In this supplementary material, we present additional experimental results and studies that are omitted in the main manuscript due to the lack of space.

1. Effect of the pose stream inputs

To analyze the effects of keypoint heatmaps and PAFs as inputs of the pose stream, we compare the top-1 and top-5 accuracy from pose-only models that take 1) the keypoint heatmaps, 2) the PAFs, and 3) both in Table 1. The table shows that taking both inputs achieves the best accuracy on Kinetics50 and Mimetics. The keypoint heatmaps provide the locations of each human body keypoint, which are useful in single person cases, but do not include sufficient information for differentiating each person in multi-person cases. On the other hand, the PAFs contain relationships between the keypoints from each person, which can provide information to differentiate each person in multi-person cases. We found that most of the videos in Mimetics contain a single person, which makes the heatmap-only model perform well on the action recognition. However, many videos in Kinetics contain multiple persons, and thus additional PAFs further improve the accuracy.

2. Deeper comparison with the score averaging

Most of the previous methods [1–5] use to simply average predicted action scores from the appearance-based and pose-based action recognition models in their testing stage. We compared their accuracy with ours in Table 4 of the main manuscript, and we provide a deeper comparison between ours and theirs in Figure 1. We report top-1 accuracy on Kinetics50 and Mimetics of score averaging with various averaging weights. As the figure shows, the score averaging method that performs best on Kinetics50 achieves slightly better accuracy than ours. However, it suffers from a noticeable performance drop on Mimetics. The proposed IntegralAction achieves highly robust performance on both Kinetics50 and Mimetics datasets.

3. Appearance-only, pose-only, vs. IntegralAction

In this experiment, we analyze top-1 accuracy of each action class using the appearance-only, pose-only, and the proposed IntegralAction on Kinetics50, Mimetics, and NTU-RGBD in Table 2, 3, and 4, respectively. In addition, we compare them with the oracle selection that chooses the best prediction between the appearance-only and the pose-only. We also visualize confusion matrices from the appearance-only, pose-only, and our IntegralAction in Fig. 2. As the tables and figures show, our IntegralAction produces robust action recognition over action classes of the three datasets, while the appearance-only and pose-only fail on Mimetics and Kinetics50, respectively. The proposed IntegralAction achieves the best average accuracy on Mimetics and NTU-RGBD. In addition, it significantly outperforms the pose-only and achieves comparable average accuracy with the appearance-only on Kinetics50.

Figures 3, 4, 5, 6, and 7 show qualitative results from the appearance-only, pose-only, and the proposed IntegralAction. Interestingly, our IntegralAction often succeeds in recognizing correct actions even when both the appearance-only and pose-only fail and so does the oracle.
selection, as shown in Fig. 7. We found that this happens when the appearance-only model is fooled by focusing on the contextual information such as background scene and objects, and the pose-only model suffers from the context ambiguity because the input pose sequence can be mapped to multiple action classes. For example, the second example of Fig. 7 shows that the appearance-only model is fooled by drinking people, and the pose-only model suffers from the context ambiguity. As the input pose sequence does not contain finger keypoints, the pose-only model predicts the input pose is about playing volleyball based on the given body keypoints. The input pose sequence may need to contain richer geometric information of the human body for better performance, for example, finger keypoints and finally, a 3D mesh of the human. Also, improving the integration part to more effectively combine the context from the appearance stream and the human motion from the pose stream should also be studied.

4. Network architecture of IntegralAction

In this section, we provide the detailed network architectures used in our paper. Table 5 shows the network architecture we used in Section 4.2 of the main manuscript, while Table 6 shows the network architecture we used in Section 4.3 and 4.4 of the main manuscript.
Table 2: The top-1 accuracy for each action comparison between appearance-only, pose-only, our IntegralAction, and the oracle selection on Kinetics50.

