

# GENERAL SKELETON SEMANTICS LEARNING WITH PROBABILISTIC MASKED CONTEXT RECONSTRUCTION FOR SKELETON-BASED PERSON RE-IDENTIFICATION – APPENDIX I EXPERIMENTS

**Anonymous authors**

Paper under double-blind review

## APPENDIX OUTLINE

The overview for this appendix is presented as follows.

- In Sec. A, we provide details of experimental settings, including dataset description (Table 1), CASIA-B evaluation settings (Sec. A.1), dataset preprocessing strategy (Sec. A.2), probe/gallery settings (Sec. A.3), experimental setup details (Sec. A.4), and utilized computational resources (Sec. A.5)<sup>1</sup>.
- In Sec. B, we provide full experimental results for ablation study (Table 2), effects of hyper-parameters (Sec. B.1), multi-shot performance with different sequence lengths  $f$  (Sec. B.2), and pseudo codes of our method (Sec. B.3).
- In Sec. C, we provide additional visualization and analysis of training metrics (different losses) (Sec. C.1), skeleton representations (Sec. C.2), and confusion metrics (Sec. C.3).
- In Sec. D, we further discuss the broader application and impacts of our method.
- In Sec. E of Appendix I, we provide additional experimental results and analyses based on reviewers' constructive comments and valuable suggestions, including:
  - Table 9: We provide an additional comparison of key differences and similarity between our method (i.e., skeleton-based person re-ID) and skeleton-based gait recognition methods (for Reviewer iRXh).
  - Table 10: We evaluate the performance of different state-of-the-art gait recognition methods (SkeletonGait, GaitTR, GPGait) on all datasets and compare them with our method (for Reviewer iRXh).
  - Table 11: We provide an additional performance comparison of different SSL tasks (DR, MIC, STPR, Prompter) under different skeleton levels (Joint-Level, Part-Level, Body-Level) on different datasets (for Reviewer DvW6).
  - Fig. 11, Fig. 12, and Fig. 13: We offer qualitative examples and analyses for the cross-domain person re-ID performance, including confusion matrices and t-SNE feature visualization (for Reviewer DvW6).
  - Table 12: We provide an additional overview of state-of-the-art skeleton semantics learning (SSL) tasks, their source method, and method types (for Reviewer v2zj).
  - Sec. E.3.1: We offer a detailed comparison between our method and existing state-of-the-art masking strategies (for Reviewers iRXh, BHkC, v2zj).
  - Table 13: We integrate the proposed Prompter into the representative state-of-the-art gait recognition method GPGait, and compare its performance with the original base model on different datasets (for Reviewer DUn6).
  - Table 14: We additionally evaluated the representative state-of-the-art gait recognition method and action recognition method ST-GCN on our benchmark datasets, and integrated the proposed Prompter into them to verify its general applicability (for Reviewers BHkC, iRXh, DvW6).

<sup>1</sup>Our anonymized codes are publicly available at <https://github.com/Anonymous-9273/Prompter>.

Table 1: Overview of datasets (K: thousand). Different testing splits are used to construct gallery sets and probe sets (see Sec. A.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

## A SUPPLEMENTARY EXPERIMENTAL SETTINGS

### A.1 EVALUATION SETTINGS OF CASIA-B

In general, 3D skeleton data in existing skeleton-based person re-ID benchmarks are collected with Kinect (Shotton et al., 2011). 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* (Yu et al., 2006), 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 (Liu et al., 2015), 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 (Liao et al., 2020), we exploit pre-trained pose estimation models (Chen & Ramanan, 2017; Cao et al., 2019) 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 (Cao et al., 2019). Then, we follow the same configuration of estimation in (Liao et al., 2020) 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 (Chen & Ramanan, 2017) 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.

### A.2 DATASET PREPROCESSING

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 the skeleton is translation invariant (Zhao et al., 2019). Then, we split all normalized skeleton sequences in the training sets into multiple shorter skeleton sequences (*i.e.*,  $S$ ) 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  $S$  in our paper refers to those split and

Figure 1: Indices of body joints (nodes) in the estimated skeletons from CASIA-B dataset. Note: All 3D skeletons are estimated from RGB videos of CASIA-B with (Cao et al., 2019) and (Chen & Ramanan, 2017) (see Sec. A.1).

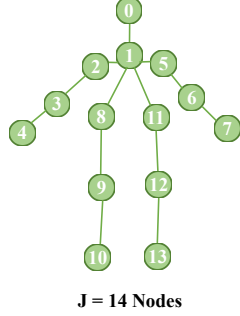


Figure 2: Body-joint indices for joint-level (20 nodes), part-level (10 nodes), and body-level (5 nodes) skeleton representations for IAS, BIWI and KGBD datasets. Our approach *only* requires joint-level skeletons for training, while we evaluate its performance on different-level skeletons following (Rao et al., 2021a; Rao & Miao, 2023) in the paper.

