TranSG: Transfomer-Based Skeleton Graph Prototype Contrastive Learning with Structure-Trajectory Prompted Reconstruction for Person Re-Identification – Appendix II Experiments

Haocong Rao Chunyan Miao*
LILY Research Center, Nanyang Technological University, Singapore
School of Computer Science and Engineering, Nanyang Technological University, Singapore
{haocong001, ascymiao}@ntu.edu.sg

1. Supplementary Experimental Settings

In this section, we provide more detailed experimental settings, including CASIA-B evaluation settings (Sec. 1.1), dataset preprocessing strategy (Sec. 1.2), probe and gallery settings (Sec. 1.3), and model implementation details (Sec. 1.4).

1.1. Evaluation Settings of CASIA-B

In general, 3D skeleton data in existing skeleton-based person re-ID benchmarks are collected with Kinect [1]. To evaluate the effectiveness of our approach when 3D skeleton data are directly estimated from RGB videos rather than depth sensors such as Kinect, we use a large-scale RGB video based dataset, *CASIA-B* [2], which contains walking sequences of 124 individuals under 11 different views and 3 conditions—pedestrians wearing a bag ("Bags"), wearing a coat ("Clothes"), and without any coat or bag ("Normal"). We follow the evaluation setup in [3], which is frequently used in the literature: First, we randomly choose half of the individuals for training and use the rest for testing. Then, to evaluate our approach under *single-condition* and *cross-condition* settings, we divide the testing sequences by the three conditions ("Bags", "Clothes", "Normal") to construct gallery and probe sets. Specifically, for the *single-condition* setting, both gallery and probe sets use the testing sequences with the same condition ($i.e.$, gallery and probe sets are the same), and we match each sequence of the probe set with the most similar sequence from the gallery set that *excludes* the original sequence. In the *cross-condition* setting, we adopt the testing sequences under bags ("Bags") or clothes condition ("Clothes") as the probe set, and use the testing sequences under normal condition ("Normal") as the gallery set.

Following [4], we exploit pre-trained pose estimation models [5, 6] to extract 3D skeletons from RGB videos of CASIA-B. We first extract eighteen 2D joints from each person in videos using the *OpenPose* model [6]. Then, we follow the same configuration of estimation in [4] and average the positions of "Nose", "Reye", "LEye", "Rear" and "Lear" as the position of "Head" to construct fourteen 2D joints, which are fed into the pose estimation method [5] to estimate corresponding 3D body joints. Thus, the number of body-joint nodes J is 14 for CASIA-B as shown in Fig. 1, and all joints in each skeleton are normalized by subtracting the neck joint.

Table 1. Overview of datasets (K: thousand). Different testing splits are used to construct gallery sets and probe sets (see Sec. 1.3). "W", "S", "A", and "B" denote BIWI-Walking, BIWI-Still, IAS-A, and IAS-B testing sets, respectively. "N", "C", and "B" represent "Normal", "Clothes", and "Bags" conditions of CASIA-B, respectively. Note: The 3D skeletons of CASIA-B are estimated from RGB videos.

# Datasets	KGBD	BIWI	KS20	IAS	CASIA-B
# Train IDs	164	50	20	11	124
# Train Skeletons	188.7K	205.8K	36.0K	89.0K	706.5K
# Probe IDs	164	28	20	11	62
# Probe Skeletons	94.1K	W: 4.9K S: 3.2K	3.3K	A: 7.0K B: 7.8K	N: 162.1K C: 54.4K B: 53.9K
# Gallery IDs	164	28	20	11	62
# Gallery Skeletons	188.7K	W: 4.9K S: 3.2K	3.3K	A: 7.0K B: 7.8K	N: 162.1K C: 54.4K B: 53.9K

1.2. Dataset Preprocssing

To avoid ineffective skeleton recording, we discard the first and last 10 skeleton frames of each original skeleton sequence. For KS20, KGBD, BIWI, and IAS datasets, all skeleton sequences are normalized by subtracting the spine joint position from each joint of the same skeleton so that

*Corresponding author

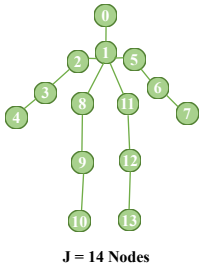


Figure 1. Node indices for graph representations of the estimated skeletons from CASIA-B dataset. Note: All 3D skeletons are estimated from RGB videos of CASIA-B with [6] and [5] (see Sec. 1.1).

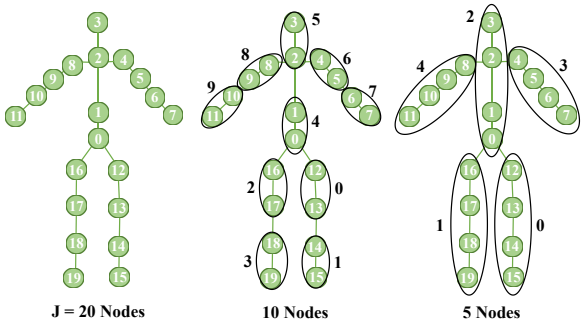


Figure 2. Node indices for joint-scale (20 nodes), part-level (10 nodes), and body-scale (5 nodes) graphs representations of skeletons from IAS, BIWI and KGBD datasets. Our approach *only* requires joint-scale graphs for training, while we evaluate its performance on different-scale graphs following [10] in the paper.

the skeleton is translation invariant [7]. Then, we spilt all normalized skeleton sequences in the training sets into multiple shorter skeleton sequences ($i.e.$, $\boldsymbol{X}$) with length f by a step of $\frac{f}{2}$, which aims to obtain as many 3D skeleton sequences as possible to train our approach. We split all skeleton sequences in the gallery and probe sets into shorter and non-overlapping sequences with length f. Unless explicitly specified, the skeleton sequence $\boldsymbol{X}$ in our paper refers to those split and normalized sequences used in learning, rather than those original skeleton sequences provided by datasets. We follow the data augmentation strategy used in [8, 9] to sample more sequences for different identities in the training set, and train our approach with randomly shuffled skeleton sequences of the training set. The details of all datasets are shown in Table 1.

