We thank the reviewers for their relevant and insightful feedback. We are also grateful for pointing out related work we have missed and we will include them in the final version of the paper. Will also be included additional comments formulated in this response page and more data points will be added to Figure 3 (relative to the performance of supervised baselines) following recommendations of Reviewer 3. Unfortunately, the updated figure and associated commentary had to be cut from this response.

[R-1 issue-1] Does the independence assumption hurt the general usefulness of our method?

In our framework, the independence between regions is conditional to a mask. Therefore, our generator can infer some information about the objects, e.g. class label, from the masks. Following the recommendation of Reviewer 1, we trained ReDO on a combination of LFW and Flowers images (without labels). This new dataset has more variability, contains different types of objects, and display a more obvious correlation between the object and the background. Re-using the same hyper-parameters as for the Flowers dataset, (probably not optimal for this fused dataset), we obtained a reasonable accuracy of 0.856 and an IoU of 0.691. Masks and generation samples are displayed in Figure 1. These preliminary results show that ReDO is not restricted to problems with a single object category. These updated experiments will be included in the final version of the paper.

Figure 1: Results on LFW + Flowers dataset, arranged as in Figure 2 of the paper. As $\mathbf{z}$ is kept constant on a column across all rows, we can observe that $\mathbf{z}$ codes for different textures depending on the class of the image even though the generator is never given this information explicitly.

[R-1 issue-2] Does our model produce a variety of objects?

The variability of the generated images is not the goal of the model and only serves as a means to recover meaningful masks. In principle, for our method to work, it should be sufficient that the variability of the pixel-values is non-zero. Anyway, experiments show that samples are quite diverse even with a fixed mask, as illustrated in Figure 2 of the submitted paper (and Figure 1 of this response page). In any case, they are diverse enough to learn to extract the masks.

[R-1 issues-3-4, R-3 issue-3] About the role and effects of the information conservation constraint

The reviewers noted the lack of an ablation study on the information conservation constraint. Reviewer 1 also noted that our model can converge to a trivial solution. Indeed, without the information conservation loss, an optimal solution to our objective function is for our system to output one mask that covers the whole image while the others are empty, for instance, $\mathbf{M}_1 = 1, \mathbf{M}_2 = 0$.

As mentioned, albeit very succinctly, in lines 178-179 of the submitted paper, we introduced the information conservation loss as a way to prevent this undesirable solution. When generating an image, if $\mathbf{M_i} = 0$ then information from $\mathbf{z_i}$ is lost and cannot be retrieved from the final image. Or equivalently, if $\mathbf{z}_i$ can be retrieved then $\mathbf{M}_i$ is non-zero. Enforcing the conservation of information of all $\mathbf{z}_k$ ensures that no region is empty.

However, as reviewer 1 also noted, generators are able to encode information in imperceptible ways, so the method is not foolproof and a mask with near-zero values is still possible. But qualitative considerations presented in Figure 2 of the submitted paper and the fact that our model is able to discover meaningful regions when information conservation loss is applied (and fail to do so otherwise) show empirically the effectiveness of the constraint despite its potential issues. These discussions will be added to the final version of the paper.

[R-2 issue-1] Why these datasets? Why not ImageNet?

The choice of datasets is motivated by two aspects: 1) the ability to fit a GAN on the dataset and to experiment with it, with reasonable time and computing power, 2) the availability of ground truth segmentation masks to evaluate our model. Unlabelled ImageNet is still a difficult setup for GANs and doesn't come with ground truth segmentation masks.

[R-3 issue-1] On the ability of the model to segment n-objects.

We have chosen to describe the model in the general case with n objects since the ReDO approach is generic and can be used in the case with $n > 1$. This ability is illustrated qualitatively in the paper on toy dataset cMNIST with $n = 3$. On that experiment, we had obtained an accuracy of 0.99 and an IoU ratio of 0.78 for the two objects. These results, however, are not entirely informative about how well ReDO would perform on a real multi-object dataset. In fact, compared to the discovery of a single object, this is clearly a more difficult setting and would require more computing power. We are focusing on this problem as an important future step.