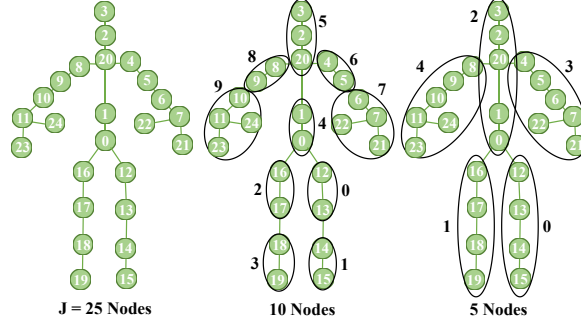
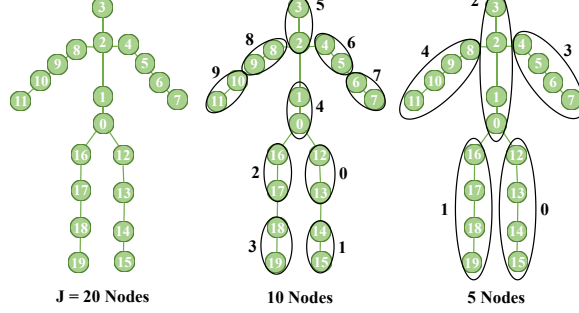


Figure 3: Body-joint indices joint-level (25 nodes), part-level (10 nodes), and body-level (5 nodes) skeleton representations for KS20 dataset. Our approach *only* requires joint-level skeletons for training, while we evaluate its performance on different-level skeleton following (Rao et al., 2021a; Rao & Miao, 2023) in the paper.

normalized sequences used in learning, rather than those original skeleton sequences provided by datasets. We follow the data augmentation strategy used in (Rao et al., 2021b;c) 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.

### A.3 PROBE AND GALLERY SETTINGS

We follow the commonly-used settings of probe and gallery in the literature (Rao & Miao, 2022; Rao & Miao, 2023): 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 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  $0^\circ$ ,  $30^\circ$ ,  $90^\circ$ ,  $130^\circ$ , and  $180^\circ$ . 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 (Liu et al., 2015) to evaluate the proposed skeleton-based approach on CASIA-B (detailed in Sec. A.1).

Table 2: Ablation study with different configurations: Random spatial masking (SM) or temporal masking (TM) with fixed mask numbers, probabilistic spatial context masking (PSCM) or probabilistic temporal context masking (PTCM). “+” indicates employing the corresponding component, and “+ PSCM + PTCM” denotes the final configuration of Prompter.

ID	Config.	KS20		IAS-A		IAS-B		BIWI-S		BIWI-W		KGBD	
		mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>
1	Baseline	42.5	71.3	31.8	48.0	37.9	56.1	26.7	66.6	25.5	31.2	18.1	57.0
2	+ SM	44.8	71.9	32.4	48.7	38.1	57.2	27.0	65.8	26.2	31.9	18.9	57.4
3	+ PSCM	46.5	73.1	33.5	49.4	42.1	58.7	29.3	66.0	25.0	32.8	16.5	56.2
4	+ TM	45.4	73.0	32.1	48.4	39.2	58.2	28.2	67.8	25.9	31.1	19.6	58.0
5	+ PTCM	46.4	73.6	33.8	49.0	42.0	58.9	29.8	67.2	24.7	33.0	18.0	56.3
6	+ SM + TM	46.2	73.6	32.8	49.2	39.4	59.1	30.1	68.7	26.9	32.7	20.2	59.0
7	+ PSCM + PTCM	48.3	74.2	34.1	49.5	43.8	60.4	30.3	66.8	27.3	34.6	21.3	59.5

Table 3: Performance of our method applied to MG-SCR (Rao et al., 2021c), SPC-MGR (Rao & Miao, 2022), and TranSG (Rao & Miao, 2023) on different datasets when setting different values for the spatial context masking probability in PSCM ( $p_s = 0.0, 0.2, 0.4, 0.5, 0.6, 0.8$ ).

