Learning Discriminative Representations for Skeleton Based Action Recognition

— *Supplementary Material* —

Huanyu Zhou[1], Qingjie Liu[1,2,3,*], Yunhong Wang[1]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China

[2]Zhongguancun Laboratory, [3]Hangzhou Innovation Institute of Beihang University

{zhysora, qingjie.liu, yhwang}@buaa.edu.cn

1. More Details

1.1. Construction of Ambiguous groups

As discussed in the main paper, we define the ambiguous groups, which collect several related ambiguous actions to verify the performance of ambiguous samples. We first pick a class as an anchor class, for example, ”writing”. Then we gather the misclassified samples on ”writing” and obtain the top-3 actions with the highest frequency, like ”reading”, ”typing on a keyboard” and ”playing with phone”. These 4 actions will be constructed as an ambiguous group. we randomly pick 60 anchor actions and constructed corresponding ambiguous groups from NTU-RGB+D 120 dataset.

The details of each ambiguous group are displayed on Tab. 5. From the description of the action names, we can find that actions in the same group are relatively similar to each other. This further demonstrate the rationality to constructing ambiguous groups.

1.2. Pseudo-code

To make our method easy to understand, we provide the PyTorch-like pseudo-code in Algorithm 1.

Algorithm 1 Pseudo-code in PyTorch-like style.

```
# x, y: input of skeleton data and labels
# blocks[i]: the i-th feature extraction block
# fc_layer: fully-connected layer
# fr_head: feature refinement head
# model.params: all trainable parameters of the model
# logits: probability scores predicted by the network
# lambda[i]: the balanced hyper-parameter of stage i
# w_cl: the balanced hyper-parameter for CL loss

for x, y in batch:  # load samples

  # 1. obtain the hidden features for refinement
  refine_feats = []
  for i in 1...10:
    x = blocks[i](x)
    # save the multi-level features
    if i in [1, 5, 8, 10]:
    refine_feats.add(x.clone())

  # 2. obtain the prediction of the network
  x = x.mean(1, 2)
  x = fc_layer(x)
  logits = softmax(x)

  # 3. obtain the loss
  loss_cl = 0
  # sum up the CL loss for each stage: Eq.(1)
  for i in 1...4:
    loss_cl += rf_head(refine_feats, logits) * lambda[i]
  # calculate the CE Loss
  loss_ce = cross_entropy_loss(logits, y)
  # calculate the full learning objective function: Eq
     .(10)
  loss_full = loss_ce + w_cl * loss_cl

  # 4. backpropagation & update params
  loss_full.backward()
  update(model.params)
```

2. More Quantitative Results

2.1. Comparison with more SOTA models

We provide experimental comparisons with more recent SOTA models on NTU-RGB+D [8], NTU-RGB+D 120 [6], and NW-UCLA [11] datasets. The quantitative results are displayed in Table 1.

Due to the different testing configurations, we do not compare our model with them in the main paper. PoseC3D [4] and LST [13] utilize extra data to augment their performance. PoseC3D [4] takes 2D skeletons from high-accuracy estimators as the input and cooperates them with RGB frames. [13] we the pretrained transformer model from CLIP [7] or BERT [3] to construct a language supervised training. InfoGCN [2] and HD-GCN [5] apply a multimodal representation of a skeleton for 6 model ensembles, while our results are based on 4 model ensembles. TCA-GCN [12] obtain comparable results with our model but need nearly 3 times of parameters. As shown in Table 1, despite the unfair testing setting, our method still obtain competitive results on all datasets. These results further prove the effectiveness and generalizability of our method.

*Corresponding author

Table 1. Performance comparison of skeleton-based action recognition in top-1 accuracy (%).