1.3. Probe and Gallery Settings

We follow the commonly-used settings of probe and gallery in the literature [11]: For the BIWI and IAS datasets, as different testing sets are non-overlapped and contain all pedestrians under different scenes, we evaluate our approach on each testing set by setting it as the probe while

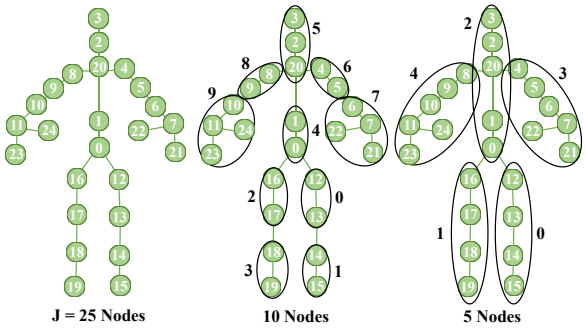


Figure 3. Node indices for joint-scale (25 nodes), part-level (10 nodes), and body-scale (5 nodes) graphs representations of skeletons from KS20 dataset. Our approach *only* requires joint-scale graphs for training, while we evaluate its performance on different-scale graphs following [10] in the paper.

the other one is adopted as the gallery. The KGBD dataset contains different skeleton videos ($i.e.$, long skeleton sequences) of each pedestrian with varying numbers of walking rounds. Since no training/testing splits are given, we randomly choose one skeleton video of each person to split skeleton sequences and construct the probe set, and equally divide the remaining videos to build the training set and gallery set. The KS20 dataset collects skeleton data of pedestrians from five different viewpoints, including $0°$, $30°$, $90°$, $130°$, and $180°$. We employ the setting of Random View Evaluation (RVE): One sequence is randomly selected from each viewpoint as the probe sequence and the remaining skeleton sequences are equally divided into gallery and training sequences. We follow the person re-ID protocols in [3] to evaluate the proposed skeleton-based approach on CASIA-B (detailed in Sec. 1.1).

1.4. Implementation Details

All the important experimental details are presented in our paper. The numbers of body joints are $J = 20$ (IAS, BIWI, KGBD) and $J = 25$ (KS20) in the original datasets. We construct corresponding skeleton graphs with the same number of body-joint nodes in the original skeletons. To verify the generality of our approach when applied to different-scale skeleton graphs, we follow [10] to construct another two scales, namely part-scale (10 nodes) and body-scale (5 nodes), by merging joints within different body partitions. The original skeleton graphs, part-scale graphs, and body-scale graphs as shown in Fig. 2 and 3. The skeleton sequence length f on four skeleton-based datasets (IAS, KS20, BIWI, KGBD) is set to 6 following [11] for a fair comparison with existing methods. As to CASIA-B, it is a large-scale dataset with roughly estimated skeleton data from RGB frames, which is intrinsically different from the previous datasets. We adopt a longer sequence length $f = 40$. The embedding size of each node representation

Table 2. Full results for ablation study with different configurations: Naïve prototype contrastive learning (PC), skeleton graph transformer (SGT) with direct supervised learning (DS) or graph prototype contrastive learning (GPC), and structure-trajectory prompted reconstruction (STPR). "+" indicates employing the component.

Configurations	BIWI-S		BIWI-W		KS20		IAS-A		IAS-B		KGBD	
	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1
Baseline	9.3	24.8	14.1	10.9	9.5	17.0	13.8	29.4	13.3	30.2	6.4	34.5
PC	11.3	38.1	18.3	21.2	20.5	64.8	17.8	39.2	21.5	40.7	11.0	53.0
SGT + DS	19.0	42.4	21.1	21.7	27.6	60.0	27.7	42.9	34.4	51.9	11.1	51.5
SGT + GPC	26.7	66.6	25.5	31.2	42.5	71.3	31.8	48.0	37.9	56.1	18.1	57.0
SGT + GPC + STPR	30.1	68.7	26.9	32.7	46.2	73.6	32.8	49.2	39.4	59.1	20.2	59.0

Table 3. The number of network parameters (million (M)) and computational complexity (giga floating-point operations (GFLOPs)) of deep learning based methods. ♠ denotes skeleton graph based methods, † indicates using hand-crafted descriptors, and ‡ refers to sequence representation learning models. Note: Both numbers of parameters and GFLOPs in the training of neural networks are counted by the Tensorflow platform [12]. Extra matrix computation is required for the clustering in SimMC and SPC-MGR (see Sec. 2.1).

Methods	# Params	GFLOPs
PoseGait† [4]	8.93M	121.60
MG-SCR♠ [9]	0.35M	6.60
AGE‡ [13]	7.15M	37.37
SGELA‡ [8]	8.47M	7.47
SM-SGE♠ [10]	5.58M	22.61
SPC-MGR♠ [14]	0.01M	0.12
SimMC‡ [11]	0.15M	0.99
TranSG♠ (Ours)	0.40M	20.22

is $d = 128$ for all datasets. We empirically set $K = 10$ for the positional encoding, and employ 2 SGT layers with $H = 8$ attention heads and $d_k = 16$ for each layer, as these settings achieve the best average performance on different datasets. We follow [15, 16] to randomly flip the sign of the eigenvectors during training to improve the model stability on small datasets (IAS, KS20, BIWI). For part-scale (10 nodes) and body-scale (5 nodes) skeleton graphs compared with SM-SGE, we correspondingly set $K = 9$ and $K = 4$ for the positional encoding. For main experiments, we equally fuse each component in our approach with $\alpha = 0.5, \beta = 0.5, \lambda = 0.5$. For the experiments with RGB-estmiated skeletons (Sec. 5 in the paper), we empirically set $\alpha = 1.0$ as it can achieve better performance. It should be noted that the models trained with RGB-estimated skeletons possess relatively large performance variations, possibly due to the noise in roughly-estimated skeletons. We thus select the models with slightly better overall performance ($i.e.$, higher mAP instead of higher Rank-1 accuracy) for the discussion in the paper. We will provide a systematic analysis for the model initializations and performance variations in our future works. We empirically set $\tau_1 = 0.07$ and $\tau_2 = 14$ for contrastive learning, while using $a = 10$ and $b = 2$ random masks for STPR. An Adam optimizer with the learning rate of of 3.5×10^{-4} is used for the model optimization, and we set batch size to 256 for all datasets.

To apply our approach to unsupervised skeleton representation learning without using ground-truth labels, we follow [14] to perform DBSCAN clustering [17] of graph representations, and leverage their pseudo classes to generate graph prototypes for contrastive learning. To avoid overfitting and achieve better generalization performance, we adopt Early Stopping [18] with a patience of 120 epochs ($i.e.$, stop the training of model after no improvement in 120 continuous epochs). The experiments are repeated for multiple time with random model parameter initialization for training, and we report the average performance for a fair comparison with existing methods. Interested readers can access our source code[1] for more details.

