Rethinking the Role of Spatial Mixing

Supplementary Material

A. The Separable Convolution

In Figure B.1, we show an illustration of the separable convolution. Throughout this work, "Full" refers to both the depthwise and pointwise filters being learned. "Space" refers to only the depthwise (spatial-mixing) filters being learned. Likewise, "Chans" refers to only the pointwise (channel-mixing) filters being learned.

B. More Pixel Un-Shuffle Results

For our CIFAR pixel un-shuffling figure in the main text (Figure 11), the "large" model had a depth of 16, width of 512, and kernel size of 7.

On the following pages, we include further results on pixel un-shuffling that were omitted from the main text due to space constraints.

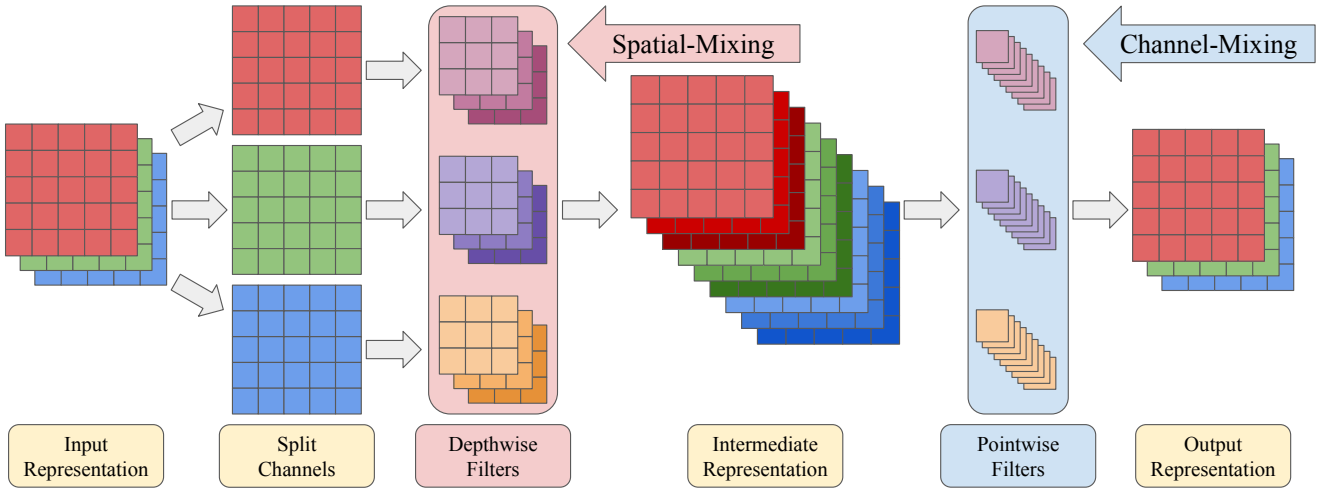
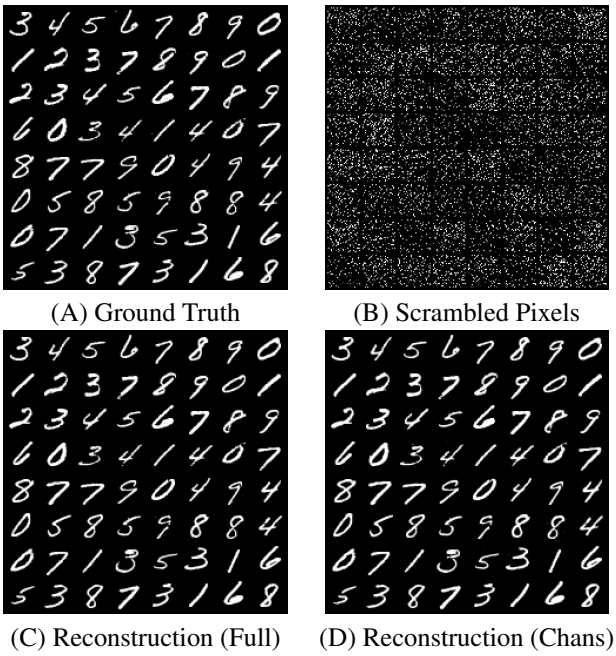


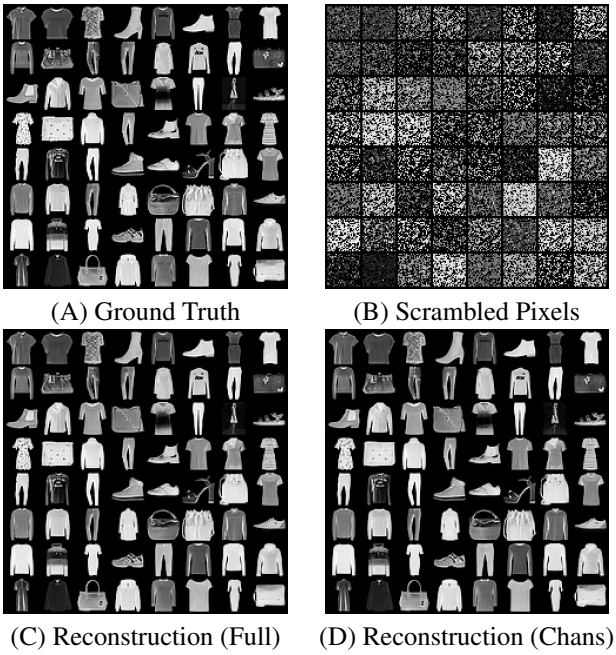
Figure B.1. Separable convolutions allow us to disentangle the roles of spatial and channel mixing in deep networks by freezing either the depthwise or pointwise filters and learning the others.



