Supplementary Material *of* "Pose-driven Feature Integration for Robust Human Action Recognition in Videos"

Gyeongsik Moon*[1] Heeseung Kwon*[2] Kyoung Mu Lee[1] Minsu Cho[2]

[1]SNU ECE & ASRI
{mks0601, kyoungmu}@snu.ac.kr

[2]POSTECH CSE & AIGS
{aruno, mscho}@postech.ac.kr

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.

Table 1: Top-1 and top-5 accuracy comparison between pose-only models that take various combinations of input on Kinetics50 and Mimetics.

settings	Kinetics50		Mimetics	
	top-1	top-5	top-1	top-5
heatmap-only	43.5	72.1	**26.0**	50.0
PAF-only	45.0	**73.2**	25.0	49.8
heatmap + PAF (ours)	**45.6**	72.9	**26.0**	**52.2**

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

* equal contribution

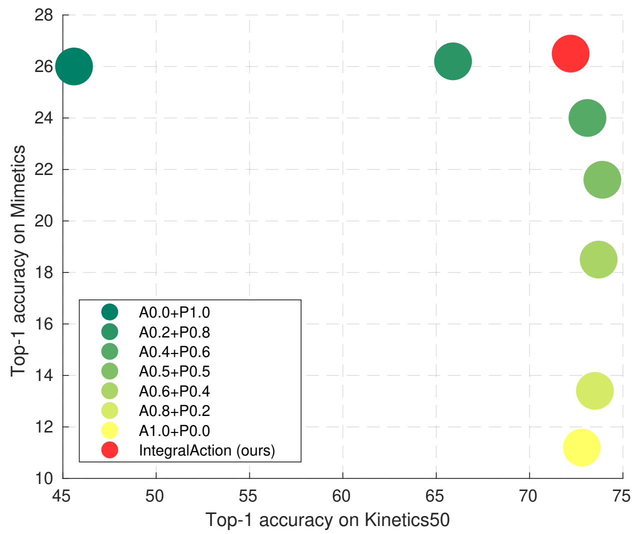

Figure 1: Top-1 accuracy on Kinetics50 and Mimetics comparison between the proposed IntegralAction and score averaging with various average ratios. The numbers to the left and right of plus sign denote averaging weight at the score from the appearance-based and pose-based models, respectively.

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.

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

Table 3: The top-1 accuracy for each action comparison between appearance-only, pose-only, our IntegralAction, and the oracle selection on Mimetics.

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

Table 4: The top-1 accuracy for each action comparison between appearance-only, pose-only, our IntegralAction, and the oracle selection on NTU-RGBD.