For all methods compared in our experiments, we select optimal model parameters for training, and use their predefined skeleton descriptors or pre-trained skeleton representations for person re-ID. It is worth noting that our re-implementations of some existing models get performance with slight variations, and the results are basically the same as the original papers under different random model initializations. For a fair comparison, we follow [8, 11] to report the average performance of all methods. Note that our approach does not use any post-processing technique, $e.g.$, re-ranking [19] or multi-query fusion [20] in the training or testing stage. To perform person re-ID, we exploit the approach to encode each original skeleton sequence of the probe set Φ_P into corresponding sequence-level graph representations, $\{\boldsymbol{S}_i^P\}_{i=1}^{N_2}$, and match it with representations, $\{\boldsymbol{S}_i^G\}_{i=1}^{N_3}$, of the same identity in the gallery set Φ_G using Euclidean distance. In the ablation study, we use the concatenation of raw skeleton sequences ($i.e.$, normalized 3D coordinates of body joints) as the baseline. For the configuration of naïve prototype contrastive learning (PC), we adopt the same setting in [11]: We leverage DBSCAN [17] to cluster original skeleton sequences in an unsupervised manner, and directly use the feature centroid in each cluster as the skeleton prototype for contrastive learning.

[1] Our codes are publicly available at https://github.com/Kali-Hac/TranSG.

Table 4. Performance of our approach on different datasets when setting different weight coefficients α to fuse sequence-level ($\mathcal{L}_{\text{GPC}}^{seq}$) and skeleton-level graph prototype contrastive learning ($\mathcal{L}_{\text{GPC}}^{ske}$) in the proposed GPC.

α	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
0.0	71.0	33.5	55.5	14.2	47.3	29.7	57.3	36.3	27.3	25.5	62.7	21.3
0.2	71.5	40.0	56.3	17.9	46.9	31.2	58.9	38.6	31.2	27.5	67.2	29.7
0.4	72.2	45.7	57.5	18.6	47.7	30.9	58.4	38.0	31.7	26.0	68.7	29.9
0.5	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
0.6	73.1	46.3	58.0	19.0	50.9	33.6	58.7	38.8	31.0	26.5	66.8	26.7
0.8	71.7	44.0	58.8	19.8	49.2	31.5	58.2	37.8	32.0	26.0	64.7	26.7
1.0	70.1	41.8	56.9	18.0	47.3	32.6	54.3	37.8	30.3	25.2	64.5	26.1

Table 5. Performance of our approach on different datasets when setting different weight coefficients β to combine graph trajectory-prompted ($\mathcal{L}_{\text{STPR}}^{tr}$) and structure-prompted reconstruction ($\mathcal{L}_{\text{STPR}}^{st}$) in the proposed STPR.

β	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
0.0	73.0	45.4	58.6	19.6	48.4	32.1	58.2	39.2	31.1	25.9	67.8	28.2
0.2	72.6	44.5	56.4	17.8	48.2	32.3	57.3	39.1	31.8	27.0	68.0	29.6
0.4	72.7	44.6	57.1	21.1	48.1	32.4	57.6	40.1	31.5	27.0	69.3	30.9
0.5	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
0.6	72.9	44.7	56.1	18.4	46.9	33.2	57.7	38.7	33.1	27.9	68.6	30.4
0.8	72.3	47.1	58.8	20.4	47.1	32.9	57.6	39.9	31.0	26.3	68.7	30.4
1.0	71.9	44.8	58.1	18.9	48.7	32.4	57.2	38.1	31.9	26.2	67.6	27.0

2. Supplementary Results

In this section, we provide full experimental results for ablation study (Table 2), model efficiency (Sec. 2.1), effects of hyper-parameters (Sec. 2.2, Table 4-12), multi-shot performance with different sequence lengths f (Sec. 2.3), effects of graph positional encoding (Sec. 2.4), visualization of body-component relations (Sec. 2.5), training metrics (different losses) (Sec. 2.6), and confusion matrices (Sec. 2.7).

2.1. Model Efficiency

We report the model efficiency in terms of model size, $i.e.$, number of network parameters, and computational complexity for existing deep learning based methods. For the model that possesses varying sizes and complexities on different datasets due to the changes of input data, we report the largest case. As shown in Table 3, the proposed approach possesses smaller model size than many existing skeleton-based person re-ID methods (PoseGait [4], AGE [13], SGELA [8], SM-SGE [10]). The number of GFLOPs in Table 3 refers to computational complexity in the training of neural networks, which is the whole[2]/main computational complexity for deep learning methods. It should be noted that the unsupervised prototype contrastive learning (SimMC [11], SPC-MGR [14]) requires extra matrix computation ($e.g.$, vector similarity query) for the clustering process , which is usually time-consuming and computationally expensive as it may require both CPU and GPU ($e.g.$, using Faiss library [21]). In contrast, our approach exploits ground-truth identities to generate graph prototypes, which can not only improve the prototype reliability but also achieve significantly faster training without requiring clustering.

[2]For representation learning methods without other learning processes ($e.g.$, clustering), the whole computational complexity of the model can be equivalent to the computational complexity of the used neural networks.

2.2. Effects of Different Hyper-Parameters

Effects of different weight coefficient α, β, and λ: As shown in Table 6, our approach can achieve the best performance in average on different datasets when equally ($i.e.$, $\lambda = 0.5$) fusing the proposed graph prototype contrastive learning (GPC) and structure-trajectory prompted reconstruction (STPR). This is also consistent with our analysis in Sec. 2.6 that GPC and STPR are compatible and can facilitate each other to learn better skeleton graph representations for person re-ID. Interestingly, only using the reconstruction mechanism (STPR) without GPC ($\lambda = 0.0$) can still learn effective skeleton graph features for person re-ID despite with significantly lower accuracy, which suggests the higher contribution of GPC and the limited ability of STPR on learning discriminative skeleton features. For GPC, we observe that an appropriate fusion ($\alpha = 0.4 - 0.6$) of sequence-level ($\mathcal{L}_{\text{GPC}}^{seq}$) and skeleton-level ($\mathcal{L}_{\text{GPC}}^{ske}$) prototype contrastive learning obtains better results than solely using them ($i.e.$, $\alpha = 0.0$ or $\alpha = 1.0$), as shown in Table 4. Our model is not sensitive to the changes of β when fusing structure-prompted and trajectory-prompted reconstruction. As shown in Table 5, $\beta = 0.5$ achieves slightly better performance in average, while a smaller value of β could benefit the model performance on some datasets such as BIWI-S. As the skeleton data of different domains ($e.g.$, datasets) are collected under different conditions, the context of skeleton structure or trajectory may have different contributions on

Table 6. Performance of our approach on different datasets when setting different weight coefficients λ to fuse the graph prototype contrastive learning ($\mathcal{L}_{\text{GPC}}$) and structure-trajectory prompted reconstruction ($\mathcal{L}_{\text{STPR}}$) for model training.