Method	Publication	Extra Data	Mode	Params.	NTU RGB+D X-Sub	NTU RGB+D X-View	NTU RGB+D 120 X-Sub	NTU RGB+D 120 X-Set	NW-UCLA
PoseC3D [4]	CVPR2022	✓	-	2.0M	94.1	97.1	86.9	90.3	-
InfoGCN [2]	CVPR2022	✕	6 ensemble	1.6M	93.0	97.1	89.8	91.2	97.0
HD-GCN [5]	Arxiv2022	✕	6 ensemble	1.7M	93.4	97.2	90.1	91.6	97.0
LST [13]	Arxiv2022	✓	4 ensemble	-	92.9	97.0	89.9	91.1	97.2
TCA-GCN [12]	Arxiv2022	✕	4 ensemble	5.7M	92.8	97.0	89.4	90.8	97.0
Ours	-	✕	4 ensemble	2.0M	92.8	96.8	89.5	90.9	96.8

Table 2. Comparison of class-wise accuracy (%) on NW-UCLA dataset with the joint input modality.

Action Name	ST-GCN [14]	2s-AGCN [9]	CTR-GCN [1]	Ours
pick up with one hand	75.00	79.17	79.17	**87.50**
pick up with two hands	**100.00**	**100.00**	**100.00**	**100.00**
drop trash	**100.00**	**100.00**	**100.00**	**100.00**
walk around	91.84	89.80	**93.88**	89.80
sit down	97.87	**100.00**	97.87	97.87
stand up	**100.00**	95.74	97.87	**100.00**
donning	**100.00**	**100.00**	97.67	**100.00**
doffing	90.70	88.37	**97.67**	90.70
throw	91.49	91.49	**95.74**	**95.74**
carry	**84.44**	80.00	82.22	82.22
Average	93.10	92.46	94.18	**94.40**

2.2. Category Performance

Tables 3, 4 and 2 present the category performance comparisons on NTU-RGB+D [8], NTU-RGB+D 120 [6] and NW-UCLA [11] datasets, respectively. We apply the X-Sub setting on both NTU-RGB+D and NTU-RGB+D 120 datasets. All experiments are conducted with only the joint input modality. We compare our method with ST-GCN [14], 2s-AGCN [9] and CTR-GCN [1]. From the results, our method improves the other baselines on most categories.

3. More Qualitative Results

Fig. 1 shows the t-SNE [10] visualization of the feature space. As illustrated, 7 anchor actions are selected to construct the corresponding ambiguous group. As discussed in the main paper, each ambiguous group contains 4 classes, including an anchor class and three ambiguous classes. We compare our method with ST-GCN [14], 2s-AGCN [9], and CTR-GCN [1].

From the results, our model obtains a more discriminative representation resulting in a compact clustering. For example, at the first and third rows in Fig. 1, we find that instances of the class indicated by red are relatively closer to each other in the feature space compared to the baselines. This further confirms the effectiveness of our proposed method.

Table 3. Comparison of class-wise accuracy (%) on NTU-RGB+D 60 dataset under the X-Sub setting with the joint input modality.