Applied Model	$p_s$	IAS-A		IAS-B		BIWI-S		BIWI-W		KS20	
		mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>
MG-SCR	0.0	18.7	51.4	17.7	48.6	10.6	37.9	12.3	31.9	12.2	52.9
	0.2	19.7	51.0	22.8	51.9	12.3	34.5	12.7	36.4	11.6	53.3
	0.4	18.7	51.2	18.7	50.8	13.3	37.2	13.2	38.0	11.8	54.7
	0.5	20.1	52.7	20.5	51.7	13.4	37.6	13.5	39.5	13.2	56.3
	0.6	19.5	51.9	18.8	49.1	12.9	36.9	13.4	38.1	13.0	55.5
	0.8	19.2	48.9	20.6	49.8	11.9	35.1	13.4	39.0	11.1	52.1
SPC-MGR	0.0	25.4	49.2	29.5	51.2	13.5	33.6	17.1	38.1	25.1	65.8
	0.2	26.0	50.1	24.2	48.5	13.7	37.7	17.4	38.5	23.6	64.5
	0.4	25.8	48.0	28.6	50.1	14.5	38.5	17.3	38.0	22.3	64.3
	0.5	27.1	49.8	28.1	51.0	15.0	37.7	18.9	37.0	23.7	65.0
	0.6	26.5	49.5	28.8	51.4	17.8	39.1	17.8	37.5	24.6	66.6
	0.8	27.5	50.5	27.0	49.5	17.3	38.5	15.5	37.1	24.3	65.8
TranSG	0.0	33.0	45.9	41.8	56.9	27.2	59.9	21.6	33.6	46.0	72.3
	0.2	34.4	49.0	42.3	57.9	27.5	64.1	29.9	33.5	48.1	74.6
	0.4	35.0	50.2	43.6	58.3	28.9	67.0	27.8	35.9	46.5	73.8
	0.5	34.1	49.5	43.8	60.4	30.3	66.8	27.3	34.6	48.3	74.2
	0.6	33.9	49.9	44.5	59.6	28.3	64.5	25.5	35.1	47.6	73.2
	0.8	34.3	49.3	44.1	57.7	27.2	62.5	23.5	32.6	43.8	71.3

#### A.4 EXPERIMENTAL SETUP DETAILS

**Skeleton Semantics Learning (SSL) Tasks.** For generality assessment, we compare all existing SSL tasks for skeleton-based person re-ID: DR, AR (Rao et al., 2020), AR + AC (Rao et al., 2021b), MSSP (Rao et al., 2021c), MSR (Rao et al., 2021a), MIC (Rao & Miao, 2022), STPR (Rao & Miao, 2023), and the proposed Prompter. For empirical evaluation on varying models and datasets (which requires the SSL task possesses co-training compatibility (CTC)), We compare our method (Prompter) with three state-of-the-art SSL tasks, including DR, MIC (Rao & Miao, 2022), STPR (Rao & Miao, 2023), which can be co-trained with different models without requiring significant architecture modification. For all compared SSL tasks, we adopt the optimal setting used in the original papers.

**Base Models.** We choose four state-of-the-art skeleton-based person re-ID models MG-SCR (Rao et al., 2021c), SPC-MGR (Rao & Miao, 2022), SimMC (Rao & Miao, 2022), and TranSG (Rao & Miao, 2023) as the base models to evaluate the effectiveness and generality of SSL tasks. For MG-SCR, SimMC, and TranSG that contain SSL tasks, we replace the original SSL task with DR, MIC, STPR or the proposed Prompter. For SPC-MGR that only possesses a downstream objective loss, we add DR, MIC, STPR or Prompter as the SSL objective to co-train the model, and we apply DR, MIC, STPR or Prompter to each level skeleton representation of SPC-MGR for reconstruction. As SimMC directly learns skeleton representations instead of body-joint representations, we slightly modify STPR and Prompter to randomly mask each skeleton representation to perform their reconstruction. Since this reconstruction way is not a standard form of STPR and Prompter, and may not fully reflect their actual effectiveness due to the *integrated* spatial-temporal skeleton reconstruction (*i.e.*, without separately reconstructing each motion trajectory), we only use their performance as a reference for comparison with other SSL tasks. It is worth noting that under this same modification we are able to

Table 4: Performance of our method applied to MG-SCR (Rao et al., 2021c), SPC-MGR (Rao & Miao, 2022), and TranSG (Rao & Miao, 2023) on different datasets when setting different values for the temporal context masking probability in PTCM ( $p_t = 0.0, 0.2, 0.4, 0.5, 0.6, 0.8$ ).