λ	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
0.0	44.9	16.8	29.8	4.3	35.3	23.2	36.4	25.9	18.8	16.0	35.9	15.3
0.2	72.9	45.9	56.4	18.4	48.2	34.0	58.1	39.4	33.5	27.0	67.4	30.3
0.4	72.4	45.1	58.7	21.0	47.4	32.3	59.1	39.9	31.4	26.3	69.1	30.6
0.5	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
0.6	71.6	44.2	58.2	19.6	49.5	33.1	59.8	40.0	32.0	24.6	66.4	30.8
0.8	71.8	41.7	58.1	19.9	48.5	32.7	58.7	39.3	32.4	28.2	65.5	26.5
1.0	71.3	42.5	57.0	18.1	48.0	33.0	56.1	40.2	31.2	26.2	66.6	26.7

Table 7. Performance of our approach on different datasets when setting different numbers of Skeleton Graph Transformer (SGT) layers.

Layer	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
1	69.8	39.0	58.6	21.0	46.8	31.8	58.5	40.3	28.8	25.6	65.8	27.1
2	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	67.8	30.1
3	72.7	41.0	53.3	13.9	49.7	33.7	57.7	40.1	33.2	27.8	66.2	30.0
4	69.0	36.2	52.8	11.8	46.6	29.8	57.0	38.1	31.1	26.1	63.3	24.3

the reconstruction and skeleton semantics learning, thus β can be further selected to facilitate the model training.

Effects of different numbers of attention heads and SGT layers: As shown in Table 7 and 8, setting 2 SGT layers and 8 attention heads per layer enables our model to obtain the best performance on different datasets. Employing too many numbers of layers (4 layers) slightly reduces the performance as it may largely expand the model scale and learn more redundant information. The results also show that our model trained on the large dataset such as KGBD is more sensitive to the layer variation. Since our approach under different numbers of attention heads achieves similar performance, we empirically select $H = 8$ heads to achieve a better trade-off between computational cost and performance.

Effects of other parameters: As shown in Table 9, 11 and 12, our approach is not sensitive to changes of some parameters such as the temperature τ_1 and random mask numbers (a, b). In practice, we select $a = 10$ and $b = 2$ as this setting achieves slightly better performance on different datasets. Although setting different τ_1 value may obtain similar results, we observe that their scales could influence the training stability ($i.e.$, setting too small or too large values induces more evident performance variations). We therefore choose a moderate value for the temperature τ_1. The results in 10 show that TranSG achieves higher performance when setting a relatively higher value for the temperature τ_2, which also improves the training stability ($i.e.$, smaller loss fluctuation) of our model on different datasets. In our experiments, the temperatures are empirically set to $\tau_1 = 0.07$ and $\tau_2 = 14$, and they could be further tuned for better performance.

2.3. Multi-Shot Performance with Different Lengths f

We evaluate the multi-shot performance of our approach with different settings of sequence lengths f ($i.e.$, f-shot person re-ID). Since skeleton sequences contain more pattern features as f increases, our approach is capable of learning more effective skeleton graph representations to achieve larger performance improvement in most cases as shown in 13. Nevertheless, it is interesting to note that using shorter sequences sometimes performs better than longer sequences on small datasets such as IAS-B and BIWI-S, implying that a larger size of available training sequences under smaller f settings could help learn better representations on those datasets. It should be noted that in our paper, we evaluate all compared methods under the same sequence length ($f = 6$) following the literature [11, 14].

2.4. Effects of Positional Encoding

The positional encoding used in our SGT helps preserve the unique positional information of nodes based on the graph structure $i.e.$, structurally nearby nodes are endowed with similar positional features while the farther nodes possess more different positional features [15, 16]. As shown in Table 14, removing the positional encoding can reduce our model performance in most cases, which demonstrates the important role of positional information in the skeleton graph learning of our approach, as it encourages capturing richer structural graph context for relation learning and graph reconstruction. Interestingly, our model achieves similar (slightly lower) performance on the KS20 dataset when removing positional encoding. As each skeleton of KS20 contains more body-joint nodes than that of other datasets, there could be two possible reasons for this result: (1) The positional encoding based on a small number of eigenvectors ($K = 10$) might be insufficient to character-

Table 8. Performance of our approach on different datasets when setting different numbers of attention heads per SGT layer.

H	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
4	73.2	46.2	56.9	18.4	46.7	33.9	57.2	38.0	31.2	26.5	67.5	29.9
8	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
16	72.7	46.4	56.6	17.5	48.5	33.0	58.6	39.5	32.7	26.7	68.5	30.2
32	72.9	44.7	52.6	15.4	49.1	32.0	58.5	39.9	32.5	27.0	66.3	29.5

Table 9. Performance of our approach on different datasets when setting different temperature τ_1.

τ_1	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
0.01	73.1	42.2	59.2	18.1	48.1	32.0	59.4	40.7	31.9	26.6	68.0	27.2
0.1	73.4	45.2	58.0	20.4	49.3	32.6	59.8	41.8	31.5	27.1	69.3	29.5
1.0	72.7	43.9	57.8	19.6	49.6	35.8	59.1	39.6	32.3	27.0	69.1	32.1
10	73.4	41.1	57.0	18.4	50.5	32.0	59.7	40.3	32.7	26.7	66.7	30.2

ize the unique node information from a large skeleton graph, and we can improve K for better learning. (2) Directly introducing positional information into the learning of larger graphs might not be suitable, so a more effective skeleton-based positional encoding mechanism should be devised.

In practice, we found that setting different K ranging from 6 to 14 achieves similar performance, while using $K = 10$ obtains slightly better average performance on different datasets. The results show that our model is not sensitive to the change of eigenvector number. We will further explore the issue (2) in the future work.

2.5. Visualization of Body Relations

As shown in Fig. 9-12, we visualize three learned full-relation (FR) matrices for the same skeleton sequence in different datasets. Note that there are totally $H = 8$ learned relation matrices in our approach and here we only visualize 3 of them. Since each FR head computes f relation matrices corresponding to f skeleton graphs in a sequence, we average them into a matrix to show the mean relations of body-joint nodes. We can observe that FR heads can capture different correlations between different nodes, and they can individually focus on patterns of the same body part correlated with other parts. For example, the 4^{th} FR head trained on BIWI focuses on salient relations between nodes 6-9 and nodes 14-17 (see Fig. 11 (b)), while the 8^{th} head pays more attention to patterns between nodes 6-9 and other body components ($i.e.$, nodes 12-13 and 18-19 (see Fig. 11 (c)). These results demonstrate that the multiple FR heads in SGT can capture different body and motion relations of nodes from different representation subspaces to facilitate learning a better skeleton graph representation.

