing secondary rays. We will discuss this in Sec. 5 and add an extended discussion to the supplement.

Limitations of pixel-independent decoding (R2) There are two key limitations: (1) 2D CNNs perform well in generating high-frequency patterns. The current pixel generator cannot exploit this strength, but guarantees view consistency. (2) A per-point formulation requires the LSTM, renderer and SRN to propagate features proportional to the number of pixels, which is expensive. We will add this discussion of limitations to Sec. 5.

Applications outside of vision, broader context (R2) We think that SRNs have high potential for applications outside of vision, such as robotics, physics modeling, and even medical imaging. In this manuscript, we chose to investigate a single application in order to fully explore their fundamental properties. We will highlight this as part of future work in this emerging area. We will rephrase discussion, abstract and introduction to contextualize this work with more applications outside of vision.

Raymarching vs. other renderers (R3) We will add a brief discussion of other rendering techniques to Sec. 3.2.1, alluding to an in-depth discussion in the supplement.

Full-model figure (R3) We will rework Fig. 1 to make it more legible, more clearly label the three key aspects of the model, and make their interaction more comprehensive. We will add an additional figure to the supplement that illustrates SRNs without spatial constraints.

Re-work abstract, introduction (R3) We will rephrase lines 1-5 and lines 17-21 to provide a more contextualized introduction to the problem of scene representations.

Expanding discussion / implementation details (R1, R3) We agree that many details from the supplement would significantly add to the main paper. *We will follow all such requests made in the summary of review.* However, the manuscript is at the page limit. Adding anything will require moving something else into the supplement. We believe that all current content is important to communicate key aspects of SRNs.