Applied Model	$p_t$	IAS-A		IAS-B		BIWI-S		BIWI-W		KS20	
		mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>
MG-SCR	0.0	19.0	49.4	16.7	47.7	13.1	37.3	13.0	37.7	11.4	51.4
	0.2	22.5	52.4	18.7	48.4	12.7	36.8	12.9	37.1	12.2	53.3
	0.4	22.3	52.9	20.2	50.2	14.4	39.2	13.3	39.5	11.5	53.3
	0.5	22.1	53.1	20.5	51.7	13.4	37.6	13.5	39.5	13.2	56.3
	0.6	22.5	52.6	18.4	49.4	13.2	37.8	13.0	37.2	12.4	54.7
	0.8	19.1	48.9	18.2	47.8	13.8	38.6	14.4	38.2	13.0	55.0
SPC-MGR	0.0	26.7	50.5	24.0	46.6	13.8	37.5	14.7	36.4	22.6	61.7
	0.2	29.1	50.8	26.5	51.2	13.8	36.0	19.0	38.3	24.7	66.0
	0.4	28.7	50.2	28.7	52.2	16.7	38.1	16.7	37.7	24.9	66.2
	0.5	27.1	49.8	28.1	51.0	15.0	37.7	18.9	37.0	23.7	65.0
	0.6	24.4	47.2	26.0	49.9	14.6	38.5	17.2	35.9	23.8	64.3
	0.8	25.2	47.5	27.3	50.6	16.5	37.5	16.9	37.1	22.8	62.9
TranSG	0.0	33.0	48.0	40.7	54.8	30.0	60.6	19.0	34.5	45.4	73.1
	0.2	33.7	48.8	43.4	61.1	31.0	67.1	18.1	33.7	47.1	74.2
	0.4	34.2	49.2	44.2	59.3	29.5	65.5	21.1	35.0	46.8	73.5
	0.5	34.1	49.5	43.8	60.4	30.3	66.8	27.3	34.6	48.3	74.2
	0.6	34.2	50.1	43.9	58.9	30.1	65.9	18.6	34.7	48.8	72.7
	0.8	32.0	49.7	41.8	56.0	27.9	62.2	22.8	35.6	47.9	72.9

Table 5: Performance of our method applied to MG-SCR (Rao et al., 2021c), SPC-MGR (Rao & Miao, 2022), and TranSG (Rao & Miao, 2023) on different datasets when setting different coefficients ( $\alpha = 0.0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0$ ) to combine spatial and temporal context reconstruction.

Applied Model	$\alpha$	IAS-A		IAS-B		BIWI-S		BIWI-W		KS20	
		mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>
MG-SCR	0.0	20.6	52.0	17.9	50.5	13.9	38.0	13.7	38.3	14.9	56.5
	0.2	18.8	51.2	18.2	49.5	14.2	38.1	13.1	37.6	12.6	55.7
	0.4	19.6	52.0	19.7	50.9	12.9	36.7	13.8	38.9	12.5	56.4
	0.5	22.1	53.1	20.5	51.7	13.4	37.6	13.5	39.5	13.2	56.3
	0.6	22.5	54.3	19.3	49.7	13.4	38.5	12.9	37.8	12.6	54.1
	0.8	22.2	54.2	18.8	48.1	12.5	36.2	12.8	36.8	12.1	51.6
	1.0	22.4	55.4	18.8	49.4	13.8	36.5	13.5	36.6	11.5	50.4
SPC-MGR	0.0	27.1	50.6	28.3	52.5	12.8	33.4	15.1	36.7	23.8	63.5
	0.2	27.7	50.2	25.0	49.7	15.5	36.5	15.3	35.6	23.5	63.1
	0.4	27.2	49.3	29.4	50.9	17.5	38.0	17.5	38.5	24.2	64.5
	0.5	27.1	49.8	28.1	51.0	15.0	37.7	18.9	37.0	23.7	65.0
	0.6	25.0	48.7	28.8	49.9	14.8	36.6	15.8	37.4	23.6	66.2
	0.8	25.4	48.1	28.5	50.9	15.8	38.9	15.1	37.7	22.3	63.1
	1.0	27.1	48.2	28.1	49.5	16.3	38.5	14.5	34.5	24.2	63.9
TranSG	0.0	35.1	49.4	43.9	59.5	31.3	61.7	20.7	33.8	46.3	73.6
	0.2	34.8	49.3	42.5	57.0	29.2	64.3	20.8	35.5	46.6	72.9
	0.4	32.0	49.7	42.2	59.0	28.4	64.4	25.7	34.5	47.0	71.9
	0.5	34.1	49.5	43.8	60.4	30.3	66.8	27.3	34.6	48.3	74.2
	0.6	34.0	50.0	43.3	59.7	30.3	65.4	24.5	35.0	48.5	74.4
	0.8	34.2	49.2	45.0	58.2	28.6	63.4	20.9	35.5	49.0	73.1
	1.0	33.3	48.5	42.1	58.7	28.8	63.9	20.2	34.1	46.5	73.1

compare the effectiveness of the key component of Prompter, probabilistic random context masking, with the direct random masking of the state-of-the-art SSL task STPR.

**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. To verify the generality of our method when applied to different-level skeleton representations, we follow (Rao et al., 2021a; Rao & Miao, 2023) to construct another two levels, namely part-level (10 nodes) and body-level (5 nodes), by merging joints within different body partitions. The original skeletons, part-level skeletons, and body-level skeletons are 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 (Rao & Miao, 2022) 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  $f = 40$ . We construct the reconstructing and inferring models by using MLP networks with one hidden layer, and the embedding size is set

Table 6: Performance of our method applied to MG-SCR (Rao et al., 2021c), SPC-MGR (Rao & Miao, 2022), and TranSG (Rao & Miao, 2023) on different datasets when employing different sequence length ( $f = 4, 6, 8, 10$ ).