2.6. Visualization of Training Process

We visualize the total training loss $\mathcal{L}$ in Fig. 4, and the results suggest that our model learning can converge very fast in the first 100 optimization epochs. Meanwhile, the graph prototype contrastive (GPC) loss $\mathcal{L}_{\text{GPC}}$ and structure-trajectory prompted reconstruction (STPR) loss curves $\mathcal{L}_{\text{STPR}}$ show similar learning effects with $\mathcal{L}$, as individually presented in Fig. 5 and 6. This validates our intuition that the graph semantics learning during skeleton reconstruction (STPR) and the discriminative feature learning in the supervised contrastive learning (GPC) are compatible and they can be combined to facilitate the model training. To provide a further analysis of the learned skeleton representations, we follow [14] to estimate the *mean intra-class tightness (mACT)* and *mean inter-class looseness (mRCL)* of the learned skeleton graph representations $w.r.t.$ the ground-truth classes. The mACT and mRCL can serve as effective evaluation metrics of the contrastive representation learning and identity-associated semantics learning[3]. As shown in Fig. 7 and 8, the training of our approach progressively and significantly improves both mACT and mRCL of the learned skeleton graph representations on different datasets, which demonstrates that the proposed TranSG can encourage the model to capture effective class-related semantics ($e.g.$, inter-class differences) to learn more discriminative skeleton representations for person re-ID.

2.7. Confusion Matrix Visualization

As shown in Fig. 13, we visualize the confusion matrices of our approach when performing person re-ID with the Rank-1 matching ($i.e.$, predicting the identity of each probe sequence using the Rank-1 gallery sequence that has the smallest Euclidean distance) on all testing sets (probe sets). Fig. 13 (a)-(f) show that each confusion matrix possesses an evident alignment between the predicted identities and the ground-truth identities on the diagonal line. This suggests that skeleton sequences in most classes can be correctly matched between the probe set and gallery set in each dataset. Moreover, it can be seen that the ratios of classes with high accuracy ($i.e.$, ratios of red grids on the

[3]According to the criterion in [14], a good model should satisfy: The same-class representations are gathered closer (higher mACT) while different-class representations possess larger distances (higher mRCL).

Table 10. Performance of our approach on different datasets when setting different temperature τ_2.

τ_2	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
0.01	70.1	31.0	55.5	13.2	49.5	31.9	55.6	35.1	32.8	26.8	61.1	15.5
0.1	72.3	39.1	56.8	13.1	49.3	29.8	57.8	36.7	29.6	26.6	67.9	27.2
1.0	73.2	43.3	57.0	16.2	48.4	33.3	59.1	37.8	32.8	27.5	68.3	29.6
10	73.4	46.0	58.7	18.5	49.0	32.6	59.0	40.9	33.0	27.2	69.0	31.3

Table 11. Performance of our approach on different datasets when setting different numbers of random structure masks.

a	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
4	72.2	46.3	58.0	19.8	47.2	31.3	58.4	39.6	32.9	28.8	67.8	30.2
8	73.2	47.0	58.2	19.5	48.5	32.9	58.6	41.8	33.5	27.4	69.3	31.0
10	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
12	72.7	45.7	58.8	20.3	48.4	32.6	58.8	40.6	33.1	27.0	69.0	30.5

diagonal line) in KS20 and BIWI-Still are larger than that in IAS-A, IAS-B, KGBD, and BIWI-Walking. The larger numbers of white and red grids diffused *around* the diagonal lines, which represent the higher proportions of false matches, on the matrices of IAS-A (see Fig. 13 (c)) and BIWI-Walking (see Fig. 13 (f)) imply that our model tends to confuse skeleton sequences of more different identities on these datasets. These results are consistent with the performance results shown in the paper.

References

[1] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. J. Finocchio, R. Moore, A. A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1297–1304, 2011. 1

[2] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *International Conference on Pattern Recognition (ICPR)*, vol. 4, pp. 441–444, IEEE, 2006. 1

[3] Z. Liu, Z. Zhang, Q. Wu, and Y. Wang, "Enhancing person re-identification by integrating gait biometric," *Neurocomputing*, vol. 168, pp. 1144–1156, 2015. 1, 2

[4] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognition*, vol. 98, p. 107069, 2020. 1, 3, 4

[5] C.-H. Chen and D. Ramanan, "3D human pose estimation= 2D pose estimation+ matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7035–7043, 2017. 1, 2

[6] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2019. 1, 2

[7] R. Zhao, K. Wang, H. Su, and Q. Ji, "Bayesian graph convolution lstm for skeleton based action recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 6882–6892, 2019. 2

[8] H. Rao, S. Wang, X. Hu, M. Tan, Y. Guo, J. Cheng, X. Liu, and B. Hu, "A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 01, pp. 1–1, 2021. 2, 3, 4

[9] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Multi-level graph encoding with structural-collaborative relation learning for skeleton-based person re-identification," in *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 973–980, 2021. 2, 3

[10] H. Rao, X. Hu, J. Cheng, and B. Hu, "SM-SGE: A self-supervised multi-scale skeleton graph encoding framework for person re-identification," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1812–1820, 2021. 2, 3, 4

[11] H. Rao and C. Miao, "SimMC: Simple masked contrastive learning of skeleton representations for unsupervised person re-identification," in *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1290–1297, 2022. 2, 3, 4, 5

[12] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation OSDI 16)*, pp. 265–283, 2016. 3

[13] H. Rao, S. Wang, X. Hu, M. Tan, H. Da, J. Cheng, and B. Hu, "Self-supervised gait encoding with locality-aware attention for person re-identification," in *International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 1, pp. 898–905, 2020. 3, 4

[14] H. Rao and C. Miao, "Skeleton prototype contrastive learning with multi-level graph relation modeling for unsupervised person re-identification," *arXiv preprint arXiv:2208.11814*, 2022. 3, 4, 5, 6

[15] V. P. Dwivedi, C. K. Joshi, T. Laurent, Y. Bengio, and X. Bresson, "Benchmarking graph neural networks," *arXiv preprint arXiv:2003.00982*, 2020. 3, 5

Table 12. Performance of our approach on different datasets when setting different numbers of random trajectory masks.

b	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	$\mathbf{R}_1$	mAP	$\mathbf{R}_1$	mAP	$\mathbf{R}_1$	mAP	$\mathbf{R}_1$	mAP	$\mathbf{R}_1$	mAP	$\mathbf{R}_1$	mAP
1	72.7	45.3	59.1	20.0	48.7	31.9	57.5	38.6	32.5	26.6	69.2	31.8
2	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
3	73.4	47.3	59.3	19.9	48.5	33.2	58.1	39.6	32.6	27.0	68.5	30.4
4	72.9	44.9	58.7	19.8	49.3	32.5	57.5	38.8	33.0	28.4	69.0	31.8

Table 13. Performance of our approach on different datasets when employing different sequence length f.