Action Name	ST-GCN [14]	2s-AGCN [9]	CTR-GCN [1]	Ours
drink water	83.21	**85.77**	82.85	85.04
eat meal/snack	69.82	70.91	70.55	**74.55**
brushing teeth	**83.88**	79.85	81.68	79.49
brushing hair	89.74	89.38	91.58	**92.67**
drop	85.45	**89.09**	**89.09**	84.36
pickup	97.09	98.55	97.45	**98.91**
throw	92.00	**94.91**	91.27	**94.91**
sitting down	95.97	96.70	96.70	**97.07**
standing up	97.80	**98.53**	98.17	**98.53**
clapping	83.15	79.12	84.98	**88.28**
reading	65.57	62.27	**73.99**	69.60
writing	56.62	**60.29**	56.62	59.19
tear up paper	91.51	88.56	90.77	**94.83**
wear jacket	96.73	**98.18**	**98.18**	**98.18**
take off jacket	**98.55**	98.19	97.83	**98.55**
wear a shoe	**83.15**	76.19	79.49	78.75
take off a shoe	73.72	**80.29**	75.55	77.74
wear on glasses	92.31	**95.60**	94.14	92.67
take off glasses	93.80	**96.72**	95.26	95.26
put on a hat/cap	95.22	94.85	94.49	**96.32**
take off a hat/cap	**97.80**	**97.80**	97.44	97.44
cheer up	**93.80**	93.43	92.34	**93.80**
hand waving	92.34	90.88	92.34	**92.70**
kicking something	92.39	94.20	94.20	**97.46**
reach into pocket	82.12	82.48	**85.40**	81.75
hopping	98.18	**98.91**	97.82	97.45
jump up	**100.00**	99.64	99.64	**100.00**
make a phone call/answer phone	86.55	86.91	**88.73**	87.27
playing with phone/tablet	66.18	70.55	69.09	**73.82**
typing on a keyboard	**72.36**	59.27	69.45	67.27
pointing to something with finger	83.33	**85.51**	83.70	84.42
taking a selfie	89.86	89.86	**90.58**	87.68
check time	86.59	87.68	88.41	**89.13**
rub two hands together	89.13	89.49	89.49	**93.48**
nod head/bow	96.38	96.01	96.01	**96.74**
shake head	95.27	**96.00**	94.55	95.27
wipe face	**88.04**	82.25	83.70	84.42
salute	92.75	92.75	93.48	**94.93**
put the palms together	**93.48**	93.12	**93.48**	91.67
cross hands in front	**95.29**	94.20	92.75	94.93
sneeze/cough	75.72	**81.88**	79.35	75.36
staggering	**99.28**	**99.28**	98.91	98.91
falling	98.18	98.55	**98.91**	**98.91**
touch head	81.16	**82.97**	81.52	**82.97**
touch chest	90.22	94.20	93.84	**95.29**
touch back	90.22	**92.39**	**92.39**	91.30
touch neck	86.96	83.33	88.04	**89.13**
nausea or vomiting condition	85.82	78.55	**87.27**	86.55
use a fan /feeling warm	91.64	91.64	**95.64**	94.55
punching/slapping other person	91.61	91.24	91.61	**92.70**
kicking other person	94.20	94.93	**95.65**	95.29
pushing other person	96.38	**98.19**	96.74	97.10
pat on back of other person	**93.84**	89.49	**93.84**	90.22
point finger at the other person	92.39	89.86	92.03	**93.48**
hugging other person	98.54	98.54	98.54	**98.91**
giving something to other person	**95.29**	94.20	93.48	93.84
touch other person’s pocket	92.73	94.91	95.27	**95.64**
handshaking	95.65	**96.01**	95.29	95.65
walking towards each other	99.63	99.27	99.63	**100.00**
walking apart from each other	**96.74**	96.38	96.01	**96.74**
Average	89.40	89.36	89.96	**90.33**

Table 4. Comparison of class-wise accuracy (%) on NTU-RGB+D 120 dataset under the X-Sub setting with the joint input modality.

