We thank all the reviewers for their valuable feedback. We hope that our answers below clarify all concerns.

Evaluation on a labeled dataset. In the paper we used LSUN datasets because it provides many examples of images for a single category and this is important for the quality of GAN. Since this dataset is not annotated, we used Mask R-CNN to get approximated ground truth segmentation on LSUN Car dataset. To address concerns of ***reviewer 1*** and ***reviewer 2***, we trained a model on the annotated Caltech-UCSD Birds-200-2011 dataset (*Wah et al., The Caltech-UCSD Birds-200-2011 Dataset, California Institute of Technology, 2011*). We used the parameters that worked best on car dataset in ablation studies. We then trained encoders in similar fashion as we did for LSUN Car and ran the evaluation, obtaining **mean IoU = 0.380**, while the reference box IoU (for the mask covering the entire image) is **0.132**.

Object-centric datasets / robustness. We show failure cases in Fig. 4 (b). Our experiments on LSUN datasets show that the training of our generator is sufficiently robust to foreground objects appearing at different sizes, viewpoints, locations and where several images have object parts only. Also, see the experiment below with images from two categories. However, because of current GAN limitations, we may not be able to capture all the data distribution modes (small scales, extreme shifts, etc.).

Dataset with more than 1 object category. ***Reviewer 1*** noticed that a huge amount of effort goes into collecting a dataset that only contains the specific category of interest. We argue that our method should work with multiple object categories when the GANs improve and are able to produce realistic images on diverse datasets. To show this, we trained our model on a dataset consisting of 50k images from LSUN Car and 50k images from LSUN Horse images. Although the quality of generated images on such a dataset is lower, our model is able to generate segmented scenes for this dataset consisting of two object categories, as presented in the image below.

However, we could not work on the MS COCO dataset because current GANs are not able to deal with its complexity.

Importance of η parameter. ***Reviewer 2*** expressed concerns about the importance of the η parameter, which defines the minimum area of the foreground object. We ran an additional ablation experiment with $\eta = 5\%$ and found that the results are similar to our default parameters with $\eta = 25\%$, as presented in the table below. The loss term for the minimum area of the foreground object helps avoid the degenerate solution (empty masks), but the generated objects must look real and this is what determines their size.

	64×64			128×128		
Setting	**mIoU**	**reference mIoU**	**detected cars**	**mIoU**	**reference mIoU**	**detected cars**
Default parameters $\eta = 25\%$	0.685	0.440	6293	0.533	0.432	7090
$\eta = 5\%$	0.693	0.458	6202	0.552	0.43	7256

Evaluation of image generation quality. As ***reviewer 2*** suggested, we ran an evaluation of the quality of images generated by our model. We compute FID using 10k real and 10k generated composite images from our model and compare it with a standard StyleGAN producing the entire images at once, trained for the same number of iterations. The results are presented in the table below. The difference in the FIDs may be explained by the more demanding constraints of our model, which may hinder the GAN training.

Setting	**FID** 64×64	**FID** 128×128
Single output GAN	27.807	21.665
Our GAN	31.409	30.867

Prevention from all-ones masks. ***Reviewer 4*** mentioned the danger of getting a degenerate solution in which all of the masks are ones. In our approach random shifts of the foreground object prevent this solution. If the mask consists of all ones, a shift will create a detectable transition between the exposed background and the foreground.

Training the encoder. ***Reviewer 4:*** To obtain the segmentations, we split the dataset into sets of 100 images and train a separate encoder for each set. We then use the encoder to extract the latent codes for its set. We found the training to be more challenging as we added more images.

Literature review. While we focused on prior work on unsupervised and weakly supervised segmentation, ***reviewer 2*** suggested including more works on layered generative models. We will add those and others to better underline the contribution and novelty of our method, which works on real images and without supervision.

Comparison with cut-and-paste, other baselines. We could not prepare the comparison in time for the rebuttal but we will add it in the camera-ready paper if accepted.