**Revision summary:** We thank all the reviewers for their insightful comments. We have added experiments on ResNet/DenseNet backbones, and optimized a more efficient C2sp model for ImageNet.

**Reviewer #1 Comment 1:** I would like to increase the score if the authors can show some discriminative results.

**Response:** We have tried our best to implement and train C2sp + Faster-RCNN models on MSCOCO, it’s still on-going and hard to complete in such limited time. Nevertheless, we’d like to validate it through two evidences: (1) GAN training is also very sensitive to spatial information and transformations, and our results (main text Table 3) have significantly exposed the asymmetric problem. When trained with C3 discriminators, C2 generators directly lead to non-convergence, and C4sp generators are much better than C4. (2) A paper [1] replaces Conv3D with temporal shift + Conv2D, which achieves efficient video understanding and also generalizes to other modalities, e.g., optical flow.

**Reviewer #2 Comment 2:** CIFAR10/100 use different backbones. Report: (1) the performances of ALL backbones with C2, C2sp, C4 and C4sp. (2) the performance improvement of symmetric padding against network depth.

**Response:** The performances of ALL backbones are shown in Figure R1. The accuracy gaps between C2 and C2sp are larger in deeper networks. As for C3 and C4, we have claimed (main text Line 172) that the degradation is dominated by “edge effect”, so C4sp only slightly improves the accuracy (but significantly in GANs). Different backbones can cross-validate the generality and consistency since architectures may affect the results (e.g., concerns in Comment 1).

**Reviewer #3 Comment 1:** To address this concern of cherry-picking, I recommend the authors to explain which channels are selected and show more channels in a figure.

**Response:** They are not cherry-picked. Since each single channel is very stochastic and hard to interpret (examples in Figure R2), 9 channels for 3 stages, C2, the activations in Figure 1 are the average values of all channels, i.e., 16, 32, and 64 channels.

**Reviewer #3 Comment 2:** An experiment using C3 with asymmetric padding.

**Response:** We test ResNet-38 (#channel 18-36-72) on CIFAR10 with four settings: C3 {1111}, C3sp (9 symmetric groups {0202}, {0211},..., {2020}), C3ap3 (3 asymmetric groups {0111}, {1102}, {0202}) and C3ap1 {0202}. The error rates (%) are 5.51 ± 0.08, 5.94 ± 0.05, 6.21 ± 0.03 and 7.27 ± 0.28. The asymmetry gains, the accuracy degrades. C3sp has expended RF 5×5 and is restricted by the “edge effect”, as with C4&C5.

**Reviewer #3 Comment 3:** The performance degradation in C4 is dominated by “edge effect” rather than the shift, I recommend the authors to provide more convincing arguments on their issues, e.g., perhaps by comparing with C5.

**Response:** In Figure RT, error rates C5≫C4>C4sp≫C3, which is consistent with the “edge effect”. Although C4sp provides minor improvement in classifications, it is much better than C4 in GANs, where the “edge effect” is negligible regarding the network depth and image resolution. In summary, the symmetric padding eliminates the shifting problem, and simultaneously expands the reception field. The former is critical, the latter is limited on some occasions.

---
