We thank the reviewers for their time and the overall positive feedback. We address each reviewer separately:

• **R1:** "the proposed representation can only capture fixed topology [sphere] and struggles at [..] high frequency details. [..] would be interesting to see which improvement could be achieved by using a stronger prior on geometry (as suggested by the authors [..], enforcing a stronger prior on geometry through synthetic datasets would be helpful [..])" This is a valid point. Although our method can work with other topologies as long as they are fixed, dynamically varying the topology is challenging. Nonetheless, we believe our approach represents a first step in the right direction and the framework is flexible enough to support extensions (e.g. semi-supervision), as both R1 and we observe. "how the simple orthographic projection models impacts the alignment of geometry and texture " We partly discuss this in the appendix A.2, 3rd paragraph. On Pascal3D+ cars, we add a learned perspective correction term to address foreshortening (unnecessary on CUB as objects are distant). Geometry/texture are always aligned since the initial reconstruction model optimizes for this task, but inaccurate camera assumptions might lead to distorted meshes. We will discuss this more in depth in the revision. We also point out that our method can work with *any* projection model if the camera parameters are known or can be estimated.

• **R2:** "comparisons [..] 1) generating 3D voxels with color channels using 3D CNNs; 2) [..] branch to infer the color for each vertex." We would first like to emphasize that voxels do not support texture mapping, which is a key aspect of our work, and which requires a triangle mesh setting. Textured triangle meshes are widespread in graphics (e.g. movies, games) and are crucial for achieving high detail, since simple meshes can be coupled with high-res textures; Additionally, *(i)* textures are more general than vertex colors (in our method, a 32×32 texture is equivalent to vertex colors); *(ii)* while it is easy to implement the proposed baselines using 3D supervision from synthetic data, it seems unclear how to translate them to a GAN framework using solely 2D supervision (our setting). Some voxel approaches use 2D supervision, but they only focus on geometry (no texture/colors) or reconstruction (not generation). *(iii)* In voxels, shape and color are not disentangled, which limits their generalization and flexibility in ways that are not captured by quantitative metrics (e.g. texture transfer is not possible). "shapes with complex topologies [..] complicated objects like chairs." We agree this is challenging. One potential direction is to use semi-supervision, as we observe in the conclusion (also mentioned by **R1**). Nonetheless, we observe that our model can at least handle some complex features (e.g. beaks, tails and wings in CUB).

• **R3** does not observe any major issue ("I cannot find any significant limitation of the paper") and appreciates our design choices ("I like the rationales behind choosing UV-sphere for mesh representation [..] well motivated and intuitive", "I like the usage of positional encoding which yields higher FID scores as shown in the ablations", "it is a very good paper and will definitely justify acceptance to NeurIPS"). "Some remarks [Prior work]" We thank R3 for those references and we will discuss them in the revision. "Typos" Thank you, we will take care of that in the revision.

• **R4:** "While I find the combination of single-image 3D reconstruction and GAN interesting, I am concerned about the technical contribution" We thank R4 for their constructive critical feedback. We think however that the raised concerns are not directed at our core contributions, which include *(i)* a novel convolutional mesh representation whose strengths include smoothness, semantic alignment, and the ability to be modeled using existing 2D convolutional GAN strategies. Thanks to the latter, our method can be easily adopted by the community and can benefit from future advances in the field. *(ii)* A way to learn textures directly from image pixels using an inverse rendering approach and masking. *(iii)* A *full* framework for textured mesh generation with supervision from natural images, whereas prior work has focused on more limited settings. Our contributions are also acknowledged by **R1** ("correct, novel, and interesting", "interesting for the community at large", "relevant for future work"), **R2** ("work is a pioneer attempt", "network architecture is novel", "results [..] are obviously better than previous works", "approach well advances such a emerging and timely research topic"), and **R3** ("approach is novel and makes sense", "great interest to the community", "convolutional mesh representation is novel and intuitive"). We would be grateful if R4 could reconsider their position in light of our response. "**[1/4]** Another solution to the proposed task [..] first training a 2D GAN to generate new 2D images [..] then directly run the single-image reconstruction network such as [24]" This is an interesting baseline, although we find it has some limitations compared to our native 3D representation (see [2/4]). First, we investigated it empirically: *(i)* running [24] on CUB *training images* achieves a FID of 85.8 on reconstructions, which is already worse than our scores and sets a lower bound; *(ii)* when running [24] on CUB images produced by StyleGAN, the FID further degrades to 101.9. The setting is the same as ours (no background). "**[2/4]** why the proposed framework would outperform this baseline" The baseline would not have the properties of a true 3D representation, e.g. pose disentangled from shape. The reconstruction model would perform suboptimally on occlusions (the model can only reliably infer information visible in the 2D image), and interpolating in the latent space of the 2D GAN would affect the 3D result unevenly. Contrarily, with our pose-independent UV representation, textures/vertices are semantically aligned across different images (i.e. parts such as wheels always show up in the same position), which greatly facilitates learning. "**[3/4]** the performance of the proposed method is also bounded by the [..] single-image 3D reconstruction network." Our reconstruction network is only used for the geometry, but textures are learned directly from images in a pure GAN setting (while reconstruction methods often use a VGG loss to avoid blurry textures). "**[4/4]** this alternative would be more flexible [..] since you can use arbitrary GAN to generate 2D images without re-training the reconstruction network." For best results with the proposed alternative, the reconstruction model would still need to be retrained on the same dataset. "Why [..] sinusoidal encoding [..]? [..] directly using (u, v) coordinates" Our UV map has circular boundary conditions, so concatenating sin,cos(π(u,v)) is more principled as it smoothly wraps around the edges. This is however a minor technical point and concatenating plain (u,v) coordinates might also work.