Applied Model	$f$	IAS-A		IAS-B		BIWI-S		BIWI-W		KS20	
		mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>
MG-SCR	4	21.1	53.8	18.6	50.2	13.8	38.9	13.2	39.9	11.4	54.6
	6	20.1	52.7	20.5	51.7	13.4	37.6	13.5	39.5	13.2	56.3
	8	21.3	52.5	18.1	49.7	15.0	39.0	14.1	34.3	13.5	57.4
	10	21.3	49.9	20.4	50.1	16.0	41.0	14.2	33.1	15.5	60.9
SPC-MGR	4	21.9	44.6	23.6	48.2	15.4	40.0	13.3	32.2	24.6	65.0
	6	27.1	49.8	28.1	51.0	15.0	37.7	18.9	37.0	23.7	65.0
	8	28.7	50.2	31.2	53.2	15.0	36.0	18.5	38.5	25.0	68.4
	10	26.7	49.9	29.7	51.7	16.0	37.5	20.9	38.7	26.4	68.4
TranSG	4	35.4	46.8	40.3	55.0	29.0	62.9	22.9	37.7	49.6	72.8
	6	34.1	49.5	43.8	60.4	30.3	66.8	27.3	34.6	48.3	74.2
	8	40.2	51.4	43.2	61.1	34.0	69.7	21.1	37.2	53.7	74.2
	10	40.4	48.6	46.0	57.9	47.6	62.1	35.1	39.8	43.9	77.3

Table 7: Performance of our method applied to TranSG (Rao & Miao, 2023) on different datasets when setting different coefficients ( $\lambda = 0.00, 0.25, 0.50, 0.75, 1.00$ ) to fuse downstream task objective loss (GPC) and SSL objective loss (Prompter).

$\lambda$	IAS-A		IAS-B		BIWI-S		BIWI-W		KS20	
	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>
0.00	22.0	35.0	23.0	34.5	12.8	31.8	13.1	26.2	16.4	47.3
0.25	33.8	49.7	40.9	58.3	29.9	66.0	20.6	35.7	47.3	72.7
0.50	34.1	49.5	43.8	60.4	30.3	66.8	27.3	34.6	48.3	74.2
0.75	34.3	49.5	43.4	57.5	29.7	65.7	28.0	34.8	47.6	73.8
1.00	31.8	48.0	37.9	56.1	26.7	66.6	25.5	31.2	42.5	71.3

to the same value as the skeleton feature size used in the original models. The probabilities for spatial and temporal context masking are empirically set to  $p_s = p_t = 0.5$ , and we use  $\alpha = 0.5$  to equally combine spatial and temporal skeleton context reconstruction. We empirically adopt  $\lambda = 0.5$  to combine the SSL objective and the downstream objective to co-train the model, as this setting achieves the best average performance on different datasets. 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 stable 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. An Adam optimizer with the learning rate of  $3.5 \times 10^{-4}$  is used for the model optimization. For batch sizes, we follow the setting used in the original models: In TranSG (Rao & Miao, 2023), batch size is 256 for all datasets; In SPC-MGR (Rao & Miao, 2022) and MG-SCR (Rao et al., 2021c), batch size is set to 256 for KGBD and 128 for other datasets. To avoid over-fitting and achieve better generalization performance, we adopt Early Stopping (Prechelt, 1998) with a patience of 150 epochs (*i.e.*, stop the training of model after no improvement in 150 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.

For all methods compared in our experiments, we select optimal model parameters for training, and use their pre-defined 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 (Rao et al., 2021b; Rao & Miao, 2022) to report the average performance of all methods. Note that our approach does not use any post-processing technique, *e.g.*, re-ranking (Zhong et al., 2017) or multi-query fusion (Zheng et al., 2015) in the training or testing stage. To perform person re-ID, we encode each original skeleton sequence of the probe set  $\Phi_P$  without using masking into corresponding sequence-level graph representations,  $\{\mathbf{V}_i^P\}_{i=1}^{N_2}$ , and match it with representations,  $\{\mathbf{V}_i^G\}_{i=1}^{N_3}$ , of the same identity in the gallery set  $\Phi_G$  using Euclidean distance.

**Generality Assessment Details.** For co-training compatibility (CTC) score ( $G_C$ ) computation, we evaluate the qualified SSL tasks (DR, MIC, STPR, Prompter) on four state-of-the-art models

Table 8: Performance of our method applied to SPC-MGR (Rao & Miao, 2022) on different datasets when setting different coefficients ( $\lambda = 0.00, 0.25, 0.50, 0.75, 1.00$ ) to fuse downstream task objective loss (SPC) and SSL objective loss (Prompter).