<table>
<thead>
<tr>
<th>classes</th>
<th>appearance-only</th>
<th>pose-only</th>
<th>IntegralAction (ours)</th>
<th>oracle selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>surfing water</td>
<td>87.5</td>
<td>25.0</td>
<td>91.7</td>
<td>87.5</td>
</tr>
<tr>
<td>shooting goal (soccer)</td>
<td><strong>51.0</strong></td>
<td>10.2</td>
<td>49.0</td>
<td>51.0</td>
</tr>
<tr>
<td>hitting baseball</td>
<td>84.0</td>
<td>54.0</td>
<td><strong>88.0</strong></td>
<td>88.0</td>
</tr>
<tr>
<td>playing bass guitar</td>
<td>86.0</td>
<td>44.0</td>
<td>78.0</td>
<td>88.0</td>
</tr>
<tr>
<td>reading book</td>
<td>66.0</td>
<td>38.0</td>
<td>60.0</td>
<td>70.0</td>
</tr>
<tr>
<td>juggling soccer ball</td>
<td>58.0</td>
<td>66.0</td>
<td><strong>70.0</strong></td>
<td>80.0</td>
</tr>
<tr>
<td>dribbling basketball</td>
<td>74.0</td>
<td>58.0</td>
<td>72.0</td>
<td>82.0</td>
</tr>
<tr>
<td>playing accordion</td>
<td><strong>91.8</strong></td>
<td>75.5</td>
<td>89.8</td>
<td>91.8</td>
</tr>
<tr>
<td>catching or throwing baseball</td>
<td>39.6</td>
<td>10.4</td>
<td><strong>41.7</strong></td>
<td>50.0</td>
</tr>
<tr>
<td>archery</td>
<td>77.6</td>
<td>30.6</td>
<td>75.5</td>
<td>79.6</td>
</tr>
<tr>
<td>tying tie</td>
<td><strong>86.0</strong></td>
<td>44.0</td>
<td><strong>86.0</strong></td>
<td>88.0</td>
</tr>
<tr>
<td>skiing (not slalom or crosscountry)</td>
<td><strong>95.9</strong></td>
<td>63.3</td>
<td><strong>95.9</strong></td>
<td>95.9</td>
</tr>
<tr>
<td>brushing hair</td>
<td>62.0</td>
<td>34.0</td>
<td>60.0</td>
<td>68.0</td>
</tr>
<tr>
<td>hurdling</td>
<td>92.0</td>
<td>66.0</td>
<td>90.0</td>
<td>94.0</td>
</tr>
<tr>
<td>playing violin</td>
<td>76.0</td>
<td>60.0</td>
<td>74.0</td>
<td>86.0</td>
</tr>
<tr>
<td>playing volleyball</td>
<td><strong>79.2</strong></td>
<td>45.8</td>
<td>77.1</td>
<td>81.2</td>
</tr>
<tr>
<td>deadlifting</td>
<td>87.8</td>
<td>87.8</td>
<td><strong>91.8</strong></td>
<td>93.9</td>
</tr>
<tr>
<td>skipping rope</td>
<td>67.3</td>
<td>77.6</td>
<td><strong>85.7</strong></td>
<td>85.7</td>
</tr>
<tr>
<td>playing piano</td>
<td>78.0</td>
<td>38.0</td>
<td>76.0</td>
<td>80.0</td>
</tr>
<tr>
<td>writing</td>
<td>72.0</td>
<td>22.0</td>
<td><strong>72.0</strong></td>
<td>74.0</td>
</tr>
<tr>
<td>climbing a rope</td>
<td>78.0</td>
<td>82.0</td>
<td>78.0</td>
<td>88.0</td>
</tr>
<tr>
<td>dunking basketball</td>
<td>56.2</td>
<td>41.7</td>
<td><strong>60.4</strong></td>
<td>68.8</td>
</tr>
<tr>
<td>playing basketball</td>
<td><strong>58.0</strong></td>
<td>32.0</td>
<td><strong>58.0</strong></td>
<td>66.0</td>
</tr>
<tr>
<td>brushing teeth</td>
<td>66.0</td>
<td>44.0</td>
<td><strong>68.0</strong></td>
<td>72.0</td>
</tr>
<tr>
<td>drinking</td>
<td>30.6</td>
<td>14.3</td>
<td><strong>32.7</strong></td>
<td>36.7</td>
</tr>
<tr>
<td>driving car</td>
<td><strong>91.7</strong></td>
<td>39.6</td>
<td>87.5</td>
<td>91.7</td>
</tr>
<tr>
<td>walking the dog</td>
<td><strong>93.9</strong></td>
<td>65.3</td>
<td>91.8</td>
<td><strong>95.9</strong></td>
</tr>
<tr>