Action Name	*ST-GCN [14]*	*2s-AGCN [9]*	*CTR-GCN [1]*	*Ours*	*Action Name*	*ST-GCN [14]*	*2s-AGCN [9]*	*CTR-GCN [1]*	*Ours*
drink water	83.58	85.04	**87.59**	82.85	eat meal/snack	68.36	69.09	70.18	**72.36**
brushing teeth	**82.42**	79.12	80.95	81.68	brushing hair	85.71	87.55	86.81	**89.01**
drop	84.00	82.91	86.55	**90.91**	pickup	**99.27**	98.91	98.55	98.55
throw	**93.45**	**93.45**	90.91	91.64	sitting down	**96.70**	95.97	**96.70**	94.87
standing up	**97.80**	**97.80**	**97.80**	97.44	clapping	74.73	**85.71**	81.68	83.52
reading	55.68	62.27	61.90	**70.33**	writing	**59.93**	56.99	56.99	55.51
tear up paper	90.04	**90.41**	88.93	88.19	wear jacket	96.73	**97.82**	97.45	**97.82**
take off jacket	97.83	97.83	97.83	**98.55**	wear a shoe	76.92	83.52	**83.88**	73.26
take off a shoe	74.82	75.55	75.91	**81.75**	wear on glasses	92.67	92.67	91.21	**94.14**
take off glasses	**94.16**	90.51	92.34	93.80	put on a hat/cap	94.49	95.59	94.49	**96.32**
take off a hat/cap	97.44	97.44	97.07	**98.17**	cheer up	91.97	**95.62**	94.53	92.70
hand waving	91.61	91.61	**93.07**	91.97	kicking something	94.20	92.75	**95.29**	94.93
reach into pocket	85.04	83.94	83.21	**85.40**	hopping	**98.18**	96.73	96.73	97.82
jump up	**100.00**	**100.00**	99.64	**100.00**	make a phone call/answer phone	88.73	87.64	85.09	**89.09**
playing with phone/tablet	52.36	50.55	**61.82**	57.82	typing on a keyboard	67.27	58.91	64.00	**68.00**
pointing to something with finger	**81.88**	76.81	78.26	75.36	taking a selfie	**90.94**	89.86	88.77	89.86
check time	**91.30**	88.77	**91.30**	89.13	rub two hands together	86.59	87.32	88.41	**89.49**
nod head/bow	96.38	95.29	95.65	**97.83**	shake head	92.73	91.64	94.55	**96.00**
wipe face	84.78	85.51	85.87	**87.32**	salute	92.39	92.39	90.58	**93.48**
put the palms together	92.39	91.67	89.86	**93.84**	cross hands in front	95.29	95.65	**96.01**	93.12
sneeze/cough	73.55	**77.90**	75.00	72.10	staggering	98.91	**99.28**	98.55	98.91
falling	98.18	**98.55**	97.09	**98.55**	touch head	81.16	78.62	**83.33**	82.61
touch chest	92.39	91.67	92.03	**93.12**	touch back	90.22	89.49	89.86	**93.12**
touch neck	88.41	**90.58**	87.68	89.13	nausea or vomiting condition	86.18	86.55	**86.91**	86.55
use a fan /feeling warm	91.27	92.00	91.27	**94.55**	punching/slapping other person	**89.42**	87.23	87.23	87.59
kicking other person	93.84	93.84	93.12	**94.93**	pushing other person	97.46	**97.83**	96.38	96.74
pat on back of other person	**95.29**	91.67	90.94	92.39	point finger at the other person	**92.03**	90.22	90.58	89.13
hugging other person	**99.27**	98.54	98.54	98.91	giving something to other person	89.13	**90.58**	86.23	**90.58**
touch other person's pocket	91.27	**94.55**	92.00	92.73	handshaking	**96.38**	95.29	95.29	**96.38**
walking towards each other	98.17	**99.27**	**99.27**	98.90	walking apart from each other	96.74	96.74	96.38	**97.10**
