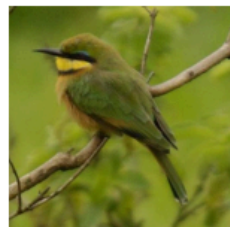
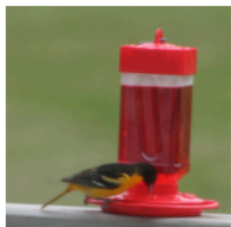
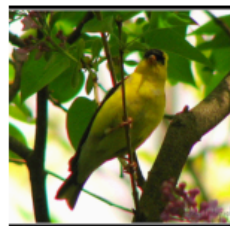
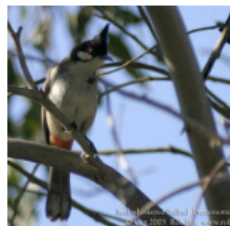
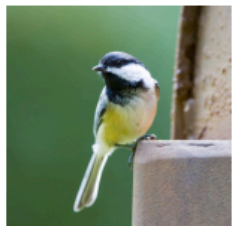


Layer: blocks.10.hook\_mlp\_out, Feature: 6547

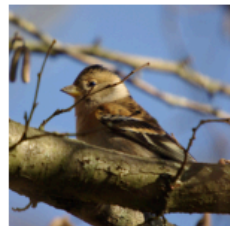
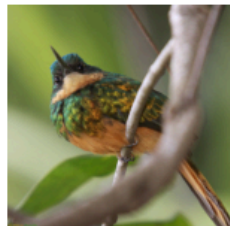
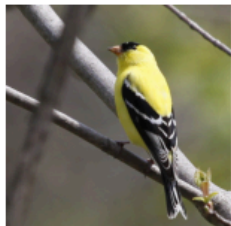
Img idx: 5995 Act: 2.597 Img idx: 6985 Act: 2.434 Img idx: 7598 Act: 2.351 Img idx: 4342 Act: 2.192



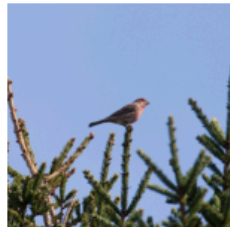
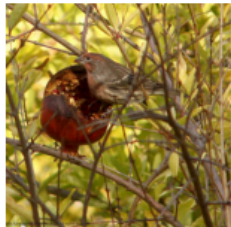
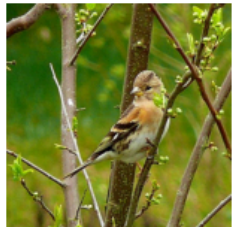
Img idx: 5545 Act: 2.167 Img idx: 6357 Act: 2.156 Img idx: 363 Act: 2.097 Img idx: 1194 Act: 2.078



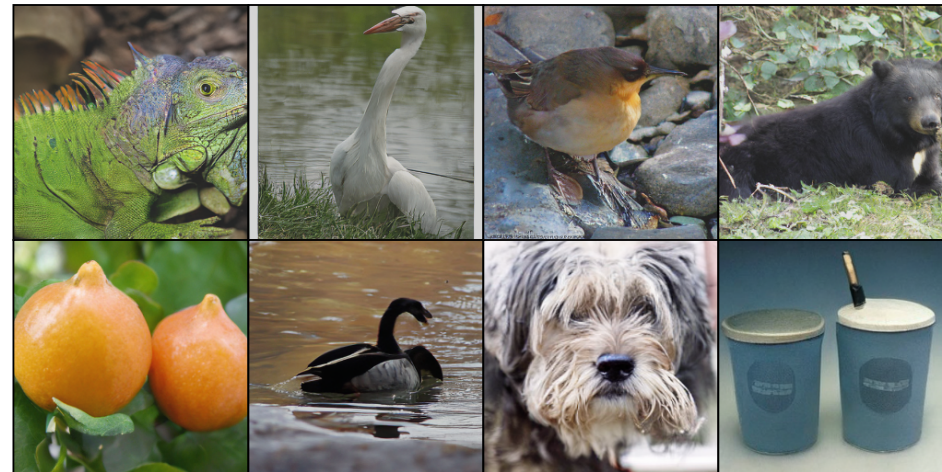
Img idx: 1959 Act: 2.047 Img idx: 4335 Act: 2.047 Img idx: 1533 Act: 1.970 Img idx: 2859 Act: 1.947



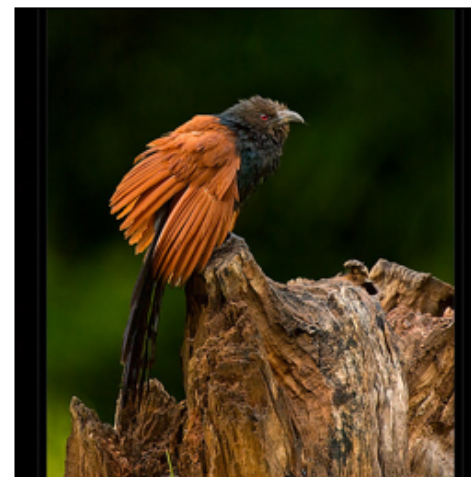
Img idx: 1492 Act: 1.940 Img idx: 6883 Act: 1.938 Img idx: 7469 Act: 1.938 Img idx: 6362 Act: 1.934



example of images activating the same feature as source image



generated invariant samples



source image