<td>playing saxophone</td>
<td><strong>80.0</strong></td>
<td>54.0</td>
<td>78.0</td>
<td>86.0</td>
</tr>
<tr>
<td>playing piano</td>
<td><strong>83.7</strong></td>
<td>57.1</td>
<td><strong>83.7</strong></td>
<td>85.7</td>
</tr>
<tr>
<td>bowling</td>
<td><strong>93.9</strong></td>
<td>38.8</td>
<td>87.8</td>
<td><strong>93.9</strong></td>
</tr>
<tr>
<td>punching person (boxing)</td>
<td><strong>79.2</strong></td>
<td>60.4</td>
<td>75.0</td>
<td><strong>83.3</strong></td>
</tr>
<tr>
<td>cleaning windows</td>
<td><strong>76.0</strong></td>
<td>16.0</td>
<td><strong>76.0</strong></td>
<td>80.0</td>
</tr>
<tr>
<td>clean and jerk</td>
<td><strong>91.8</strong></td>
<td>87.8</td>
<td><strong>91.8</strong></td>
<td>93.9</td>
</tr>
<tr>
<td>eating cake</td>
<td><strong>58.0</strong></td>
<td>22.0</td>
<td>48.0</td>
<td>64.0</td>
</tr>
<tr>
<td>flying kite</td>
<td><strong>90.0</strong></td>
<td>54.0</td>
<td><strong>90.0</strong></td>
<td>96.0</td>
</tr>
<tr>
<td>opening bottle</td>
<td>52.0</td>
<td>18.0</td>
<td><strong>70.0</strong></td>
<td>58.0</td>
</tr>
<tr>
<td>canoeing or kayaking</td>
<td><strong>94.0</strong></td>
<td>38.0</td>
<td>90.0</td>
<td>94.0</td>
</tr>
<tr>
<td>reading newspaper</td>
<td><strong>54.0</strong></td>
<td>8.0</td>
<td>38.0</td>
<td><strong>54.0</strong></td>
</tr>
<tr>
<td>skiing slalom</td>
<td><strong>86.0</strong></td>
<td>76.0</td>
<td><strong>86.0</strong></td>
<td>92.0</td>
</tr>
<tr>
<td>playing guitar</td>
<td><strong>80.0</strong></td>
<td>56.0</td>
<td>76.0</td>
<td>84.0</td>
</tr>
<tr>
<td>eating ice cream</td>
<td><strong>54.0</strong></td>
<td>20.0</td>
<td>46.0</td>
<td>66.0</td>
</tr>
<tr>
<td>climbing ladder</td>
<td>68.0</td>
<td>40.0</td>
<td><strong>72.0</strong></td>
<td>74.0</td>
</tr>
<tr>
<td>juggling balls</td>
<td>81.6</td>
<td>79.6</td>
<td><strong>85.7</strong></td>
<td>91.8</td>
</tr>
<tr>
<td>shooting basketball</td>
<td><strong>30.6</strong></td>
<td>16.3</td>
<td>22.4</td>
<td>40.8</td>
</tr>
<tr>
<td>catching or throwing frisbee</td>
<td><strong>56.0</strong></td>
<td>8.0</td>
<td>48.0</td>
<td><strong>58.0</strong></td>
</tr>
<tr>
<td>sweeping floor</td>
<td><strong>72.0</strong></td>
<td>42.0</td>
<td><strong>72.0</strong></td>
<td>76.0</td>
</tr>
<tr>
<td>playing tennis</td>
<td><strong>93.9</strong></td>
<td>69.4</td>
<td><strong>93.9</strong></td>
<td><strong>98.0</strong></td>
</tr>
<tr>
<td>sword fighting</td>
<td>38.8</td>
<td>32.7</td>
<td><strong>40.8</strong></td>
<td>53.1</td>
</tr>
<tr>
<td>smoking</td>
<td><strong>55.1</strong></td>
<td>36.7</td>
<td>53.1</td>
<td>65.3</td>
</tr>
<tr>
<td>golf driving</td>
<td><strong>84.0</strong></td>
<td>80.0</td>
<td><strong>86.0</strong></td>
<td><strong>86.0</strong></td>
</tr>
<tr>
<td>average</td>
<td><strong>72.8</strong></td>
<td>45.6</td>
<td><strong>72.2</strong></td>
<td><strong>78.2</strong></td>
</tr>
</tbody>
</table>
Table 3: The top-1 accuracy for each action comparison between appearance-only, pose-only, our IntegralAction, and the oracle selection on Mimetics.