f	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	$\mathbf{R}_1$	mAP	$\mathbf{R}_1$	mAP	$\mathbf{R}_1$	mAP	$\mathbf{R}_1$	mAP	$\mathbf{R}_1$	mAP	$\mathbf{R}_1$	mAP
4	70.1	47.5	59.3	22.9	49.1	35.3	56.4	35.1	32.3	27.2	66.8	27.5
6	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
8	75.8	52.6	58.3	18.9	47.9	41.7	61.2	43.7	32.2	33.8	76.6	39.1
10	80.5	49.3	58.5	20.9	49.6	37.2	57.3	50.0	34.8	34.1	63.3	37.4

[16] V. P. Dwivedi and X. Bresson, "A generalization of transformer networks to graphs," in *AAAI Conference on Artificial Intelligence (AAAI) Workshop*, 2021. 3, 5

[17] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, vol. 96, pp. 226–231, 1996. 3

[18] L. Prechelt, "Early stopping-but when?," in *Advances in Neural Information Processing Systems (NeurIPS) Workshop*, pp. 55–69, 1998. 3

[19] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1318–1327, 2017. 3

[20] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, 2015. 3

[21] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, pp. 535–547, 2019. 4

Table 14. Performance of our approach on different datasets when the SGT uses (w/) positional encoding or without (w/o) positional encoding.

Pos. Enc.	KS20		KGBD		IAS-A		IAS-B		BIWI-W		BIWI-S	
	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
w/	73.6	46.2	59.0	20.2	49.2	32.8	59.1	39.4	32.7	26.9	68.7	30.1
w/o	74.1	48.3	49.2	13.2	44.8	29.1	53.3	37.6	32.6	27.6	66.1	28.4

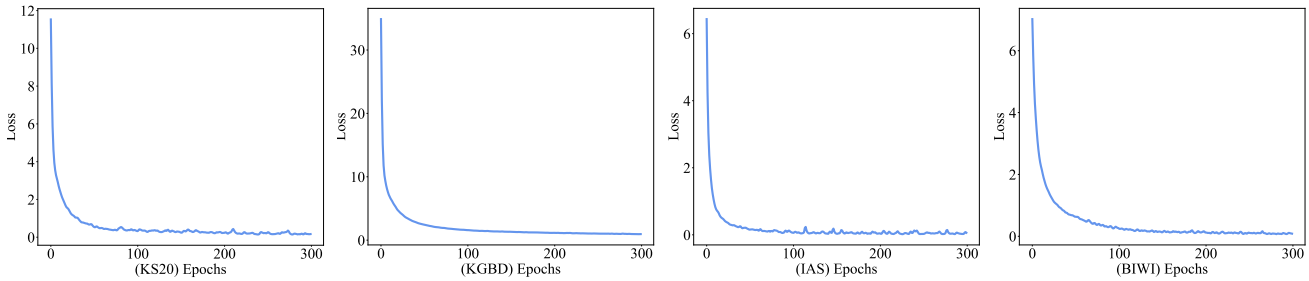


Figure 4. The total training loss curves on different training datasets.

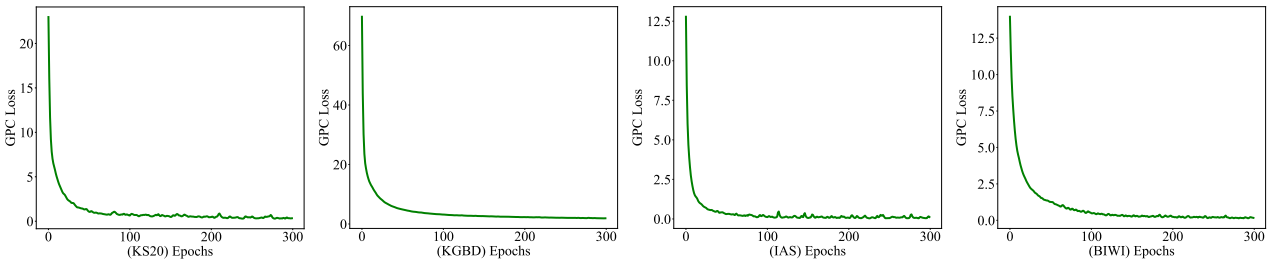


Figure 5. The graph prototype contrastive learning loss ($\mathcal{L}_{\text{GPC}}$) curves on different training datasets.

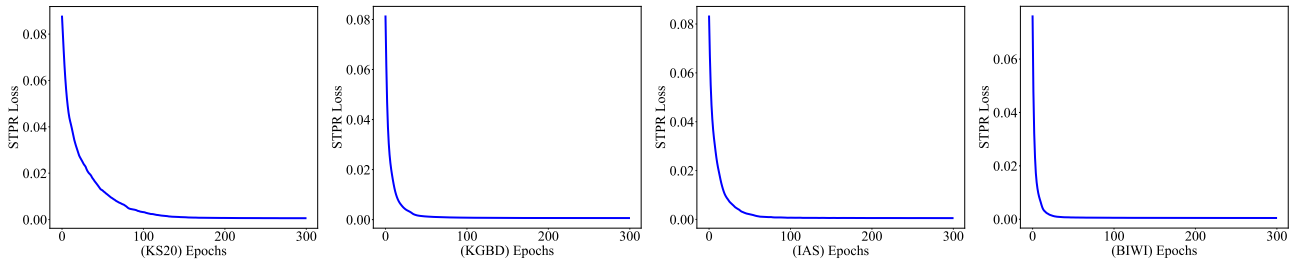


Figure 6. The graph structure-trajectory prompted reconstruction loss ($\mathcal{L}_{\text{STPR}}$) curves on different training datasets.

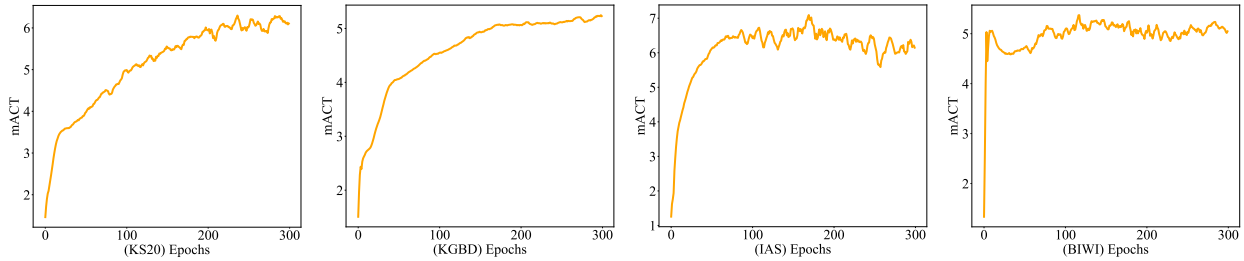


Figure 7. The mean intra-class tightness (mACT) of skeleton representations learned by our approach on different training datasets.

