Im2Flow: Motion Hallucination from Static Images for Action Recognition (Supplementary Materials)

Ruohan Gao
UT Austin
rhgao@cs.utexas.edu

Bo Xiong
UT Austin
bxiong@cs.utexas.edu

Kristen Grauman
UT Austin
grauman@cs.utexas.edu

The supplementary materials consist of:

A. Details of Im2Flow network architecture.

B. Details of the metrics we use for flow prediction evaluation

C. Quantitative results of flow prediction on HMDB-51 and Weizmann.

D. Comparison of different encoding schemes for motion prediction.

E. Additional qualitative results.

F. Ablation study.

A. Details of Im2Flow Network Architecture

Our Im2Flow network architecture is adapted from those in [3, 5] with some modifications. Let C_k denote a Convolution-BatchNorm-ReLU layer with k filters. CD_k denotes a Convolution-BatchNorm-Dropout-ReLU layer with a dropout rate of 50%. The encoder uses dilated-convolutions [7], and the decoder uses up-convolutions. All dilated-convolutions and up-convolutions use 4 x 4 spatial filters applied with stride 2. After the last layer in the decoder, an up-convolution is followed by a Tanh layer to produce the flow image. BatchNorm is not applied to the first layer in the encoder. The encoder uses leaky ReLUs with a slope of 0.2, while ReLUs in the decoder are not leaky. Skip connections are added between each layer i in the encoder and layer $n-i$ in the decoder, where n is the total number of layers. The skip connections concatenate activations from layer i to layer $n-i$.

Encoder:

C_{64}-C_{128}-C_{256}-C_{512}-C_{512}-C_{512}-C_{512}-C_{512}

Decoder:

CD_{512}-CD_{1024}-CD_{1024}-C_{1024}-C_{1024}-C_{512}-C_{256}-C_{128}

The motion content loss network is the first two residual blocks of ResNet18 [2], which produces an activation map of size $128 \times 28 \times 28$.

B. Details of the metrics we use for flow prediction evaluation

We employ a suite of metrics, following prior work in this area [4, 6]: End-Point-Error (EPE), Direction Similarity (DS), and Orientation Similarity (OS). EPE computes the Euclidean distance between the end point of the predicted optical flow vector and the ground-truth vector. It is a direct error measure, but is weak if the motion is small and ambiguous in direction [4, 6]. DS is the cosine similarity between the prediction and the ground-truth, and OS (unsigned version of DS) measures how parallel the predicted and ground-truth flow vectors are (see [6] for details).

$$DS = \frac{u_1u_2 + v_1v_2}{\sqrt{u_1^2+v_1^2}\sqrt{u_2^2+v_2^2}},\ OS = \frac{|u_1u_2 + v_1v_2|}{\sqrt{u_1^2+v_1^2}\sqrt{u_2^2+v_2^2}}. \quad (1)$$

OS is useful to evaluate predicted motions whose exact direction may be ambiguous (e.g., push-ups in Fig. 1 in the main paper).

C. Quantitative Results of Flow Prediction on HMDB-51 and Weizmann

Table 1 shows the flow prediction results on HMDB-51 and Weizmann, as a supplement to Table 1 in the main paper, where due to space constraints we could show only the results for UCF-101. Because the authors' model for [6] is trained only for UCF, for fair comparison, we compare with only Pintea *et al*. [4] on HMDB-51 and Weizmann. On these two datasets, our method still outperforms the prior approach and the Nearest Neighbor baseline consistently by a large margin across all metrics. These results are using the same metrics as in the main paper. The results show the effectiveness of our proposed motion encoding and Im2Flow network.

HMDB-51	EPE ↓	EPE-Canny	EPE-FG	DS ↑	DS-Canny	DS-FG	OS ↑	OS-Canny	OS-FG
Pintea *et al*. [4]	2.621	2.683	3.576	0.001	0.000	-0.008	0.498	0.532	0.552
NN-pool5	3.635	3.847	4.124	-0.004	-0.005	-0.043	0.643	0.641	0.649
Ours	**2.571**	**2.629**	**3.389**	**0.086**	**0.079**	**0.085**	**0.676**	**0.666**	**0.674**
Weizmann	EPE ↓	EPE-Canny	EPE-FG	DS ↑	DS-Canny	DS-FG	OS ↑	OS-Canny	OS-FG
Pintea *et al*. [4]	0.562	1.867	5.955	0.101	0.088	0.095	0.712	0.701	0.723
NN-pool5	0.588	2.212	6.324	0.002	0.003	0.005	0.689	0.688	0.695
Ours	**0.512**	**1.739**	**5.455**	**0.380**	**0.366**	**0.375**	**0.801**	**0.789**	**0.824**

Table 1. Quantitative results of dense optical flow prediction on HMDB-51 and Weizmann. ↓ lower better, ↑ higher better. Across all measures, our method outperforms the baseline methods by a large margin. Because the model provided by Walker *et al*. [6] is trained on UCF-101, we don't compare with them on HMDB-51 and Weizmann for fairness.