<table>
<thead>
<tr>
<th>classes</th>
<th>appearance-only</th>
<th>pose-only</th>
<th>IntegralAction (ours)</th>
<th>oracle selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>surfing water</td>
<td>0.0</td>
<td>0.0</td>
<td>12.5</td>
<td>0.0</td>
</tr>
<tr>
<td>shooting goal (soccer)</td>
<td>18.2</td>
<td>9.1</td>
<td>18.2</td>
<td>27.3</td>
</tr>
<tr>
<td>hitting baseball</td>
<td>0.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
</tr>
<tr>
<td>playing bass guitar</td>
<td>9.1</td>
<td>18.2</td>
<td>27.3</td>
<td>27.3</td>
</tr>
<tr>
<td>reading book</td>
<td>11.1</td>
<td>11.1</td>
<td>0.0</td>
<td>22.2</td>
</tr>
<tr>
<td>juggling soccer ball</td>
<td>16.7</td>
<td>41.7</td>
<td>33.3</td>
<td>41.7</td>
</tr>
<tr>
<td>dribbling basketball</td>
<td>0.0</td>
<td>61.5</td>
<td>23.1</td>
<td>61.5</td>
</tr>
<tr>
<td>playing accordion</td>
<td>10.0</td>
<td>40.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>catching or throwing baseball</td>
<td>14.3</td>
<td>14.3</td>
<td>0.0</td>
<td>28.6</td>
</tr>
<tr>
<td>archery</td>
<td>6.7</td>
<td>33.3</td>
<td>20.0</td>
<td>33.3</td>
</tr>
<tr>
<td>tying tie</td>
<td>16.7</td>
<td>16.7</td>
<td>16.7</td>
<td>16.7</td>
</tr>
<tr>
<td>skiing (not slalom or crosscountry)</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
<td>12.5</td>
</tr>
<tr>
<td>brushing hair</td>
<td>11.8</td>
<td>41.2</td>
<td>47.1</td>
<td>41.2</td>
</tr>
<tr>
<td>hurdling</td>
<td>11.1</td>
<td>55.6</td>
<td>22.2</td>
<td>55.6</td>
</tr>
<tr>
<td>playing violin</td>
<td>11.8</td>
<td>17.6</td>
<td>23.5</td>
<td>17.6</td>
</tr>
<tr>
<td>playing volleyball</td>
<td>23.1</td>
<td>23.1</td>
<td>38.5</td>
<td>38.5</td>
</tr>
<tr>
<td>deadlifting</td>
<td>11.1</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>skipping rope</td>
<td>25.0</td>
<td>83.3</td>
<td>75.0</td>
<td>83.3</td>
</tr>
<tr>
<td>playing piano</td>
<td>11.8</td>
<td>11.8</td>
<td>11.8</td>
<td>17.6</td>
</tr>
<tr>
<td>writing</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>climbing a rope</td>
<td>0.0</td>
<td>42.9</td>
<td>50.0</td>
<td>42.9</td>
</tr>
<tr>
<td>dunking basketball</td>
<td>0.0</td>
<td>33.3</td>
<td>33.3</td>
<td>33.3</td>
</tr>
<tr>
<td>playing basketball</td>
<td>23.1</td>
<td>23.1</td>
<td>30.8</td>
<td>30.8</td>
</tr>
<tr>
<td>brushing teeth</td>
<td>35.7</td>
<td>35.7</td>
<td>42.9</td>
<td>50.0</td>
</tr>
<tr>
<td>drinking</td>
<td>10.0</td>
<td>20.0</td>
<td>30.0</td>
<td>25.0</td>
</tr>
<tr>
<td>driving car</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>walking the dog</td>
<td>7.7</td>
<td>7.7</td>
<td>7.7</td>
<td>7.7</td>
</tr>
<tr>
<td>playing saxophone</td>
<td>0.0</td>
<td>7.7</td>
<td>15.4</td>
<td>7.7</td>
</tr>
<tr>
<td>playing trumpet</td>
<td>0.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>bowling</td>
<td>8.3</td>
<td>0.0</td>
<td>8.3</td>
<td>8.3</td>
</tr>
<tr>
<td>punching person (boxing)</td>
<td>9.1</td>
<td>36.4</td>
<td>45.5</td>
<td>36.4</td>
</tr>
<tr>
<td>cleaning windows</td>
<td>6.7</td>
<td>0.0</td>
<td>0.0</td>
<td>6.7</td>
</tr>
<tr>
<td>clean and jerk</td>
<td>15.4</td>
<td>92.3</td>
<td>92.3</td>
<td>92.3</td>
</tr>
<tr>
<td>eating cake</td>
<td>0.0</td>
<td>11.8</td>
<td>0.0</td>
<td>11.8</td>
</tr>
<tr>
<td>flying kite</td>
<td>10.0</td>
<td>0.0</td>
<td>20.0</td>
<td>10.0</td>
</tr>
<tr>
<td>opening bottle</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>canoeing or kayaking</td>
<td>0.0</td>
<td>21.4</td>
<td>14.3</td>
<td>21.4</td>
</tr>
<tr>
<td>reading newspaper</td>
<td>0.0</td>
<td>11.1</td>
<td>0.0</td>
<td>11.1</td>
</tr>
<tr>
<td>skiing slalom</td>
<td>0.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
</tr>
<tr>
<td>playing guitar</td>
<td>7.1</td>
<td>14.3</td>
<td>14.3</td>
<td>14.3</td>
</tr>
<tr>
<td>eating ice cream</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>climbing ladder</td>
<td>7.7</td>
<td>15.4</td>
<td>7.7</td>
<td>23.1</td>
</tr>
<tr>
<td>juggling balls</td>
<td>42.9</td>
<td>57.1</td>
<td>57.1</td>
<td>71.4</td>
</tr>
<tr>
<td>shooting basketball</td>
<td>8.3</td>
<td>8.3</td>
<td>16.7</td>
<td>16.7</td>
</tr>
<tr>
<td>catching or throwing frisbee</td>
<td>50.0</td>
<td>0.0</td>
<td>40.0</td>
<td>50.0</td>
</tr>
<tr>
<td>sweeping floor</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>playing tennis</td>
<td>5.6</td>
<td>16.7</td>
<td>11.1</td>
<td>22.2</td>
</tr>
<tr>
<td>sword fighting</td>
<td>46.7</td>
<td>66.7</td>
<td>66.7</td>
<td>73.3</td>
</tr>
<tr>
<td>smoking</td>
<td>26.7</td>
<td>20.0</td>
<td>33.3</td>
<td>33.3</td>
</tr>
<tr>
<td>golf driving</td>
<td>14.3</td>
<td>64.3</td>
<td>50.0</td>
<td>64.3</td>
</tr>
<tr>
<td>average</td>
<td>11.2</td>
<td>26.0</td>
<td>26.5</td>
<td>30.7</td>
</tr>
</tbody>
</table>
Table 4: The top-1 accuracy for each action comparison between appearance-only, pose-only, our IntegralAction, and the oracle selection on NTU-RGBD.

