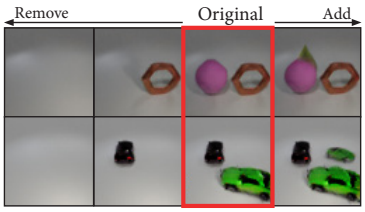
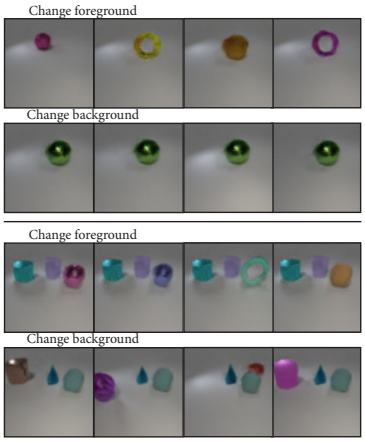
We thank the reviewers for their encouraging feedback. We will revise our writing as suggested (R1, R2 and R4), and discuss the prior work mentioned by the reviewers in the final manuscript (R1, R2 and R4).

BlockGAN is a generative model that learns 3D **object-aware** scene representations using only *unlabelled* images. We show that BlockGAN works with both synthetic datasets with simple backgrounds and real images with complex background and lighting. We evaluated BlockGAN on 64×64 images, which are common for this line of work [1,2]. Due to resource constraints, we did not train BlockGAN on images with higher resolution, but instead chose to focus on pushing the complexity, both in terms of number of objects and, especially, in terms of texture and cluttered background, of the datasets. We believe that this is the first demonstration that deep 3D **object** representations can be learnt directly from natural images without any template geometry, pretrained object detector, or multi-view input.

R1: *Claims:* Although pose is an input to our model, no GT pose labels were used for training. Hence, we maintain our claim that we do *not* need any pose *supervision*. In addition to synthetic images with a simple background, we also train BlockGAN on the real CAR dataset with complex, natural backgrounds (see Figures 5 and 6). Changing the background object, in this case, changes not only lighting but also colour and texture.
Learnt renderer: Yes, it is category-specific. We discuss a shared renderer as future work in line 131 of the supplement.
Objects' appearance interaction: As the foreground object moves, its appearance and shadow move accordingly, depending on where the object is in relation to the camera view (specularity) and lighting positions (shadows) – this can be observed most clearly in the animated results. Indeed, we leave more complex effects, such as inter-object reflection between foreground objects, as future works as discussed in line 272 in the paper.
3D vs 2D convolutions: Our goal is to perform 3D transformations on deep 3D features (including the background). Performing scene combination in 3D allows representing geometry and appearance independent of camera specification.
Removing objects: We show adding and removing objects on the right →
ShapeNet dataset: ShapeNet contains only a limited number of textured models, most of which are low quality, e.g., no specularity. Note that our SYNTH-CHAIR dataset contains ShapeNet chairs with high-quality textures from PhotoShape.



R2: In the final version of the paper, we will also add a discussion on the projection/depth composition method, in addition to the rendering function. *Number of objects:* On the right, we show BlockGAN with 2 foreground (FG) object generators trained with images containing 1 or 3 FG objects. **1 object (top):** Changing either FG object changes the object's appearance and pose; changing the background works as expected. **3 objects (bottom):** Changing one FG object changes one object as expected; changing the background changes one FG object and the background. *Compositing function:* We perform max pooling across objects. This does not require any learning, and, more importantly, is agnostic to the number of inputs, allowing any number of objects to be added at test time. We have set up an experiment with a learned linear weight per voxel, as suggested by R2; however, the training collapsed, and we could not get it to work during the rebuttal period.



R4: *Performance of voxel grids:* Voxel grids can be more memory efficient when adopting warping fields [Neural Volumes, SIGGRAPH 2019], a multi-resolution strategy [Lighthouse, CVPR 2020] or sparsity [Neural Sparse Voxel Fields, arXiv 2020]. Moreover, HoloGAN [ICCV 2019] showed that voxel grids with low spatial resolution but high feature dimension can be very expressive. Note that the choice of voxel grid does not affect whether shape and appearance can be separated – both voxel grids [HoloGAN] and implicit functions [Texture Fields, CVPR 2020] can separate shape and appearance.
Image encoding: Many GAN models (apart from ALI and BiGAN) lack an inference (image encoding) mechanism. However, recent work on training image encoders, such as Image2StyleGAN [ICCV 2019], exploit the representations learnt by GANs to great effect. We look forward to extending BlockGAN towards this direction in future work.
Alternative representations: We thank the reviewer for pointing out [3], which we will cite and discuss. However, this work requires predefined 3D mesh templates for each category (which are not always available) and a pretrained Mask R-CNN to detect objects in images, while BlockGAN learns to disentangle and represent objects using only unlabelled 2D images. Scene Representation Networks [NeurIPS 2019] need many images with labelled poses to learn a good representation for each scene. More importantly, only Visual Object Networks [NeurIPS 2018] and HoloGAN [ICCV 2019], which both use voxel grids, are trained successfully in an unsupervised manner, and can work across multiple scenes. This makes voxel grids a reasonable and effective choice to achieve the goals of our paper.

References: [1] Investigating object compositionality in Generative Adversarial Networks. Neural Networks 2020. [2] Towards Unsupervised Learning of Generative Models for 3D Controllable Image Synthesis, CVPR 2020. [3] 3D-Aware Scene Manipulation via Inverse Graphics. NeurIPS 2018.