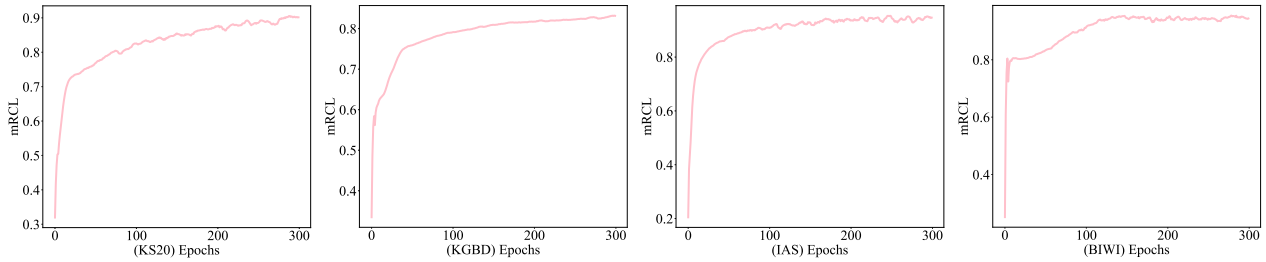


Figure 8. The mean inter-class looseness (mRCL) of skeleton representations learned by our approach on different training datasets.

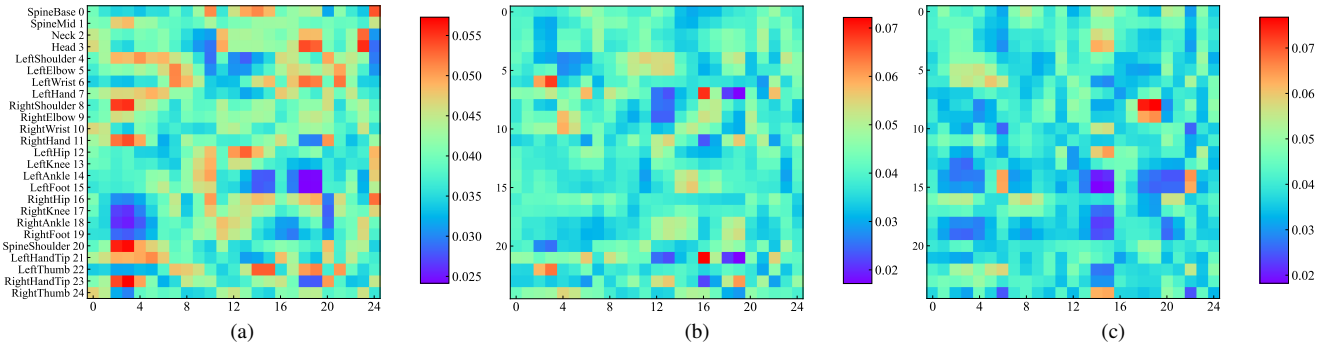


Figure 9. (a) Visualization of full-relation (FR) among body-joint nodes for testing skeletons in KS20. (a)-(b) represent the relations learned by the 1^{st}, 4^{th}, and 8^{th} FR heads. Note that the abscissa and ordinate denote indices of nodes (see Sec. 1.4).

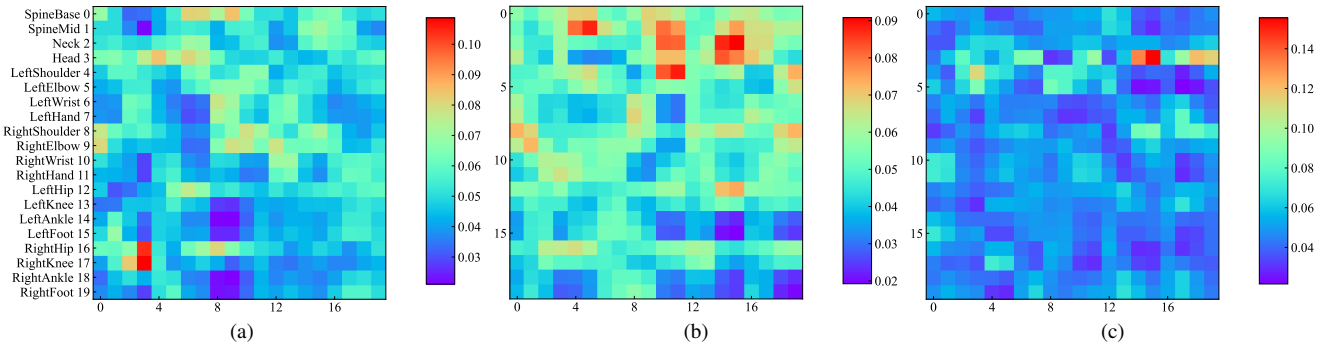


Figure 10. (a) Visualization of full-relation (FR) among body-joint nodes for testing skeletons in IAS. (a)-(b) represent the relations learned by the 1^{st}, 4^{th}, and 8^{th} FR heads. Note that the abscissa and ordinate denote indices of nodes (see Sec. 1.4).

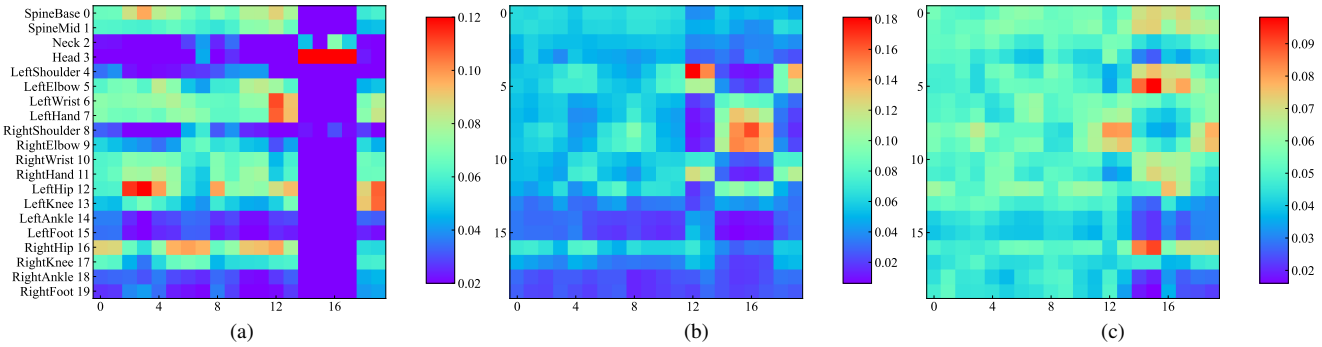


Figure 11. (a) Visualization of full-relation (FR) among body-joint nodes for testing skeletons in BIWI. (a)-(b) represent the relations learned by the 1^{st}, 4^{th}, and 8^{th} FR heads. Note that the abscissa and ordinate denote indices of nodes (see Sec. 1.4).