<table>
<thead>
<tr>
<th>classes</th>
<th>appearance-only</th>
<th>pose-only</th>
<th>IntegralAction (ours)</th>
<th>oracle selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>drink water</td>
<td>90.9</td>
<td>82.8</td>
<td>92.0</td>
<td>96.4</td>
</tr>
<tr>
<td>eat meal</td>
<td>80.0</td>
<td>66.9</td>
<td>82.5</td>
<td>82.9</td>
</tr>
<tr>
<td>brush teeth</td>
<td>92.3</td>
<td>80.1</td>
<td>91.5</td>
<td>95.6</td>
</tr>
<tr>
<td>brush hair</td>
<td>95.2</td>
<td>87.5</td>
<td>96.7</td>
<td>97.4</td>
</tr>
<tr>
<td>drop</td>
<td>97.1</td>
<td>78.5</td>
<td>97.1</td>
<td>98.5</td>
</tr>
<tr>
<td>pick up</td>
<td>96.7</td>
<td>93.8</td>
<td>98.2</td>
<td>98.2</td>
</tr>
<tr>
<td>throw</td>
<td>90.2</td>
<td>83.3</td>
<td>90.2</td>
<td>94.2</td>
</tr>
<tr>
<td>sit down</td>
<td>95.6</td>
<td>96.0</td>
<td>98.5</td>
<td>98.9</td>
</tr>
<tr>
<td>stand up</td>
<td>98.5</td>
<td>96.7</td>
<td><strong>99.3</strong></td>
<td>99.6</td>
</tr>
<tr>
<td>clapping</td>
<td>79.5</td>
<td>64.8</td>
<td>78.8</td>
<td>89.4</td>
</tr>
<tr>
<td>reading</td>
<td>73.9</td>
<td>48.2</td>
<td>70.6</td>
<td>82.0</td>
</tr>
<tr>
<td>writing</td>
<td>64.3</td>
<td>40.4</td>
<td>64.7</td>
<td>77.2</td>
</tr>
<tr>
<td>tear up paper</td>
<td>93.0</td>
<td>80.8</td>
<td><strong>93.7</strong></td>
<td>95.9</td>
</tr>
<tr>
<td>put on jacket</td>
<td><strong>100.0</strong></td>
<td>97.8</td>
<td>99.6</td>
<td>100.0</td>
</tr>
<tr>
<td>take off jacket</td>
<td>98.2</td>
<td>94.6</td>
<td>96.4</td>
<td>98.9</td>
</tr>
<tr>
<td>put on a shoe</td>
<td>90.8</td>
<td>58.6</td>
<td>83.2</td>
<td>92.7</td>
</tr>
<tr>
<td>take off a shoe</td>
<td>82.8</td>
<td>55.8</td>
<td>73.5</td>
<td>89.8</td>
</tr>
<tr>
<td>put on glasses</td>
<td>85.7</td>
<td>85.3</td>
<td><strong>92.6</strong></td>
<td>96.0</td>
</tr>
<tr>
<td>take off glasses</td>
<td>89.8</td>
<td>86.1</td>
<td>92.3</td>
<td>93.4</td>
</tr>
<tr>
<td>put on a hat/cap</td>
<td>99.6</td>
<td>92.3</td>
<td>98.2</td>
<td>100.0</td>
</tr>
<tr>
<td>take off a hat/cap</td>
<td>98.2</td>
<td>94.1</td>
<td><strong>98.5</strong></td>
<td>99.3</td>
</tr>
<tr>
<td>cheer up</td>
<td>94.5</td>
<td>90.9</td>
<td>93.8</td>
<td>96.7</td>
</tr>
<tr>
<td>hand waving</td>
<td>87.2</td>
<td>87.2</td>
<td>90.9</td>
<td><strong>93.4</strong></td>
</tr>
<tr>
<td>kicking something</td>
<td>98.2</td>
<td>93.1</td>
<td><strong>98.2</strong></td>
<td>99.3</td>
</tr>
<tr>
<td>reach into pocket</td>
<td>81.8</td>
<td>72.6</td>
<td>81.4</td>
<td>85.8</td>
</tr>
<tr>
<td>hopping</td>
<td>94.9</td>
<td>96.4</td>
<td><strong>96.4</strong></td>
<td>96.7</td>
</tr>
<tr>
<td>jump up</td>
<td><strong>100.0</strong></td>
<td>99.6</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>phone call</td>
<td>89.1</td>
<td>78.9</td>
<td>78.2</td>
<td>94.9</td>
</tr>
<tr>
<td>play with phone/tablet</td>
<td>70.2</td>
<td>58.2</td>
<td>76.4</td>
<td>81.5</td>
</tr>
<tr>
<td>type on a keyboard</td>
<td>88.4</td>
<td>66.5</td>
<td>89.8</td>
<td>93.8</td>
</tr>
<tr>
<td>point to something</td>
<td>88.8</td>
<td>75.7</td>
<td><strong>90.6</strong></td>
<td>92.4</td>
</tr>
<tr>
<td>taking a selfie</td>
<td>88.0</td>
<td>85.5</td>
<td><strong>93.1</strong></td>
<td>94.6</td>
</tr>