classes	appearance-only	pose-only	**IntegralAction (ours)**	oracle selection
drink water	90.9	82.8	**92.0**	96.4
eat meal	80.0	66.9	**82.5**	82.9
brush teeth	**92.3**	80.1	91.5	95.6
brush hair	95.2	87.5	**96.7**	97.4
drop	**97.1**	78.5	**97.1**	98.5
pick up	96.7	93.8	**98.2**	98.2
throw	**90.2**	83.3	**90.2**	94.2
sit down	95.6	96.0	**98.5**	98.9
stand up	98.5	96.7	**99.3**	99.6
clapping	**79.5**	64.8	78.8	89.4
reading	**73.9**	48.2	70.6	82.0
writing	64.3	40.4	**64.7**	77.2
tear up paper	93.0	80.8	**93.7**	95.9
put on jacket	**100.0**	97.8	99.6	100.0
take off jacket	**98.2**	94.6	96.4	98.9
put on a shoe	**90.8**	58.6	83.2	92.7
take off a shoe	**82.8**	55.8	75.5	89.8
put on glasses	85.7	85.3	**92.6**	96.0
take off glasses	89.8	86.1	**92.3**	93.4
put on a hat/cap	**99.6**	92.3	98.2	100.0
take off a hat/cap	98.2	94.1	**98.5**	99.3
cheer up	**94.5**	90.9	93.8	96.7
hand waving	87.2	87.2	**90.9**	93.4
kicking something	**98.2**	93.1	**98.2**	99.3
reach into pocket	**81.8**	72.6	81.4	85.8
hopping	94.9	**96.4**	**96.4**	96.7
jump up	**100.0**	99.6	**100.0**	100.0
phone call	**89.1**	78.9	78.2	94.9
play with phone/tablet	70.2	58.2	**76.4**	81.5
type on a keyboard	88.4	66.5	**89.8**	93.8
point to something	88.8	75.7	**90.6**	92.4
taking a selfie	88.0	85.5	**93.1**	94.6
check time (from watch)	88.4	84.8	**92.8**	97.1
rub two hands	72.5	75.4	**79.7**	91.7
nod head/bow	93.8	90.2	**95.7**	97.1
shake head	94.9	83.2	**98.2**	97.8
wipe face	83.7	76.4	**91.7**	95.7
salute	**95.7**	90.6	**95.7**	97.8
put palms together	82.2	88.8	**92.4**	95.7
cross hands in front	96.0	94.9	**97.5**	97.8
sneeze/cough	68.5	69.6	**77.2**	80.4
staggering	98.6	96.7	**98.9**	99.6
falling down	98.2	96.0	**98.9**	98.5
headache	68.5	68.1	**75.0**	84.1
chest pain	88.0	88.0	**92.0**	96.4
back pain	95.3	85.5	**97.1**	98.9
neck pain	83.7	77.2	**89.1**	92.8
nausea/vomiting	87.6	83.6	**90.5**	93.1
fan self	86.5	86.9	**88.4**	96.0
punch/slap	88.3	89.1	**92.0**	95.3
kicking	**98.6**	91.7	97.8	99.3
pushing	97.5	94.9	**98.5**	99.6
pat on back	97.8	85.5	**98.2**	99.3
point finger	97.5	89.1	**97.8**	98.6
hugging	99.3	97.4	**99.6**	100.0
giving object	**95.3**	87.3	94.2	97.1
touch pocket	**96.7**	92.0	96.0	99.3
shaking hands	98.2	94.6	**100.0**	99.6
walking towards	**100.0**	98.9	**100.0**	100.0
walking apart	**100.0**	97.1	98.9	100.0
average	90.4	83.7	**91.7**	95.1