$\lambda$	IAS-A		IAS-B		BIWI-S		BIWI-W		KS20	
	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>	mAP	R <sub>1</sub>
<b>0.00</b>	16.8	34.6	19.8	45.6	11.3	36.1	17.8	17.8	14.4	47.5
<b>0.25</b>	26.1	47.3	27.3	52.6	14.2	38.0	16.7	36.6	23.9	64.8
<b>0.50</b>	27.1	49.8	28.1	51.0	15.0	37.7	18.9	37.0	23.7	65.0
<b>0.75</b>	23.5	43.8	28.4	51.6	14.1	38.2	16.6	37.3	24.3	66.0
<b>1.00</b>	24.2	41.9	24.1	43.3	16.0	34.1	19.4	18.9	21.7	59.0

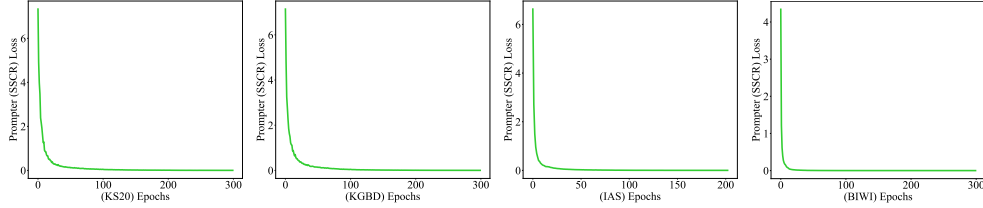


Figure 4: The SSL loss (Prompter (SSCR) loss) curves on different training datasets when applying Prompter to TranSG (Rao & Miao, 2023).

(MG-SCR, SPC-MGR, SimMC, TranSG) on all benchmark datasets (KS20, BIWI, IAS, KGBD) following the empirical evaluation in the paper. For spatial-temporal effectiveness (STE) score ( $G_{ST}$ ) computation, we evaluate STPR and Prompter under the same setting of base model TranSG on all benchmark datasets (KS20, BIWI, IAS, KGBD).

**Ablation Study Details.** In the ablation study, we adopt the state-of-the-art model TranSG (Rao & Miao, 2023) without employing any SSL tasks as the baseline. For the configurations of “+ SM” and “+ TM”, we follow (Rao & Miao, 2023) to set the optimal mask numbers for random spatial masking or temporal masking. The “+ PSCM” or “+ PTCM” denotes only using the spatial (structural locations) or temporal context (motion trajectories) reconstruction with the proposed probabilistic masking in Prompter.

**Cross-Domain Person Re-ID Details.** The model applying a certain SSL task (STPR, MIC, DR, Prompter) is trained on the training set of source dataset, and tested on the testing sets (*i.e.*, probe set and gallery set) of target dataset without model fine-tuning, following the cross-domain (also termed domain-generalized) person re-ID protocol in (Rao et al., 2021b; Rao & Miao, 2022). As both BIWI and IAS datasets possess the same structure of input skeleton data, the pre-trained model can be directly transferred to other datasets. For example, “IAS $\rightarrow$ W” denotes training the model on the IAS training set and evaluating it on the probe set BIWI-W (corresponding to the gallery set BIWI-S). We adopt the state-of-the-art model TranSG with the original SSL task (STPR) as the base model for comparison in the paper.

**Other Details.** In the experiments of Transferring Prompter to Different Skeleton Modeling, we follow (Rao et al., 2021a; Rao & Miao, 2023) to construct different level skeleton representations, namely joint-level (corresponding to joint-scale in (Rao et al., 2021a)), part-level (corresponding to part-scale), body-level (corresponding to body-scale) representations, as visualized in Fig. 2 and Fig. 3. We train the base model TranSG on each individual skeleton level by applying MIC, DR, STPR or Prompter to compare their performance. For the loss curves provided in the last part, we visualize the average loss of each continuous training batch when applying *only* Prompter to TranSG on KS20. The DR and STPR losses are directly computed under the training of Prompter, instead of using them as optimization objectives. We repeat experiments for multiple time to report their average loss changes. It is worth noting that for a more comprehensive and intuitive comparison, we follow the STPR and Prompter (SSCR) losses to equally combine spatial and temporal skeleton reconstruction losses as the final DR loss.



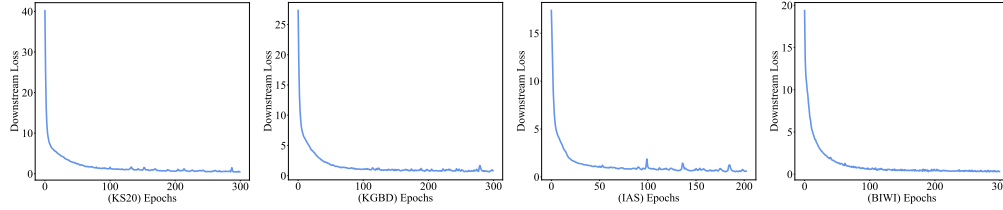


Figure 5: The downstream loss (GPC loss) curves on different training datasets when applying Prompter to TranSG (Rao & Miao, 2023).

