Supplementary Material: Optical Flow Estimation using a Spatial Pyramid Network

1. Experiment with Identical Frames

In order to validate our network performance on zero motion, we experiment by passing identical frames as the input to our network. We use the test set of Sintel Clean, which was never seen by our network as inputs. We observe an average flow magnitude of 0.06 pixels for identical frames, which is significantly smaller than the average flow magnitude in Sintel. The average flow magnitude in Sintel is 18.926. We show qualitative results in Figure 2.

2. Network and Training Details

Architecture Choices We found that including the flow field $u(V_{k-1})$ as inputs to the networks G_k improves the accuracy. At Level 2, the EPE was 0.54 without and 0.51 with. We suspect that network is able to exploit the spatial structure in $u(V_{k-1})$.

Weight Sharing We found that using a single G_k across all levels, or using shared weights, decreased the accuracy. Hence, we avoid weight sharing and train a different convnet at each pyramid level.

Higher Pyramid Configurations While adding extra levels to the pyramid for testing on high resolution images like frames of Sintel, we tried several configurations like $(G_0, G_0, G_1, G_2, G_3, G_4), (G_0, G_1, G_1, G_2, G_3, G_4)$ etc for the convnets in the pyramid. We found the best configuration to be $(G_0, G_1, G_2, G_3, G_4, G_4)$.

3. Comparison of Learned Filters

Sequential vs. End-to-end Training. We evaluate our performance on Sequential training of convnets G_k w.r.t training the entire network end to end. We found that full end-to-end training was slower by nearly a factor of 2 and gave a higher EPE (2.99 vs 2.71 on Flying Chairs). Training networks sequentially followed by cascading the networks end-to-end and fine tuning them did not improve performance but kept the EPE steady (at 2.71 on Flying Chairs).

We compare our learned spatio-temporal filters with the spatial filters learned by FlowNet [1]. We observe that while FlowNet's filters are random looking, our filters are more Gabor-like resembling cortical areas MT and V1. We show the visualizations of learned filters in Figure 1.

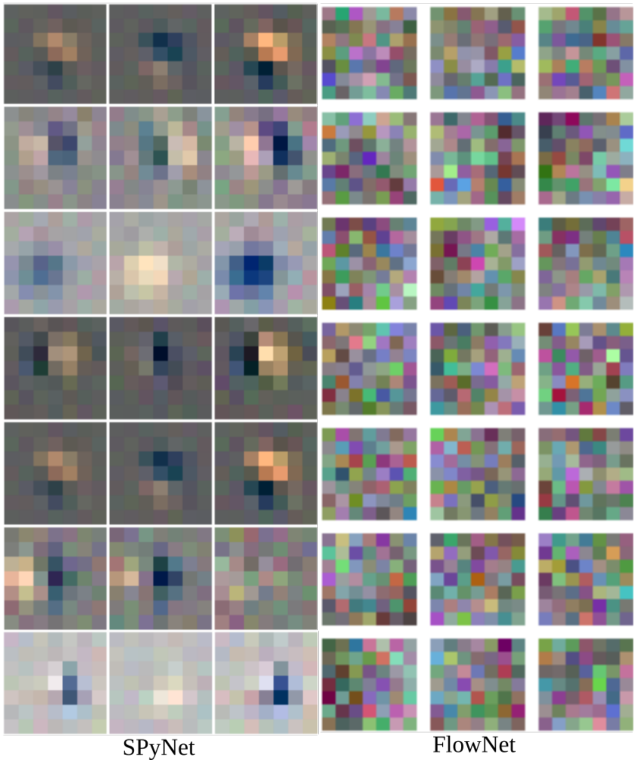
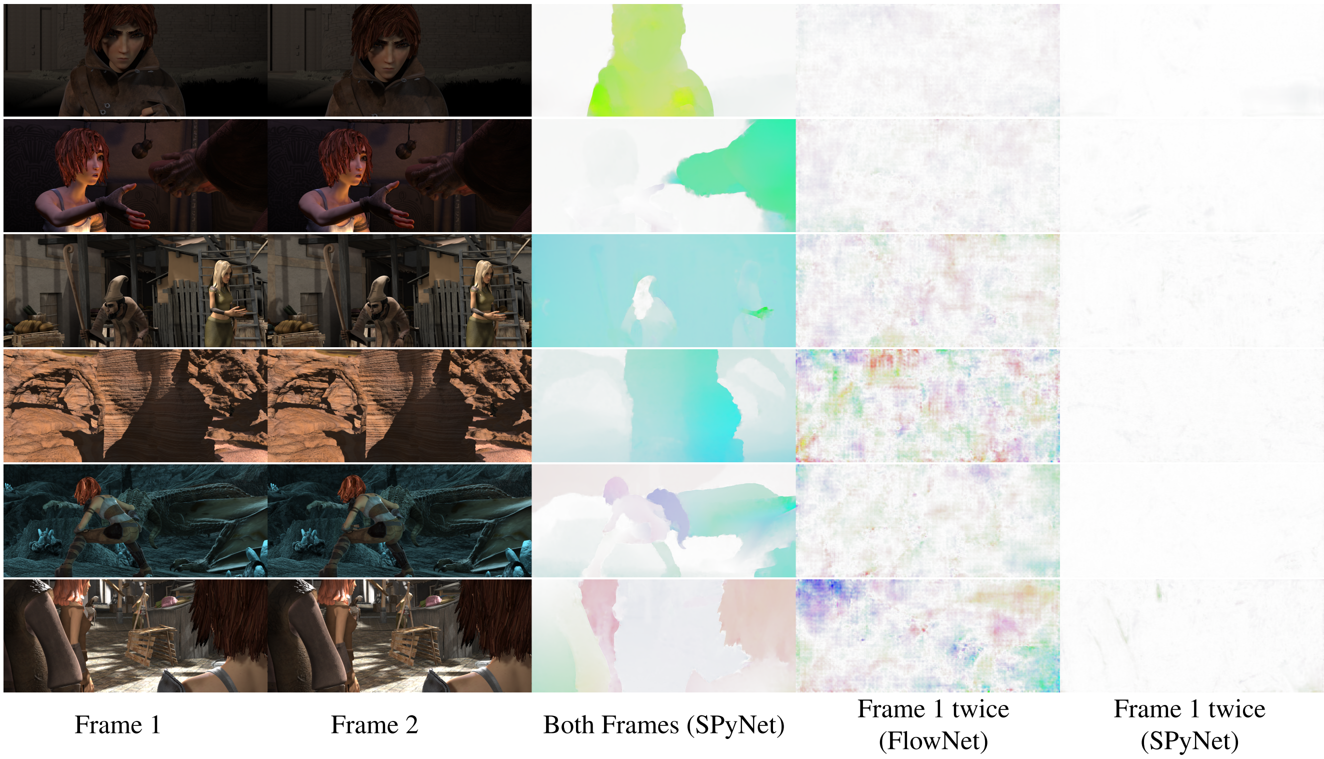


Figure 1. Comparison of learned spatio-temporal filters of SPyNet (left) and FlowNet (right)

References

[1] A. Dosovitskiy, P. Fischery, E. Ilg, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, T. Brox, et al. Flownet: Learning optical flow with convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766. IEEE, 2015. 1



Frame 1

Frame 2

Both Frames (SPyNet)

Frame 1 twice (FlowNet)

Frame 1 twice (SPyNet)

Figure 2. Optical flow using both frames and identical frames. Note that flow for the latter case for SPyNet is nearly zero. The flow on identical frames has been magnified by 10x for visualization.