put on headphone	88.24	**89.84**	88.06	89.30	take off headphone	86.75	89.05	85.69	**89.40**
shoot at the basket	85.84	**86.89**	**86.89**	**86.89**	bounce ball	**97.37**	96.32	96.32	97.19
tennis bat swing	82.93	**86.93**	86.06	86.24	juggling table tennis balls	96.68	**97.21**	95.81	**97.21**
hush	76.27	77.66	76.44	**79.76**	flick hair	77.91	76.87	81.22	**82.96**
thumb up	64.70	59.13	64.70	**67.13**	thumb down	88.00	**89.91**	**89.91**	89.39
make ok sign	50.09	48.17	47.30	**51.30**	make victory sign	39.48	**42.96**	36.87	38.26
staple book	34.68	35.90	**42.03**	38.00	counting money	56.32	59.82	52.81	**60.88**
cutting nails	59.05	58.17	**71.88**	65.03	cutting paper	56.54	58.81	60.91	**61.08**
snapping fingers	65.16	68.99	68.82	**71.08**	open bottle	72.25	66.84	69.46	**74.52**
sniff	78.26	**81.04**	77.91	79.30	squat down	98.61	98.43	98.43	**98.95**
toss a coin	88.83	**90.58**	89.18	90.23	fold paper	**69.22**	68.17	67.13	66.78
ball up paper	76.17	76.35	70.09	**77.04**	play magic cube	63.81	61.89	**65.03**	64.16
apply cream on face	82.75	85.71	85.02	**86.59**	apply cream on hand back	70.03	72.30	72.13	**77.87**
put on bag	94.43	93.74	94.43	**97.22**	take off bag	94.44	**94.79**	93.92	94.27
put something into a bag	73.39	77.39	80.00	**80.17**	take something out of a bag	85.07	85.42	81.25	**89.93**
open a box	70.03	**73.52**	73.34	69.86	move heavy objects	93.49	94.72	94.19	**95.60**
shake fist	80.56	79.17	83.16	**84.72**	throw up cap/hat	80.80	**86.21**	85.51	84.47
hands up	95.30	96.34	96.34	**96.86**	cross arms	94.78	96.52	**97.22**	**97.22**
arm circles	99.13	99.13	98.78	**99.65**	arm swings	98.95	99.13	98.78	**99.48**
running on the spot	96.87	96.70	96.52	**97.39**	butt kicks	93.73	95.47	95.64	**96.86**
cross toe touch	96.69	93.90	94.43	**97.21**	side kick	94.08	**95.47**	94.08	**95.47**
yawn	66.96	**73.04**	66.43	72.17	stretch oneself	90.97	89.93	90.97	**92.71**
blow nose	61.22	61.74	**67.13**	61.91	hit other person with something	66.26	63.83	62.78	**70.96**
wield knife towards other person	66.49	65.10	**74.65**	71.18	knock over other person	89.24	**92.36**	88.89	91.32
grab other person's stuff	90.78	91.30	91.13	**92.87**	shoot at other person with a gun	75.48	74.26	**79.65**	78.43
step on foot	93.74	93.04	**95.65**	93.57	high-five	**98.09**	97.57	97.74	97.92
cheers and drink	98.43	**99.30**	**99.30**	98.43	carry something with other person	95.49	95.49	**96.18**	95.66
take a photo of other person	**94.97**	92.36	93.40	92.71	follow other person	**96.18**	94.44	93.75	95.49
whisper in other person's ear	92.00	**92.70**	**92.70**	92.17	exchange things with other person	88.35	91.30	**92.70**	92.52
support somebody with hand	91.48	91.48	**92.52**	91.83	finger-guessing game	96.70	**97.05**	95.83	96.70
Average	83.42	84.31	84.54	**85.51**					