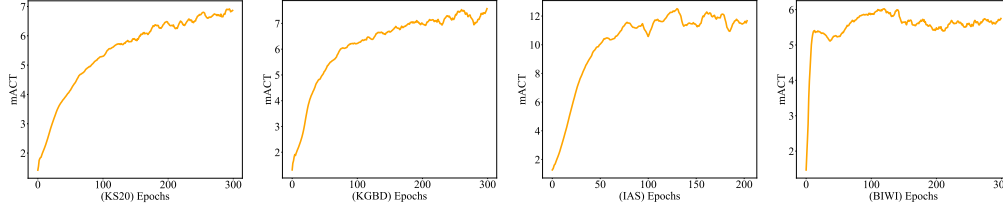


Figure 6: The mean intra-class tightness (mACT) of skeleton representations on different training datasets when applying Prompter to TranSG (Rao & Miao, 2023) .

#### A.5 UTILIZED COMPUTATIONAL RESOURCES

Our experiments are run on  $1 \times$  NVIDIA Tesla V100 (32GB) or P100 (16GB) GPUs with the  $2 \times$  Intel(R) Xeon(R) Gold 6148 CPU @2.40GHz. In practice, co-training existing skeleton-based person re-ID models with our method does not require the whole above resources, and the utilized computational resources mainly depend on the original model, as our method only introduces a small number of extra parameters (see our paper). Multiple experiments with the same or different models can be parallelly conducted on our device.

## B SUPPLEMENTARY EXPERIMENTAL RESULTS

### B.1 EFFECTS OF DIFFERENT HYPER-PARAMETERS

#### Effects of spatial context masking probability $p_s$ and temporal context masking probability $p_t$ :

As shown in Table 3 and Table 4, Prompter achieves slightly better average performance on different datasets when setting the masking spatial or temporal context masking probability to 0.5. According to this observation, we empirically set  $p_s = p_t = 0.5$  in all models. However, in practice, it is observed that two different masking probabilities (not close ones) can achieve similar performance in some cases, and training with the same masking probability could yield slightly different results in multiple experiments. This is because of the randomness nature of Prompter that it introduces more random combinations of positions into training (demonstrated in our paper). As the convergence of the model may simultaneously be influenced by the random parameter initialization and the masking probability used in Prompter, we repeat experiments with the same setting for multiple times and report the average performance.

**Effects of weight coefficient  $\alpha$ :** As presented in Table 5, using only temporal masked reconstruction (*i.e.*,  $\alpha = 0.0$ ) can achieves slightly better performance than using only spatial masked reconstruction (*i.e.*,  $\alpha = 1.0$ ) in most cases, which suggests that solely using temporal masked reconstruction is more effective than solely employing the spatial masked reconstruction on learn useful discriminative skeleton semantics for person re-ID. Notably, combining both of them with  $\alpha$  values around 0.5 can achieve higher performance in average. This is also consistent with our analysis in our paper that spatial and temporal masked reconstruction are compatible and can facilitate each other to learn better skeleton representations (*i.e.*, STE property) for person re-ID. Based on this reason, we empirically set  $\alpha = 0.5$  in our experiments. Nevertheless, as the skeleton data of different domains (*e.g.*, datasets) are collected under different conditions, the context of skeletal spatial structure or temporal trajectory



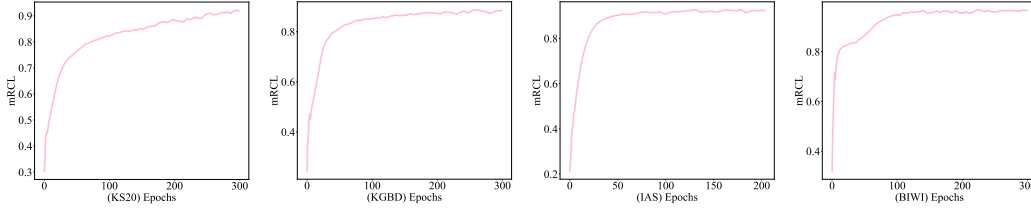


Figure 7: The mean inter-class looseness (mRCL) of skeleton representations on different training datasets when applying Prompter to TranSG (Rao & Miao, 2023).

may have different contributions on the reconstruction and skeleton semantics learning, thus  $\alpha$  can be further selected to facilitate the model training.

**Effects of weight coefficient  $\lambda$ :** The results in Table 7 and 8 show that combining the proposed SSL task Prompter and the original downstream objective with an appropriate  $\lambda$  can facilitate the model to learn more effective representations and achieve higher performance in terms of Rank-1 accuracy and mAP. In our experiments, we follow existing works to equally fuse them. Interestingly, only applying the SSL task to perform probabilistic masked spatial-temporal reconstruction without using downstream objective can still learn useful skeleton features for person re-ID despite with significantly lower accuracy. This may suggest the enhanced effectiveness of Prompter when being combined with different downstream objectives to learn more general class-agnostic skeleton semantics and discriminative skeleton features. It also shows the higher contribution of downstream tasks since they typically utilize more discriminative supervision with ground-truth labels (Rao & Miao, 2023).

