We thank the reviewers for their constructive comments. We appreciate that the reviewers find our contribution novel and timely (R4), our architecture clever (R1) and successful (R2), and our results impressive (R3). R1 and R3 highlight our systematic evaluation and all reviewers agree that the paper is clear and well-written and that the claims and method are correct. In the following, we address the comments of the reviewers.

Reviewer 1

Comparison to SRN: SRN is a method for novel view synthesis of given objects while our model allows to generate novel objects. The latter is not possible for SRN: SRN optimizes the latent code for a particular input image, but does not provide a full probabilistic generative model for drawing unconditional random samples. It is non-trivial to extend SRN to a full probabilistic generative model in order to conduct a comparison to GRAF. Moreover, SRN requires posed images for training which are not available in our setting. We will clarify these differences in the final version.

Broader Impact: We thank the reviewer for the suggestion. We will discuss potential dangers of 3D-aware generative models that can create 3D-consistent fake images. Such models might increase credibility of fake contents and potentially fool systems that rely on multi-view consistency, e.g., modern face recognition systems.

Hybrid representations: We thank the reviewer for pointing out additional related work. We will extend our related work section accordingly. While these works require 3D input and do not consider texture, extending our work to hybrid representations is indeed an interesting avenue for future research.

Reviewer 2

Comparison to Texture Fields: Texture Fields map a *3D surface point* to a color value. Radiance fields take a 3D point and a viewing direction as input and predict a color and a volume density for *any 3D point in space*. Importantly, Texture Fields require a 3D shape as input (main paper, l.64) and colored surface points as supervision, while GRAF learns from 2D supervision only. Furthermore, Texture Fields only allow for synthesizing novel textures conditioned on a particular input 3D shape while we learn a generative model for both shape and texture. Given the different settings, a fair comparison to Texture Fields is hence not possible. We will clarify these differences in the final version.

Multi-scale patch discriminator: We acknowledge that a patch discriminator with a *fixed* receptive field was previously used for GANs (main paper, l.176). Both Pix2Pix and CycleGAN use a patch discriminator to model high-frequency details but still require an additional loss on the full image. In contrast, we propose to use *multi-scale* patches to model both local and global content, thereby avoiding to generate the full image which is difficult for neural radiance fields due to the large computational and memory requirements involved in rendering a single pixel.

Loss function: We indeed train GRAF only with a non-saturating GAN loss and an R1-regularization (see Eq.(10)).

Reviewer 3

High-frequency details: Ideally, the generator learns to model high-frequency details to generate realistic patches at the finest scale $s = 1$. We agree that it is indeed interesting to use scale dependent discriminators. To circumvent excessive memory requirements we instead tried to append the scale to the discriminator input but did not observe an improvement in our experiments. However, this might become useful when scaling to very high resolutions.

Reflection effects and specularities: We render Photoshapes with a big area light. Thus, the ground truth dataset does not have strong reflections and specularities. For cars, our method is able to produce view-dependent specularities as can be seen in the supplementary video (1:50-1:55). To further improve appearance it could be beneficial to disentangle scene properties like materials and lighting in order to generalize to different lighting effects across scenes.

Ripple artifacts: We hypothesize that the ripple artifacts are a consequence of the positional encoding because it transforms the input to a periodic signal. This naturally encourages periodic structures in the output.

Reviewer 4

Quantitative evaluation of 3D consistency: We thank the reviewer for the valuable suggestion. We found that the Fréchet point cloud distance [Shu et al., ICCV2019] is very sensitive to object poses. Therefore we adopt Minimum Matching Distance (MMD) [Achlioptas et al., ICML 2018] to measure the chamfer distance (CD) between 100 reconstructed shapes and their closest shapes in the ground truth. Results on cars suggest that our multi-view consistent images lead to better 3D reconstruction compared to HoloGAN. This table will be added to the final version.

Method	Ours	HGAN	HGAN ~~3D~~
MMD-CD	**0.044**	0.109	0.092

Visual quality: We thank the reviewer for suggesting to incorporate additional inductive biases. While our method works well on simple objects like cars, we agree that more inductive biases are needed for complex real world scenes. An interesting direction could be hybrid representations that combine convolutional architectures with FC networks.

Learned pose distribution: We thank the reviewer for the suggestion to also learn the camera pose distribution. We indeed consider this as an interesting research direction and will explore it in our future work.