Table 5. Details of our selected ambiguous groups on NTU-RGB+D 120 dataset.

Anchor Action	1st Wrongest Action	2nd Wrongest Action	3rd Wrongest Action
make victory sign	make ok sign	thumb up	thumb down
make ok sign	make victory sign	thumb up	thumb down
counting money	play magic cube	cutting nails	playing with phone/tablet
writing	typing on a keyboard	reading	playing with phone/tablet
cutting paper	staple book	reading	cutting nails
reading	writing	playing with phone/tablet	cutting paper
hit other person with something	wield knife towards other person	punching/slapping other person	grab other person’s stuff
typing on a keyboard	writing	playing with phone/tablet	reading
thumb up	make ok sign	make victory sign	thumb down
play magic cube	counting money	playing with phone/tablet	cutting nails
yawn	hush	blow nose	flick hair
fold paper	ball up paper	play magic cube	counting money
blow nose	yawn	hush	sniff
snapping fingers	shake fist	thumb up	make victory sign
open bottle	play magic cube	open a box	apply cream on hand back
ball up paper	fold paper	play magic cube	counting money
eat meal/snack	brushing teeth	take off glasses	make a phone call/answer phone
apply cream on hand back	rub two hands together	open bottle	counting money
open a box	fold paper	apply cream on hand back	reading
wield knife towards other person	hit other person with something	point finger at the other person	grab other person’s stuff
sneeze/cough	touch head	nausea or vomiting condition	touch chest
take off a shoe	wear a shoe	kicking something	reading
hush	blow nose	yawn	sniff
sniff	hush	blow nose	make victory sign
pointing to something with finger	taking a selfie	thumb up	drink water
shoot at other person with a gun	point finger at the other person	take a photo of other person	wield knife towards other person
put something into a bag	take something out of a bag	open a box	take off a shoe
flick hair	brushing hair	blow nose	touch head
take something out of a bag	put something into a bag	open a box	wear a shoe
clapping	rub two hands together	playing with phone/tablet	put the palms together
shake fist	snapping fingers	hand waving	salute
reach into pocket	touch back	wear a shoe	wear on glasses
touch head	touch neck	wear on glasses	brushing teeth
wear a shoe	take off a shoe	butt kicks	hand waving
apply cream on face	wipe face	eat meal/snack	sniff
make a phone call/answer phone	apply cream on face	touch head	brushing teeth
throw up cap/hat	toss a coin	shoot at the basket	thumb up
take off headphone	take off glasses	flick hair	take off a hat/cap
wipe face	brushing hair	sneeze/cough	apply cream on face
tennis bat swing	throw up cap/hat	shoot at the basket	tear up paper
giving something to other person	exchange things with other person	handshaking	point finger at the other person
drop	tear up paper	make a phone call/answer phone	playing with phone/tablet
brushing hair	wipe face	flick hair	touch head
shoot at the basket	throw	throw up cap/hat	hands up
nausea or vomiting condition	touch chest	sneeze/cough	nod head/bow
punching/slapping other person	hit other person with something	pushing other person	pat on back of other person
drink water	eat meal/snack	brushing teeth	wear on glasses
touch neck	touch head	drink water	touch back
put on headphone	wear on glasses	sniff	take off headphone
rub two hands together	clapping	apply cream on hand back	use a fan /feeling warm
taking a selfie	pointing to something with finger	drink water	hand waving
knock over other person	step on foot	whisper in other person’s ear	wield knife towards other person
tear up paper	fold paper	rub two hands together	reading
toss a coin	thumb up	snapping fingers	make victory sign
put the palms together	cross hands in front	hugging other person	sniff
touch back	reach into pocket	touch chest	standing up
thumb down	thumb up	pointing to something with finger	make ok sign
salute	brushing teeth	touch head	yawn
point finger at the other person	shoot at other person with a gun	pat on back of other person	punching/slapping other person
throw	wear jacket	punching/slapping other person	put on bag

