Supplementary Material For D3D: Distilled 3D Networks for Video Action Recognition

1. Predicted Optical Flow Visualizations

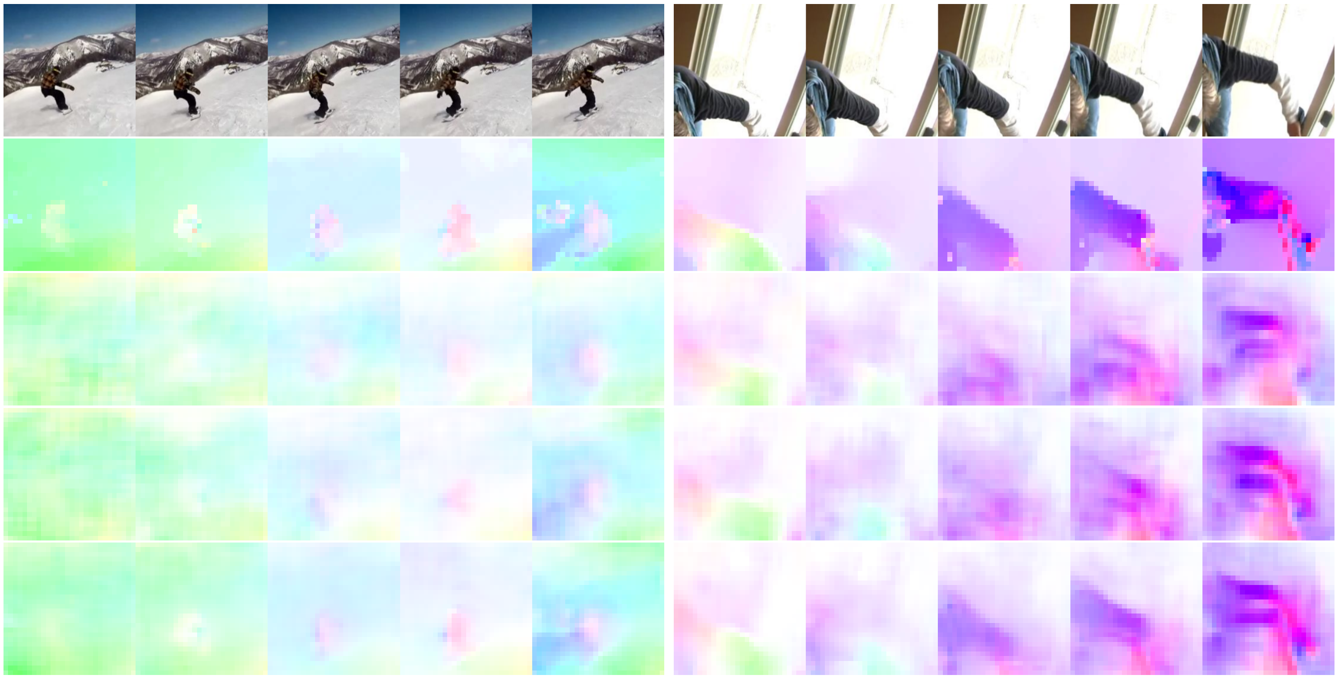

Figure 1: Examples of optical flow produced by S3DG and D3D by adding the optical flow decoder applied at layer 3A. From top to bottom: RGB Frames, TV-L1 optical flow, S3D-G flow, D3D flow, D3D flow with finetuning. TV-L1 optical flow is shown downsampled to 28 × 28 px, which is the decoder output resolution used during training.