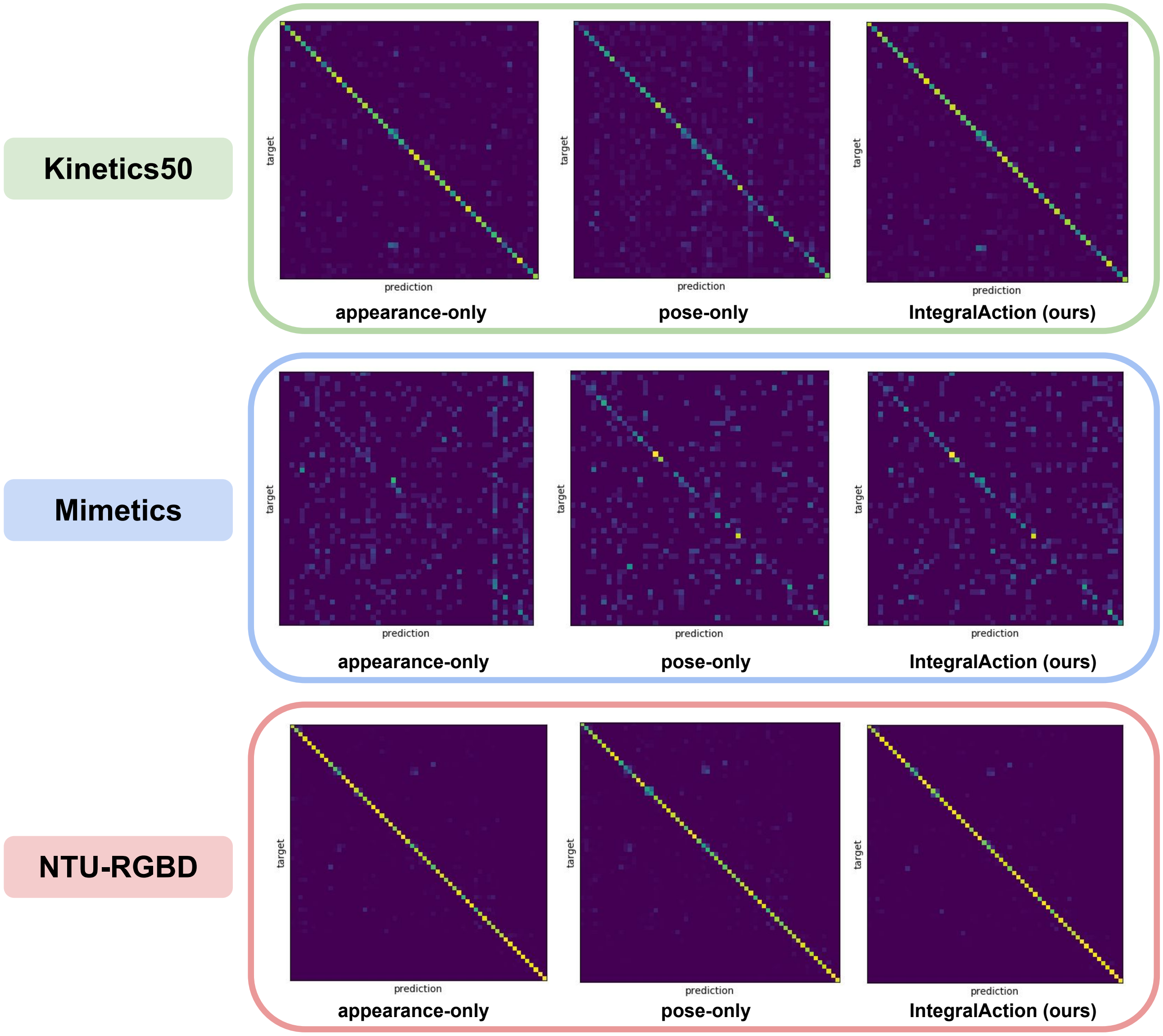

Figure 2: Visualized confusion matrices of appearance-only, pose-only, and our IntegralAction on Kinetisc50, Mimetics, and NTU-RGBD.