<tr>
<td>check time (from watch)</td>
<td>88.4</td>
<td>84.8</td>
<td><strong>92.8</strong></td>
<td>97.1</td>
</tr>
<tr>
<td>rub two hands</td>
<td>72.5</td>
<td>75.4</td>
<td>79.7</td>
<td>91.7</td>
</tr>
<tr>
<td>nod head/bow</td>
<td>93.8</td>
<td>90.2</td>
<td>95.7</td>
<td>97.1</td>
</tr>
<tr>
<td>shake head</td>
<td>94.9</td>
<td>83.2</td>
<td><strong>98.2</strong></td>
<td>97.8</td>
</tr>
<tr>
<td>wipe face</td>
<td>83.7</td>
<td>76.4</td>
<td>91.7</td>
<td>95.7</td>
</tr>
<tr>
<td>salute</td>
<td>95.7</td>
<td>90.6</td>
<td><strong>95.7</strong></td>
<td>97.8</td>
</tr>
<tr>
<td>put palms together</td>
<td>82.2</td>
<td>88.8</td>
<td>92.4</td>
<td>95.7</td>
</tr>
<tr>
<td>cross hands in front</td>
<td>96.0</td>
<td>94.9</td>
<td>97.5</td>
<td>97.8</td>
</tr>
<tr>
<td>sneeze/cough</td>
<td>68.5</td>
<td>69.6</td>
<td>77.2</td>
<td>80.4</td>
</tr>
<tr>
<td>staggering</td>
<td>98.6</td>
<td>96.7</td>
<td><strong>98.9</strong></td>
<td>99.6</td>
</tr>
<tr>
<td>falling down</td>
<td>98.2</td>
<td>96.0</td>
<td><strong>98.9</strong></td>
<td>98.5</td>
</tr>
<tr>
<td>headache</td>
<td>68.5</td>
<td>68.1</td>
<td><strong>75.0</strong></td>
<td>84.1</td>
</tr>
<tr>
<td>chest pain</td>
<td>88.0</td>
<td>88.0</td>
<td>92.0</td>
<td>96.4</td>
</tr>
<tr>
<td>back pain</td>
<td>95.3</td>
<td>85.5</td>
<td><strong>97.1</strong></td>
<td>98.9</td>
</tr>
<tr>
<td>neck pain</td>
<td>83.7</td>
<td>77.2</td>
<td><strong>89.1</strong></td>
<td>92.8</td>
</tr>
<tr>
<td>nausea/vomiting</td>
<td>87.6</td>
<td>83.6</td>
<td><strong>90.5</strong></td>
<td>93.1</td>
</tr>
<tr>
<td>fan self</td>
<td>86.5</td>
<td>86.9</td>
<td><strong>88.4</strong></td>
<td>96.0</td>
</tr>
<tr>
<td>punch/slap</td>
<td>88.3</td>
<td>89.1</td>
<td><strong>92.0</strong></td>
<td>95.3</td>
</tr>
<tr>
<td>kidney</td>
<td>98.6</td>
<td>91.7</td>
<td><strong>97.8</strong></td>
<td>99.3</td>
</tr>
<tr>
<td>pushing</td>
<td>97.5</td>
<td>94.9</td>
<td><strong>98.5</strong></td>
<td>99.6</td>
</tr>
<tr>
<td>pat on back</td>
<td>97.8</td>
<td>85.5</td>
<td><strong>98.2</strong></td>
<td>99.3</td>
</tr>
<tr>
<td>point finger</td>
<td>97.5</td>
<td>89.1</td>
<td><strong>97.8</strong></td>
<td>98.6</td>
</tr>
<tr>
<td>hugging</td>
<td>99.3</td>
<td>97.4</td>
<td><strong>99.6</strong></td>
<td>100.0</td>
</tr>
<tr>
<td>giving object</td>
<td><strong>95.3</strong></td>
<td>87.3</td>
<td>94.2</td>
<td>97.1</td>
</tr>
<tr>
<td>touch pocket</td>
<td><strong>96.7</strong></td>
<td>92.0</td>
<td>96.0</td>
<td>99.3</td>
</tr>
<tr>
<td>shaking hands</td>
<td>98.2</td>
<td>94.6</td>
<td><strong>100.0</strong></td>
<td>99.6</td>
</tr>
<tr>
<td>walking towards</td>
<td><strong>100.0</strong></td>
<td>98.9</td>
<td><strong>100.0</strong></td>
<td>100.0</td>
</tr>
<tr>
<td>walking apart</td>
<td><strong>100.0</strong></td>
<td>97.1</td>
<td>98.9</td>
<td>100.0</td>
</tr>
<tr>
<td>average</td>
<td>90.4</td>
<td>83.7</td>
<td><strong>91.7</strong></td>
<td>95.1</td>
</tr>
</tbody>
</table>
Figure 2: Visualized confusion matrices of appearance-only, pose-only, and our IntegralAction on Kinetics50, Mimetics, and NTU-RGBD.
Figure 3: Qualitative results of appearance-only, pose-only, and the proposed IntegralAction.