2. Performance on Kinetics-400 Categories

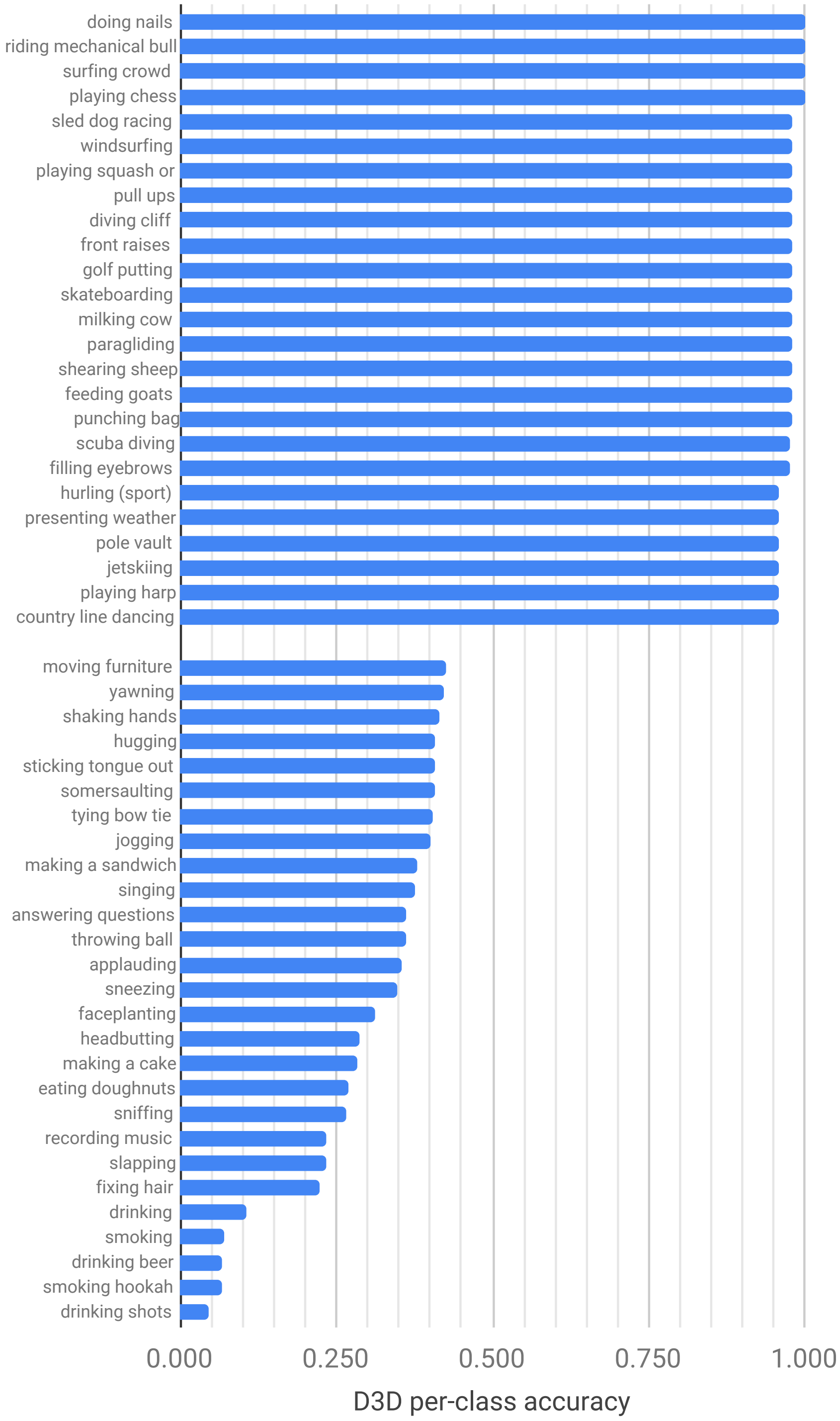

Figure 2: Accuracy on individual Kinetics-400 categories using D3D. We show the per-class accuracy for D3D trained on Kinetics-400. Only the top and bottom 25 classes are shown.