	Accuracy	mAP
Only Pixel L2 Loss	46.3	54.4
Without Action Label Supervision	49.5	57.2
Walker *et al*. [6]	21.2	29.9
Ours	51.0	58.8

Table 2. Ablation study. All results are in %.

D. Comparison of Different Encoding Schemes for Motion Prediction

As discussed in Sec. 3.1 in the main paper, in this section, we compare the motion prediction results of our encoding scheme with two other encoding schemes in the literature [1, 8] to show the advantage of our encoding. These two encoding schemes also encode optical flow as a single three-channel image. Table 3 compares the flow prediction results on UCF-101. Our encoding scheme is more suitable for flow prediction, and produces much more reliable predictions.

E. Additional Qualitative Results

We show more qualitative results in this section. In the supplementary video we attached, we show some motion prediction results of video sequences using our Im2Flow framework. We predict for each independent frame as if it were a static image, and then just concatenate the frames as laid on the video for visualization. There is no temporal smoothing between frames' estimates. Our Im2Flow network can predict motion in a variety of contexts, and the prediction is pretty fine-grained. In Fig. 1, we show more examples to illustrate how the inferred motion can help recognition, as a supplement to Fig. 5 in the main paper. While a classifier solely based on appearance can be confused by actions appearing in similar contexts, the inferred motion provides cues about the fine-grained differences among these actions to help recognition. For instance, the last image shows a man playing nunchucks. However, playing Yo-Yo, playing nunchucks, and playing juggling balls usually appear in similar contexts. Showing the hand movement of the man guides the classifier to make the correct prediction.

F. Ablation Study

As discussed in Sec. 3.2 in the main paper, we perform an ablation study to examine the impact of motion content loss in our Im2Flow framework for recognition. Table 2 compares the motion stream performance of several variants of our model on PennAction dataset. We compare our model (trained on UCF-101) with one variant that completely removes the motion content loss and only uses the pixel L_2 loss ($\lambda = 0$); one variant that uses a ResNet-18 network that is not fine-tuned for action classification. We see that motion content loss helps to preserve high-level motion features, leading to better recognition performance (Ours vs. Only Pixel L2 Loss). We only have a slight drop even if we completely remove supervision from action labels of videos (Ours vs. Without Action Label Supervision). Therefore, our Im2Flow model has the capability to learn purely from unlabeled video.

References

[1] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2, 3

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1

[4] S. L. Pintea, J. C. van Gemert, and A. W. Smeulders. Déja vu. In *ECCV*, 2014. 1, 2

[5] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1

	EPE ↓	EPE-Canny	EPE-FG	DS ↑	DS-Canny	DS-FG	OS ↑	OS-Canny	OS-FG
XYMag Encoding [1]	2.588	2.951	3.230	0.078	0.072	0.070	0.671	0.661	0.669
XYZero Encoding [8]	2.274	2.604	3.016	0.073	0.069	0.061	0.668	0.662	0.666
Our Encoding	**2.210**	**2.533**	**2.936**	**0.143**	**0.135**	**0.137**	**0.699**	**0.692**	**0.696**

Table 3. Quantitative results of dense optical flow prediction of different encoding schemes on UCF-101. ↓ lower better, ↑ higher better.

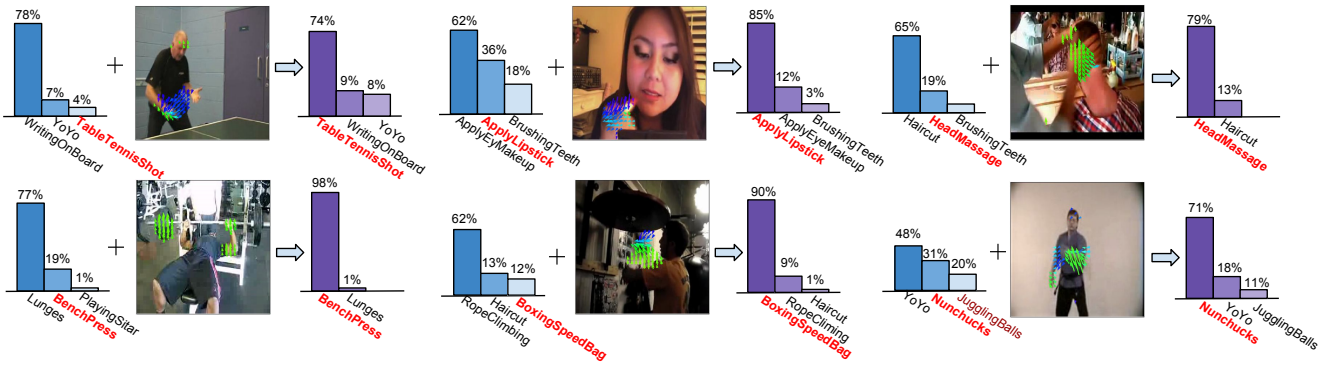

Figure 1. Examples of how the inferred motion can help static-image action recognition. For each example, the left shows the classification results of the appearance stream, and the right shows the two-stream results after incorporating the inferred motion.

[6] J. Walker, A. Gupta, and M. Hebert. Dense optical flow prediction from a static image. In *ICCV*, 2015. 1, 2

[7] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 1

[8] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015. 2, 3