Figure 4: Qualitative results of appearance-only, pose-only, and the proposed IntegralAction.
Figure 5: Qualitative results of appearance-only, pose-only, and the proposed IntegralAction.

Figure 6: Qualitative results of appearance-only, pose-only, and the proposed IntegralAction.
Figure 7: Qualitative results of appearance-only, pose-only, and the proposed IntegralAction.
Table 5: The network architecture details of IntegralAction in Section 4.2 of the main manuscript. The dimensions of kernels are denoted by \((T \times S^2, C)\) for the temporal, spatial, and channel sizes. The strides and output size are denoted by \((T \times S^2)\) for the temporal and spatial sizes.

<table>
<thead>
<tr>
<th>layers</th>
<th>appearance stream</th>
<th>pose stream</th>
<th>output size</th>
</tr>
</thead>
<tbody>
<tr>
<td>input</td>
<td>RGB frames</td>
<td>keypoint heatmaps+PAFs</td>
<td>appearance: (8 \times 224^2) pose: (32 \times 56^2)</td>
</tr>
<tr>
<td>conv1</td>
<td>(1 \times 7^2, 64, \text{stride} \ 1 \times 2^2)</td>
<td>(1 \times 3^2, 64)</td>
<td>appearance: (8 \times 112^2) pose: (32 \times 56^2)</td>
</tr>
<tr>
<td></td>
<td>(1 \times 3^2 \text{ max pool, stride} \ 1 \times 2^2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>res2</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 1^2, 256 \ 1 \times 3^2, 256 \ 1 \times 1^2, 256 \end{bmatrix} \times 3)</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 3^2, 64 \ 1 \times 3^2, 64 \end{bmatrix} \times 2)</td>
<td>appearance: (8 \times 56^2) pose: (32 \times 56^2)</td>
</tr>
<tr>
<td>res3</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 1^2, 512 \ 1 \times 3^2, 512 \ 1 \times 1^2, 512 \end{bmatrix} \times 4)</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 3^2, 128 \ 1 \times 3^2, 128 \end{bmatrix} \times 2)</td>
<td>appearance: (8 \times 28^2) pose: (32 \times 28^2)</td>
</tr>
<tr>
<td>res4</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 1^2, 1024 \ 1 \times 3^2, 1024 \ 1 \times 1^2, 1024 \end{bmatrix} \times 6)</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 3^2, 256 \ 1 \times 3^2, 256 \end{bmatrix} \times 2)</td>
<td>appearance: (8 \times 14^2) pose: (32 \times 14^2)</td>
</tr>
<tr>
<td>res5</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 1^2, 2048 \ 1 \times 3^2, 2048 \ 1 \times 1^2, 2048 \end{bmatrix} \times 3)</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 3^2, 512 \ 1 \times 3^2, 512 \end{bmatrix} \times 2)</td>
<td>appearance: (8 \times 7^2) pose: (32 \times 7^2)</td>
</tr>
<tr>
<td>pool</td>
<td>global average pool</td>
<td>global average pool</td>
<td>appearance: (8 \times 1^2) pose: (32 \times 1^2)</td>
</tr>
<tr>
<td>feature align (TCB(_A),TCB(_P))</td>
<td>(1 \times 1^2, 512) layer normalization</td>
<td>(4 \times 1^2 \text{ avg pool, stride} \ 4 \times 1^2) (1 \times 1^2, 512) layer normalization</td>
<td>both: (8 \times 1^2)</td>
</tr>
<tr>
<td>pose-driven gating (CGB)</td>
<td>((1 - G) \text{ element-wise product})</td>
<td>((G :1 \times 1^2, 512)) (G \text{ element-wise product})</td>
<td></td>
</tr>
<tr>
<td>aggregation</td>
<td>element-wise addition</td>
<td>classifier</td>
<td>fully-connected layer</td>
</tr>
</tbody>
</table>
Table 6: The network architecture details of IntegralAction in Section 4.3 and 4.4 of the main manuscript. The dimensions of kernels are denoted by \((T \times S^2, C)\) for the temporal, spatial, and channel sizes. The strides and output size are denoted by \((T \times S^2)\) for the temporal and spatial sizes.