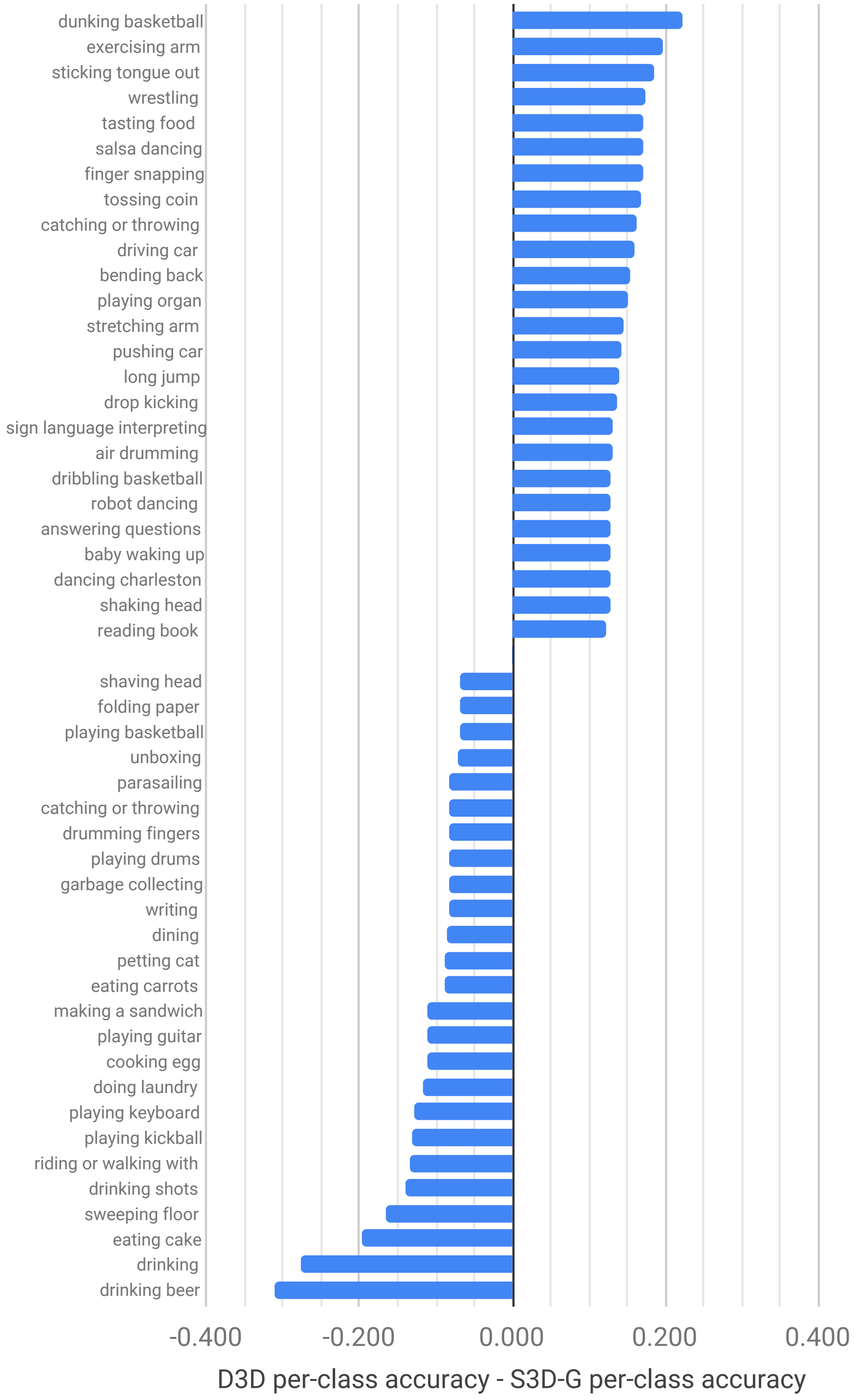

Figure 3: Accuracy difference on individual Kinetics-400 categories by adding distillation. We compare the difference between per-class accuracy for D3D and per-class accuracy for S3D-G. Only the top and bottom 25 classes are shown. In total, D3D leads to improvements on 203 of the 400 classes (50.8%) and degradations on 103 of the 400 classes (27.3%), with less than a ±.1% difference on the remaining classes.

3. Optical Flow Decoders

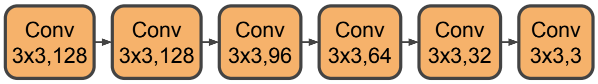

Figure 4: The optical flow decoder architecture. This is equivalent to that of PWC-Net [1], but with two changes: (1) we do not include warping or cost volume layers, and (2) the output is represented using three channels.

4. Non-Local S3D-G

For our experiments with Non-Local S3D-G (NL S3D-G), we include two non-local blocks [2] into the S3D-G architecture, immediately before blocks 5B and 5C. We make no changes to the training procedure or hyper-parameters for these experiments.

We implement non-local blocks similarly to [2], but with two known differences:

1. We do not apply batch norm inside the nonlocal block. Adding batch norm slightly reduced performance.
2. We do not use the sub-sampling trick to reduce the feature map size in the non-local block. This is because the 5X layers in S3D-G already have a small feature map size (7x7x8).

References

[1] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018. 4

[2] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018. 4