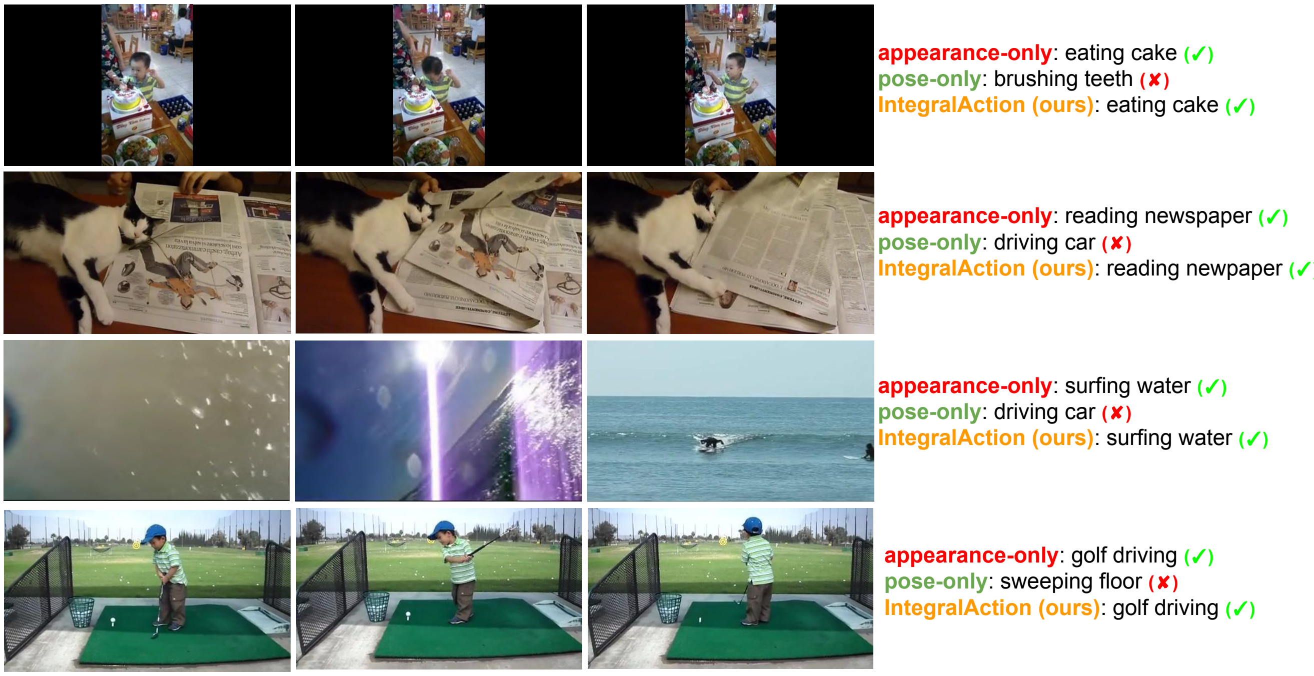

Figure 3: Qualitative results of appearance-only, pose-only, and the proposed IntegralAction.

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

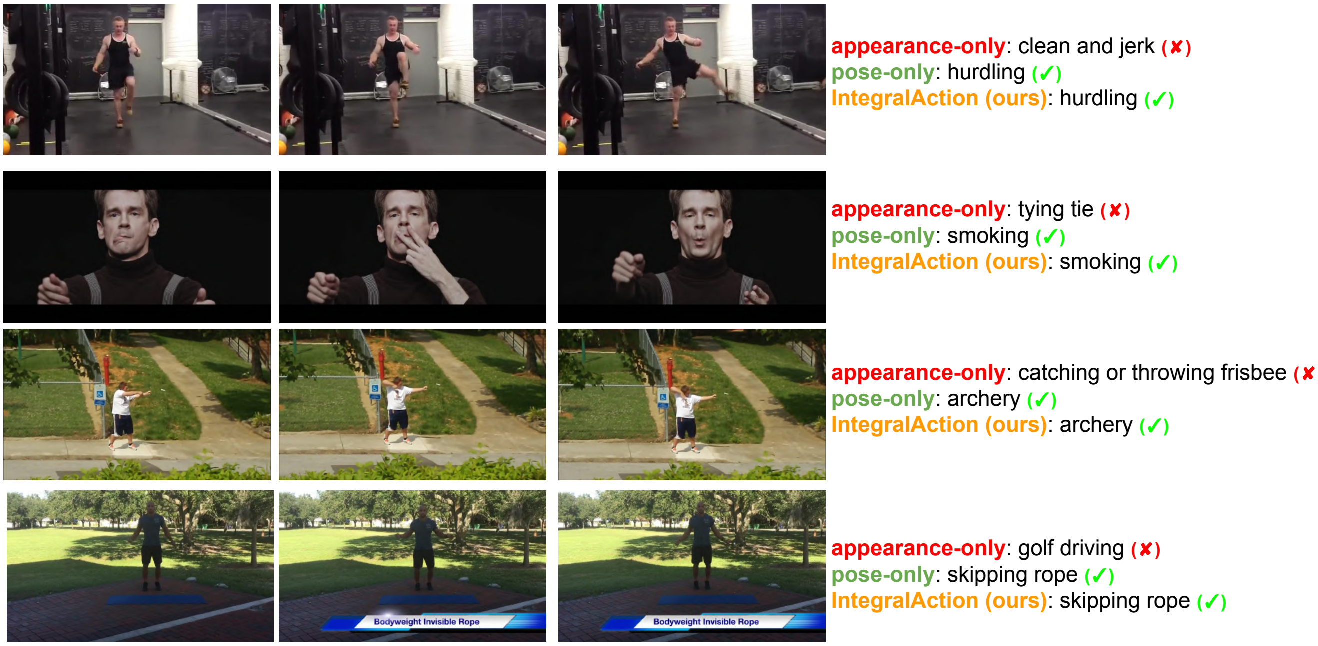

Figure 5: Qualitative results of appearance-only, pose-only, and the proposed IntegralAction.

Figure 6: Qualitative results of appearance-only, pose-only, and the proposed IntegralAction.

