We sincerely thank our reviewers for the constructive feedback. We are glad to see our paper is well-written (R1, R3, R4), has strength in visual quality (R1, R3, R4), and performs diverse (R1) and thorough (R3, R4) experiments.

**Novelty [R1, R3]** Note that we do not claim to be the first to use swapping for disentanglement. Our key contribution is in the way we *constrain* this disentanglement, by separating intra-image information (co-occurrence of patches within the same image) from inter-image information (visual content across images within the dataset). Typically, learning for image synthesis happens by gathering information across the entire image dataset (or across an image class). However, this is difficult to do correctly, as the texture statistics within each image can be quite different and incompatible. Thus the recent popular "internal learning" methods, such as InGAN[58] or SinGAN[59], show superior results by learning from the patches of just the input image, utilizing the commonality of texture statistics. But, of course, methods like [59] can't learn that much because their dataset is very limited (just one image!). What our proposed method is doing is disentangling the texture-like information useful for internal learning from the structure-like information that can profit from external learning. To separate out internal learning from external learning, we model internal information (texture code) using patch co-occurrence D (which operationalizes Julesz's 2nd order texture conjecture), which leaves the structure code representing texture-invariant (compatible with texture code of any image). So, we combine the best of both internal and external learning. The swapping is just a self-supervised pretext task to accomplish this disentanglement. Figure 6 confirms that the quality of disentanglement is superior to generative models that don't utilize internal information (Im2StyleGAN[1], StyleGAN2[38]) or external information (STROTSS[41], WCT$^2$[69]). In the current draft, this motivational story got distributed throughout the paper, but in the next draft we will be sure to summarize it concisely in the intro as well.

**Comparison to FUNIT [R3]** FUNIT is fundamentally different from our work; it requires labels to define the disentanglement of style from structure. For example, the AnimalFaces, Birds, Flowers, and Food datasets contain 149, 444, 85, and 224 classes, respectively. ***For this reason, FUNIT cannot be trained on any of the datasets in our paper***. In contrast, our method is fully unsupervised, and works because co-occurrence statistics of patches of a single image carry enough signal for learning a smooth embedding space for image editing. By doing so, our method can be easily trained on user-collected datasets such as the Flickr Mountains and Waterfalls, which is useful for long-tail image editing. In more detail, the learning objective and discriminator design are substantially different; the FUNIT discriminator requires class labels, while ours performs internal learning on unlabeled images (see Novelty). While substantially different to FUNIT's discriminator in concept, we agree our discriminator shares architectural similarities with FUNIT's encoder. To clarify, the PatchEncoder of Figure 2 of SuppMat is the feature extractor of our discriminator, not our encoder. We apologize for any confusion.

**Image Quality, Resolution, and Versatility [R1]** We show our method works well on a wider variety of datasets than most GAN-based methods, on human and animal faces, outdoor and indoor scenes, user-collected datasets, and paintings (Section 1.3 of Supp Mat). Moreover, our method supports HD resolution (e.g. the mountain example of Figure 7, the demo video and Supp Mat). Lastly, we evaluated the quality of texture transfer against two SOTA style transfer methods (non-photorealistic and photorealistic), and show that our method is competitive, if not better. Therefore, we believe our method has advantage over general texture transfer methods.

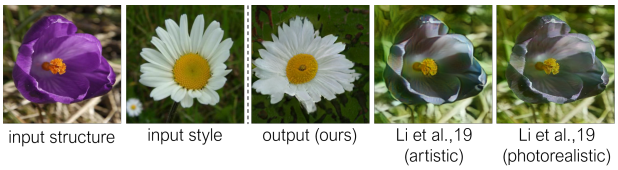
**Cite and compare to disentanglement / neural style transfer papers [R1]** Thank you for the suggestions. We will cite the suggested works. Our end goal differs in that a real, existing image must be faithfully embedded for editing, while the suggested works do not focus on accurately reconstructing input images.

**Comparison to StyleGAN2 [R3]** Our method aims to build an image editing model in an unsupervised setting. To our knowledge, the state-of-the-art unsupervised methods for GANs-based image editing are Im2StyleGAN[1] and StyleGAN2[38]. Randomly-generated StyleGAN2 images cannot be used for this application.

**Injection [R4]** We appreciate your feedback. We will modify "clearly injective (L119)" to "optimized toward injection", and also revise the Supp Map accordingly.

**About Broader Impact [R4]** We are happy to move the figure to Supp Mat based on reviewers/AC's recommendation.

**Structural Limitation [R3]** We show the result of the suggested lily / daisy example. Please see the figure below.




Structure-style mixing results on the Oxford Flower dataset. Our model can modify fine-level structure if the texture code requires it, while maintaining the coarse-level structure. We additionally show the result of a real-time style transfer method (Li et al., CVPR2019 [A6 suggested by R4], both artistic and photorealistic version) that makes only small changes.

**Comparison to recent style transfer models [R4]** Thank you for the suggestion; we will discuss our work in the context of neural style transfer more explicitly. Note that style transfer methods, even if real-time, are not suitable for *natural* image manipulation, because they often fail to produce photorealistic images (please see left) and do not allow controllable structural editing. For thoroughness, in the submission, we compared the quality of disentanglement to both SOTA non-photorealistic (STROTSS) and photorealistic (WCT$^2$) style transfer methods. While WCT$^2$ does achieve better structural preservation, it does so by barely changing the style (Figure 5).