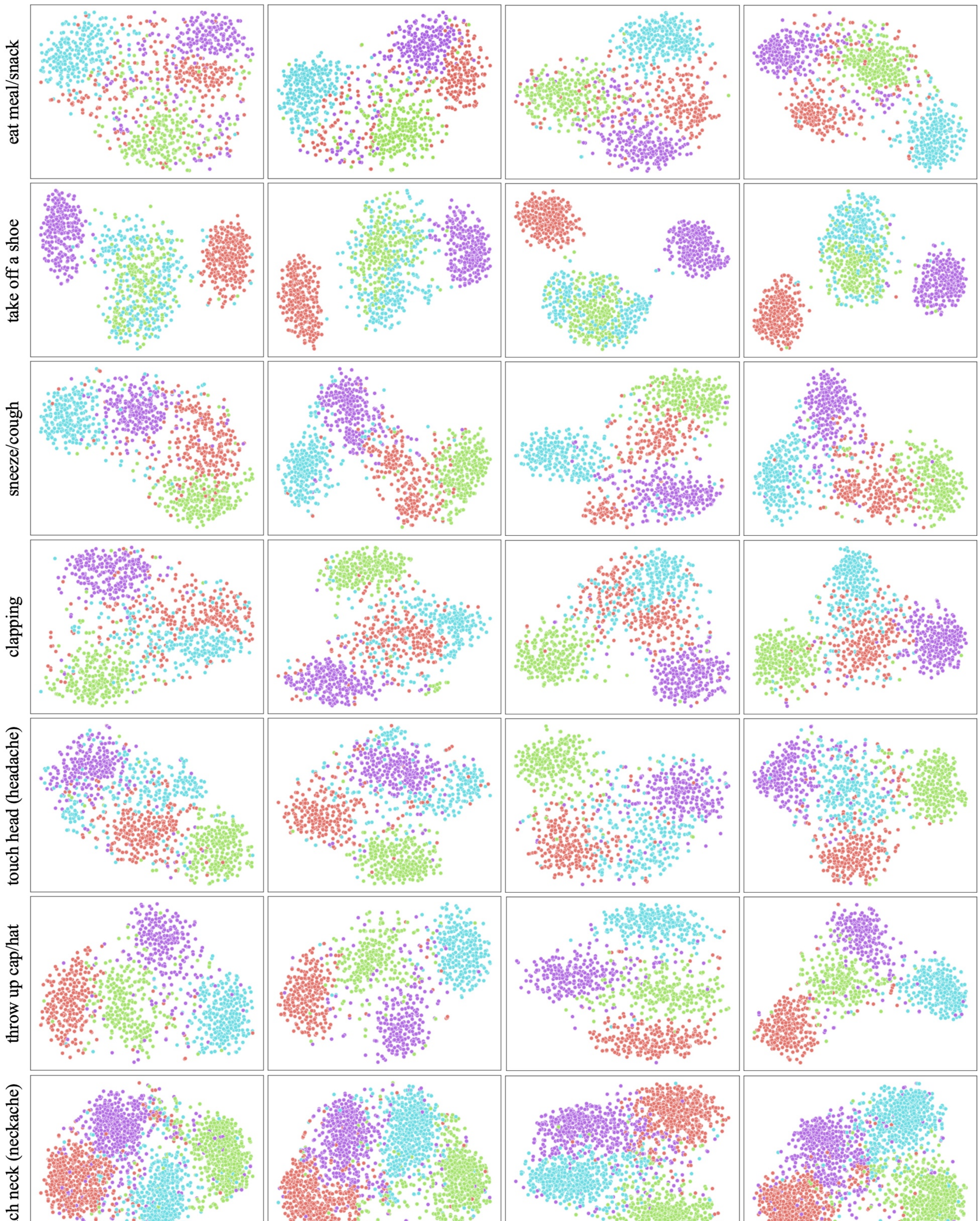

Figure 1. T-SNE visualization [10] of the feature space for ambiguous groups on NTU RGB+D 120 dataset. The texts on the left describe the anchor action of the corresponding ambiguous group. Different colors indicate different classes.

References

[1] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 13359–13368, 2021. 2, 3, 4

[2] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 20186–20196, June 2022. 1, 2

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 1

[4] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022. 1, 2

[5] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoon Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:2208.10741*, 2022. 1, 2

[6] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2019. 1, 2

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[8] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016. 1, 2

[9] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2019. 2, 3, 4

[10] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008. 2, 6

[11] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2649–2656, 2014. 1, 2

[12] Shengqin Wang, Yongji Zhang, Fenglin Wei, Kai Wang, Minghao Zhao, and Yu Jiang. Skeleton-based action recognition via temporal-channel aggregation. *arXiv preprint arXiv:2205.15936*, 2022. 1, 2

[13] Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Language supervised training for skeleton-based action recognition. *arXiv preprint arXiv:2208.05318*, 2022. 1, 2

[14] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7444–7452, 2018. 2, 3, 4