(A) Ground Truth (B) Scrambled Pixels

(C) Reconstruction (Full) (D) Reconstruction (Chans)

Figure B.2. Models that only learn channel mixing are capable of learning near perfect reconstruction of MNIST digits from their shuffled pixels.



(A) Ground Truth (B) Scrambled Pixels

(C) Reconstruction (Full) (D) Reconstruction (Chans)

Figure B.3. Even on MNIST-Fashion, a significantly more complex dataset than MNIST-Digits, the models that learn only channel mixing still achieve near perfect un-shuffling with results almost indistinguishable from those of the fully-learned models.

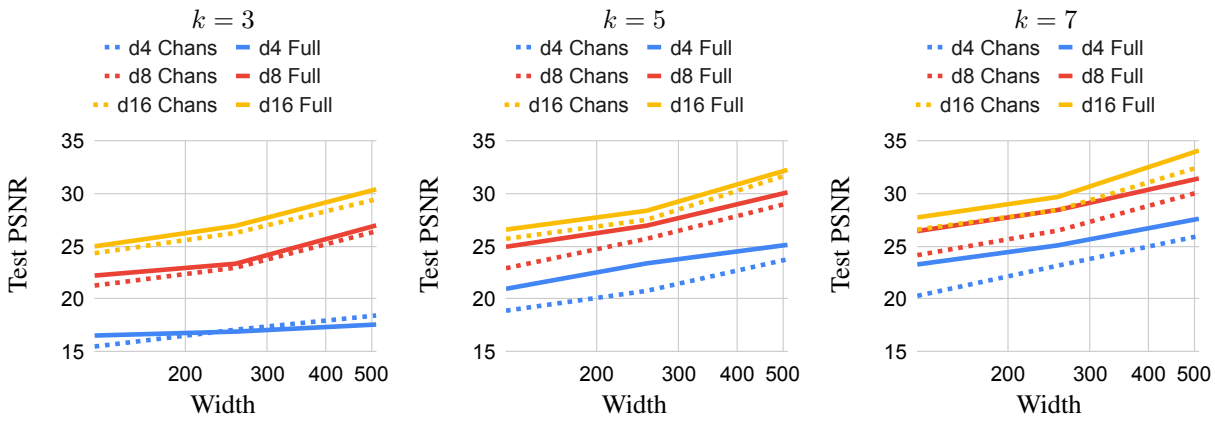


Figure B.4. **MNIST**: PSNR for Pixel Un-Shuffling.

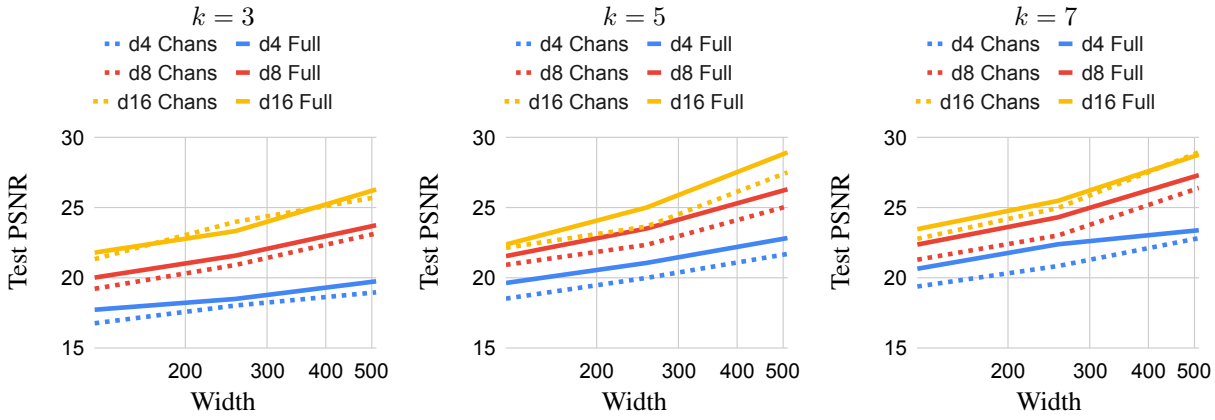


Figure B.5. **Fashion-MNIST**: PSNR for Pixel Un-Shuffling.