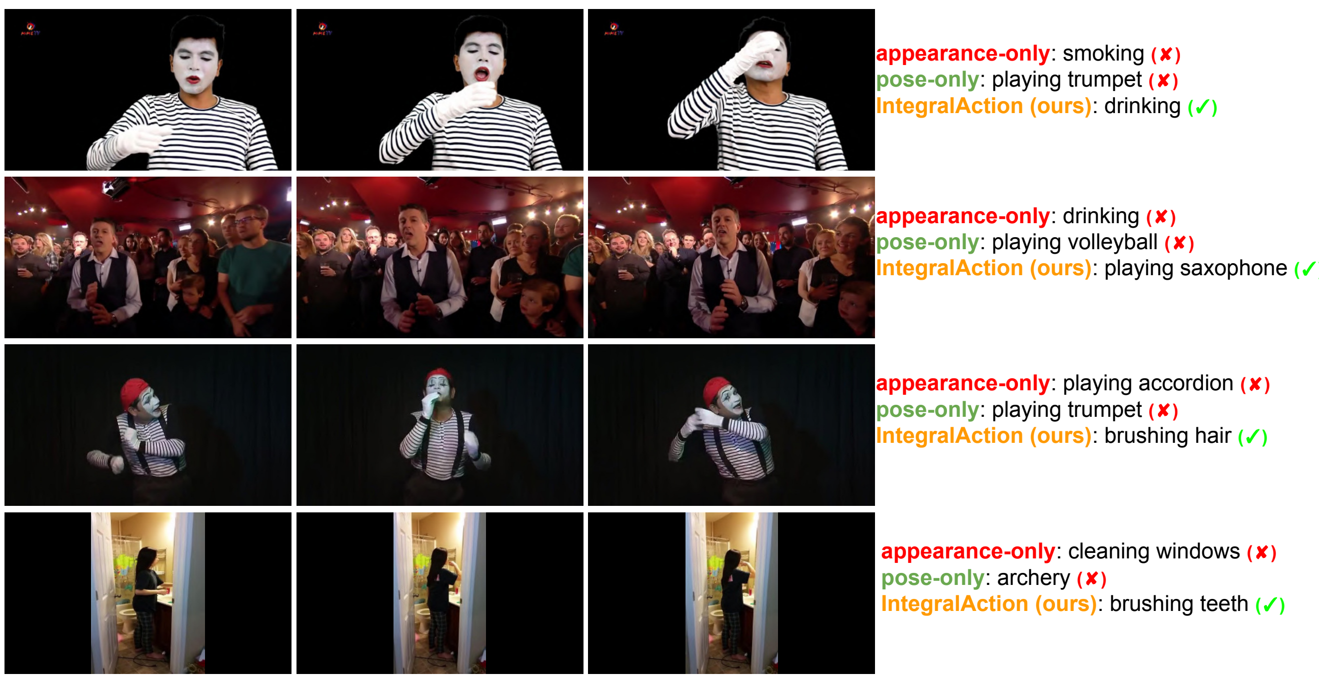

Figure 7: Qualitative results of appearance-only, pose-only, and the proposed IntegralAction.

Table 5: The network architecture details of IntegralAction in Section 4.2 of the main manuscript. The dimensions of kernels are denoted by ($\mathrm{T}\times\mathrm{S}^2$, C) for the temporal, spatial, and channel sizes. The strides and output size are denoted by ($\mathrm{T}\times\mathrm{S}^2$) for the temporal and spatial sizes.