<table>
<thead>
<tr>
<th>layers</th>
<th>appearance stream</th>
<th>pose stream</th>
<th>output size</th>
</tr>
</thead>
<tbody>
<tr>
<td>input</td>
<td>RGB frames</td>
<td>keypoint heatmaps+PAFs</td>
<td>appearance: (8 \times 224^2) pose: (8 \times 56^2)</td>
</tr>
<tr>
<td>conv_1</td>
<td>(1 \times 7^2, 64, \text{stride } 1 \times 2^2)</td>
<td>(1 \times 3^2, 64)</td>
<td>appearance: (8 \times 112^2) pose: (8 \times 56^2)</td>
</tr>
<tr>
<td>res_2</td>
<td>(1 \times 3^2) max pool, stride (1 \times 2^2)</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 3^2, 64 \ 1 \times 3^2, 64 \end{bmatrix} \times 2)</td>
<td>both: (8 \times 56^2)</td>
</tr>
<tr>
<td>res_3</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 3^2, 128 \ 1 \times 3^2, 128 \end{bmatrix} \times 2)</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 3^2, 128 \ 1 \times 3^2, 128 \end{bmatrix} \times 2)</td>
<td>both: (8 \times 28^2)</td>
</tr>
<tr>
<td>res_4</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 3^2, 256 \ 1 \times 3^2, 256 \end{bmatrix} \times 2)</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 3^2, 256 \ 1 \times 3^2, 256 \end{bmatrix} \times 2)</td>
<td>both: (8 \times 14^2)</td>
</tr>
<tr>
<td>res_5</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 3^2, 512 \ 1 \times 3^2, 512 \end{bmatrix} \times 2)</td>
<td>(\begin{bmatrix} \text{TSM} \ 1 \times 3^2, 512 \ 1 \times 3^2, 512 \end{bmatrix} \times 2)</td>
<td>both: (8 \times 7^2)</td>
</tr>
<tr>
<td>pool</td>
<td>global average pool</td>
<td>global average pool</td>
<td></td>
</tr>
<tr>
<td>feature align (TCB_A,TCB_B)</td>
<td>(1 \times 1^2, 512)</td>
<td>(1 \times 1^2, 512)</td>
<td>both: (8 \times 1^2)</td>
</tr>
<tr>
<td>pose-driven gating (CGB)</td>
<td>element-wise product</td>
<td>((G : 1 \times 1 \times 1, 512)) G element-wise product</td>
<td></td>
</tr>
<tr>
<td>aggregation</td>
<td>element-wise addition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>classifier</td>
<td>fully-connected layer</td>
<td></td>
<td># of classes</td>
</tr>
</tbody>
</table>
References


