We thank the reviewers for their detailed feedback—it will help improve our paper.

(R1) Training difficulties. Training the models was actually quite simple (involved adding only 5-6 lines of code on top of standard/boilerplate model training code), and extremely consistent and stable. In fact, our method scales easily to larger datasets (unlike GANs) and does even better on "harder", more fine-grained tasks.

(R3 W1, R3 W2) The reviewer's concern about overclaiming is well received (and appreciated). We are very much aware that our approach does not "solve CV" and did not intent to claim that. In fact, that's why we explicitly list the tasks we perform in the introduction (and our intention was to use the phrase "CV toolkit" in the same way one might describe GANs as enabling a "CV toolkit", despite not being all-powerful).

However, in the light of this feedback, we decided to make this point more clear. To this end, we will change our title to "Image Synthesis with a Single (Robust) Classifier". Similarly, we will modify the abstract, intro, and conclusion accordingly, changing "CV toolkit" to "toolkit for (class-conditioned) image synthesis tasks," etc., as suggested.

(R3, R6) Quality of synthesized images. We want to emphasize that the goal of our paper is not to directly improve on state-of-the-art for any of the tasks that we consider. Instead, our goal is to introduce robust classifiers as a new and promising framework that even "out of the box" (i.e., without any real engineering) is able to perform these tasks at a reasonable level. Crucially, our framework does not seem to have hit any fundamental obstacles preventing it from attaining state-of-the-art results. To contrast it with GAN-based approaches, one should note that:

(A) The "out-of-the-box" GAN framework (minimizing the theoretically motivated loss with standard architectures) gives results that are far from satisfactory. In particular, they are worse than what we get with an off-the-shelf robust classifier and no optimizations/regularization/proprietary datasets in Fig. 3. Attaining the current state-of-the-art results with GANs required years of effort that involved devising computational and computer vision-based optimizations. In contrast, our ImageNet results (Fig. 3) would be state-of-the-art even 3-4 years after GANs were introduced. Indeed, scaling GANs to ImageNet was possible only recently—whereas robust classifiers actually seem to *improve* with more fine-grained classification tasks. **(B)** The training dynamics of GANs are complex, unstable, and hard to interpret. In our case, the dynamics and the learned representations are just fairly natural variants of the standard (and rather well-understood) discriminative framework.

Finally, we find the R6's comment on "bragging" about the Inception Score somewhat unfair. Our paper explicitly discusses why I.S. is not a good metric to compare the two approaches, and just points out that it is unreasonably high.

Non-robust models work/Unsuprising/Novelty over distill.pub. Overall, we are surprised by these comments. We are not aware of prior work that performs all these vision tasks with an off-the-shelf discriminative model without task-specific optimization. In fact, even the work the reviewer cites `https://distill.pub/2017/feature-visualization/appendix/` notes:"In this layer [the final layer, used for feature painting] visualizations become mostly nonsensical collages...*neurons do not seem to correspond to particularly meaningful semantic ideas anymore.* [emph. added]."

Indeed, replicating our experiments using a standard network (without regularization) fails completely (this has been observed by several papers, including the linked Feature Visualization one—we will include examples in our appendix). Even when one employs quite intricate regularization methods (gradual upscaling, gradient blur, blur reduction, random shifts, etc.) the generated images from a single discriminative model are qualitatively worse than what we present. For example, compare `https://bit.ly/2JYr0yh` and `https://arxiv.org/abs/1507.02379`, (which are the most recent class visualization sans trained GAN/AE we found, via the distill.pub article), and even the image labeled "Class Logits" in the distill.pub paper, to Figure 3 in our work.

Finally, as mentioned earlier, prior work relies heavily on additional regularization—our work highlights that when using robust models none of this is necessary (and in fact, the promising advancements in regularization for non-robust networks can even be integrated with robust classifiers in the future to get even better results).

(R6) Overlap with Tsipras et al. While Tsipras et al was a motivation for this work, we disagree that there exists a significant overlap between the two works. In particular, the Tsipras et al. paper simply shows that untargeted attacks on robust models, starting from test set images, seem to change relevant features in the input. In our work, we show that one can actually leverage robust classifiers for diverse class-conditional image generation, inpainting, superresolution, etc. which as far as we are aware have never been done with a standard discriminative architecture.

Minor comments (space constrained): (R1) Our method is general and can work with arbitrary pixel masks. We will show this in revised appendix along with **(R3 C2)** our results for standard networks (which are indeed poor). **(R3 C3)** The difference is purely for implementation convenience, as the constrained and regularized version are technically equivalent (via Lagrange multiplier property) **(R3 C6)** We view robustness as a prior that allows for good synthesis results. In that sense, it is complementary to other priors, e.g., the deep convolutional prior.