layers	appearance stream	pose stream	output size
input	RGB frames	keypoint heatmaps+PAFs	appearance: 8×224^2 pose: 32×56^2
conv_1	1×7^2, 64, stride 1×2^2	1×3^2, 64	appearance: 8×112^2 pose: 32×56^2
res_2	1×3^2 max pool, stride 1×2^2 $\begin{bmatrix}\text{TSM}\\1\times1^2, 256\\1\times3^2, 256\\1\times1^2, 256\end{bmatrix}\times3$	$\begin{bmatrix}\text{TSM}\\1\times3^2, 64\\1\times3^2, 64\end{bmatrix}\times2$	appearance: 8×56^2 pose: 32×56^2
res_3	$\begin{bmatrix}\text{TSM}\\1\times1^2, 512\\1\times3^2, 512\\1\times1^2, 512\end{bmatrix}\times4$	$\begin{bmatrix}\text{TSM}\\1\times3^2, 128\\1\times3^2, 128\end{bmatrix}\times2$	appearance: 8×28^2 pose: 32×28^2
res_4	$\begin{bmatrix}\text{TSM}\\1\times1^2, 1024\\1\times3^2, 1024\\1\times1^2, 1024\end{bmatrix}\times6$	$\begin{bmatrix}\text{TSM}\\1\times3^2, 256\\1\times3^2, 256\end{bmatrix}\times2$	appearance: 8×14^2 pose: 32×14^2
res_5	$\begin{bmatrix}\text{TSM}\\1\times1^2, 2048\\1\times3^2, 2048\\1\times1^2, 2048\end{bmatrix}\times3$	$\begin{bmatrix}\text{TSM}\\1\times3^2, 512\\1\times3^2, 512\end{bmatrix}\times2$	appearance: 8×7^2 pose: 32×7^2
pool	global average pool	global average pool	appearance: 8×1^2 pose: 32×1^2
feature align (TCB_A,TCB_P)	1×1^2, 512 layer normalization	4×1^2 avg pool, stride 4×1^2 1×1^2, 512 layer normalization	both: 8×1^2
pose-driven gating (CGB)	$(1-\mathbf{G})$ element-wise product	($\mathbf{G}$:1×1^2, 512) $\mathbf{G}$ element-wise product	
aggregation	element-wise addition		
classifier	fully-connected layer		# of classes

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 ($\mathrm{T}\times\mathrm{S}^2$, C) for the temporal, spatial, and channel sizes. The strides and output size are denoted by ($\mathrm{T}\times\mathrm{S}^2$) for the temporal and spatial sizes.

layers	appearance stream	pose stream	output size
input	RGB frames	keypoint heatmaps+PAFs	appearance: 8×224^2 pose: 8×56^2
conv_1	1×7^2, 64, stride 1×2^2	1×3^2, 64	appearance: 8×112^2 pose: 8×56^2
res_2	1×3^2 max pool, stride 1×2^2 $\begin{bmatrix}\text{TSM}\\1\times3^2, 64\\1\times3^2, 64\end{bmatrix}\times2$	$\begin{bmatrix}\text{TSM}\\1\times3^2, 64\\1\times3^2, 64\end{bmatrix}\times2$	both: 8×56^2
res_3	$\begin{bmatrix}\text{TSM}\\1\times3^2, 128\\1\times3^2, 128\end{bmatrix}\times2$	$\begin{bmatrix}\text{TSM}\\1\times3^2, 128\\1\times3^2, 128\end{bmatrix}\times2$	both: 8×28^2
res_4	$\begin{bmatrix}\text{TSM}\\1\times3^2, 256\\1\times3^2, 256\end{bmatrix}\times2$	$\begin{bmatrix}\text{TSM}\\1\times3^2, 256\\1\times3^2, 256\end{bmatrix}\times2$	both: 8×14^2
res_5	$\begin{bmatrix}\text{TSM}\\1\times3^2, 512\\1\times3^2, 512\end{bmatrix}\times2$	$\begin{bmatrix}\text{TSM}\\1\times3^2, 512\\1\times3^2, 512\end{bmatrix}\times2$	both: 8×7^2
pool	global average pool	global average pool	both: 8×1^2
feature align (TCB_A,TCB_P)	1×1^2, 512 layer normalization	1×1^2, 512 layer normalization	
pose-driven gating (CGB)	$(1-\mathbf{G})$ element-wise product	($\mathbf{G}$:$1\times1\times1$, 512) $\mathbf{G}$ element-wise product	
aggregation	element-wise addition		
classifier	fully-connected layer		# of classes

References

[1] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *CVPR*, 2018.

[2] Wenbin Du, Yali Wang, and Yu Qiao. Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. In *ICCV*, 2017.

[3] Wei Wang, Jinjin Zhang, Chenyang Si, and Liang Wang. Pose-based two-stream relational networks for action recognition in videos. *arXiv preprint arXiv:1805.08484*, 2018.

[4] An Yan, Yali Wang, Zhifeng Li, and Yu Qiao. Pa3d: Pose-action 3d machine for video recognition. In *CVPR*, 2019.

[5] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018.