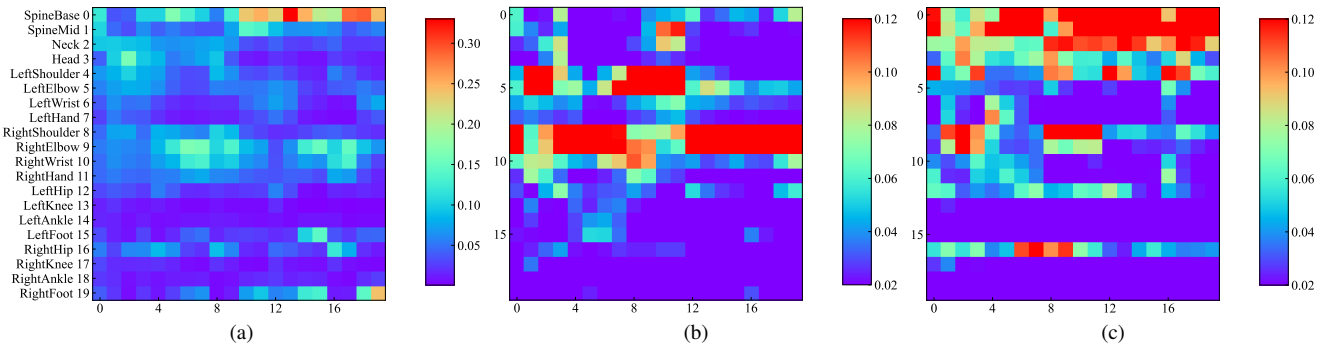
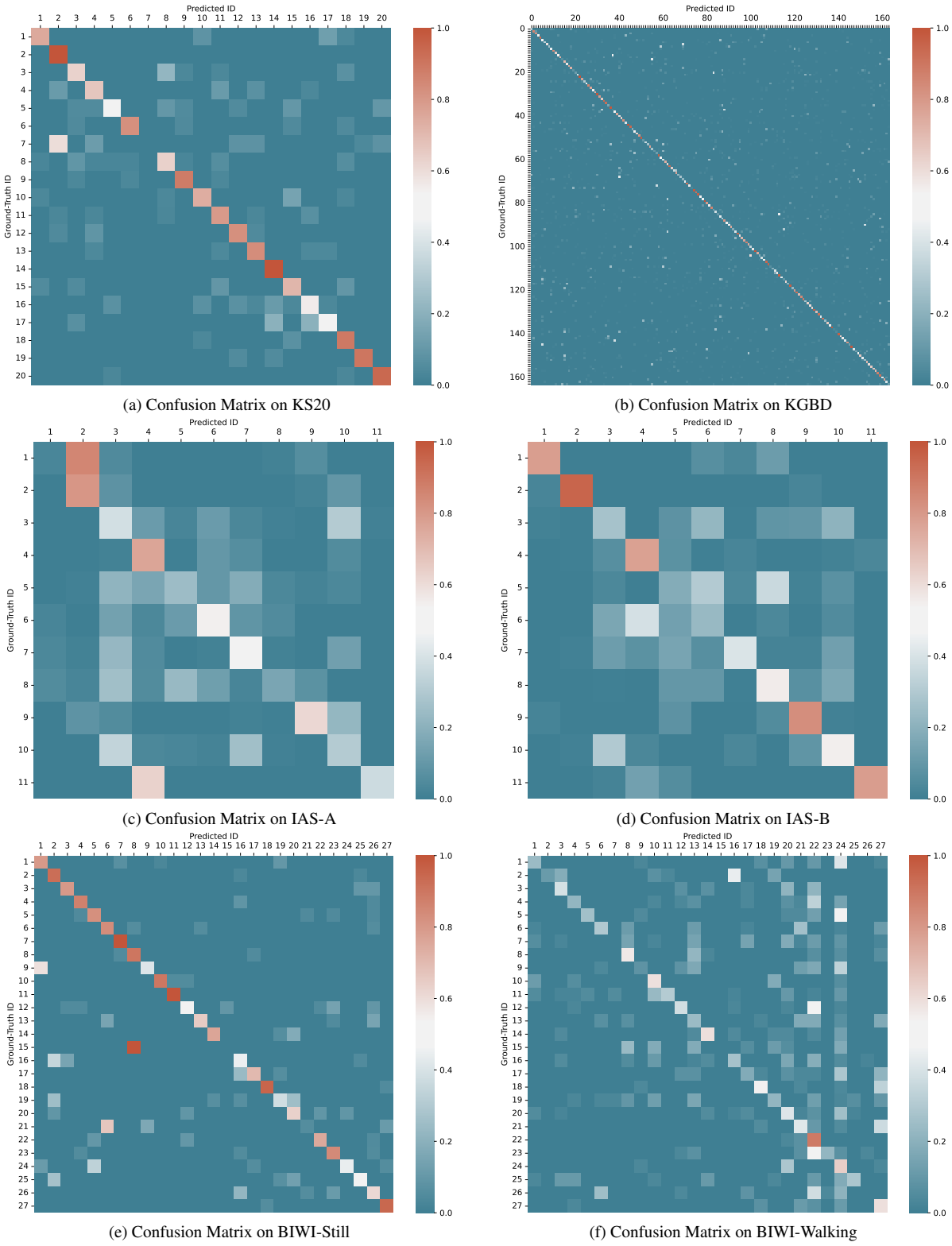


Figure 12. (a) Visualization of full-relation (FR) among body-joint nodes for testing skeletons in KGBD. (a)-(b) represent the relations learned by the 1^{st}, 4^{th}, and 8^{th} FR heads. Note that the abscissa and ordinate denote indices of nodes (see Sec. 1.4).



(a) Confusion Matrix on KS20

(b) Confusion Matrix on KGBD

(c) Confusion Matrix on IAS-A

(d) Confusion Matrix on IAS-B

(e) Confusion Matrix on BIWI-Still

(f) Confusion Matrix on BIWI-Walking

Figure 13. Visualization of confusion matrices on KS20 (a), KGBD (b), IAS-A (c), IAS-B (d), BIWI-Still (e), and BIWI-Walking (f) when using the Rank-1 matching. Note that abscissa and ordinate denote the predicted and ground-truth identities, respectively. The position in the a^{th} row and b^{th} column indicates that the testing samples belonging to the a^{th} identity is predicted as the b^{th} identity, while the corresponding value is the proportion of such samples to the same-identity samples in the testing set.