## B.2 MULTI-SHOT PERFORMANCE WITH DIFFERENT LENGTHS

We evaluate the multi-shot performance of our method when applied to different models with different settings of sequence lengths  $f$  (*i.e.*,  $f$ -shot person re-ID). Since skeleton sequences contain more pattern features as  $f$  increases, the model is capable of learning more effective skeleton representations to achieve larger performance improvement in most cases as shown in Table 6. Nevertheless, it is interesting to note that using shorter sequences performs better than longer sequences on some datasets such as BIWI-S and BIWI-W in some cases, implying that a larger size of available training sequences under smaller  $f$  settings may 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 (Rao & Miao, 2022; Rao & Miao, 2022; Rao & Miao, 2023).

## B.3 PSEUDO CODES OF PROMPTER

In this section, we provide python-style pseudo codes of our method, which is structural, concise and can be flexibly integrated into different models. The formal codes, data, and models are publicly released in <https://github.com/Anonymous-9273/Prompter>.

```
# Independently and randomly sample J spatial masks from Bernoulli(1 -
    prob_s), prob_s is the probability for spatial context masking of
    skeletal structural locations
def PSCM(prob_s):
    s_mask = np.zeros([J, ])
    # number of zero masks is from 0 to J-1
    while np.mean(s_mask) == 0:
        prob = np.random.uniform(0, 1, [J, ])
        s_mask = prob >= prob_s
    return s_mask

# Independently and randomly sample f temporal masks from Bernoulli(1 -
    prob_t), prob_t is the probability for temporal context masking of
    skeletal motion trajectories
def PTCM(prob_t):
    t_mask = np.zeros([f, ])
    # number of zero masks is from 0 to f-1
```

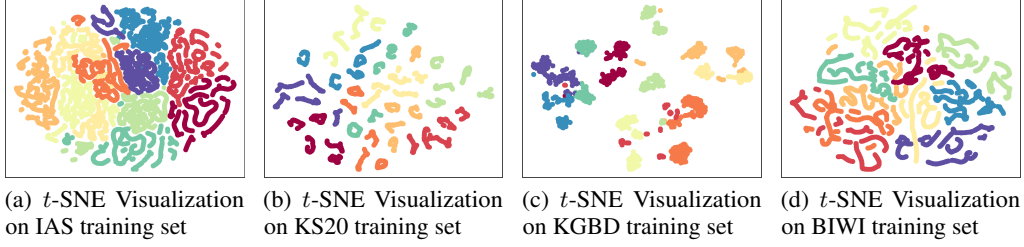


Figure 8: *t*-SNE visualization for the first ten classes of training sets: IAS (a), KS20 (b), KGBD (d), and BIWI (d). Different colors indicate representations of different classes.

```

502 while np.mean(t_mask) == 0:
503     prob = np.random.uniform(0, 1, [f, ])
504     t_mask = prob >= prob_t
505     return t_mask
506
507 # Spatial skeleton sequence reconstruction and inference based on PSCM
508 # h with shape [batch_size, f, J, H]
509 # spatial_mask with shape [batch_size, J], generated by PSCM(prob_s)
510 # gt_pos with shape [batch_size, f, J, 3]
511 def skeleton_recon_loss(h, spatial_mask, gt_pos):
512     # Apply masks to structural locations of joints in the skeleton and
513     # average unmasked representations
514     mask_h = apply_mask_and_ave(h, spatial_mask)
515     # Use MLP to predict ground-truth structural locations of joints
516     [batch_size, f, J, 3]
517     pred_pos = MLP(mask_h)
518     # Compute MSE loss between predicted structural locations and
519     # ground-truth structural locations of skeleton sequences
520     s_recon_loss = MSE_loss(pred_pos, gt_pos) / batch_size
521     return s_recon_loss
522
523 # Temporal skeleton sequence reconstruction and inference based on PTCM
524 # traj_h with shape [batch_size, J, f, H]
525 # temporal_mask with shape [batch_size, f], generated by PTCM(prob_t)
526 # gt_pos with shape [batch_size, f, J, 3]
527 def trajectory_recon_loss(traj_h, temporal_mask, gt_pos):
528     # Apply masks to motion trajectories of joints and average unmasked
529     # representations
530     mask_traj_h = apply_mask_and_ave(traj_h, temporal_mask)
531     # Use MLP to predict ground-truth motion trajectories of joints
532     [batch_size, J, f, 3]
533     pred_pos = MLP(mask_traj_h)
534     # Transpose shape [batch_size, J, f, 3] to shape [batch_size, f, J,
535     # 3] to match the original skeleton sequence shape
536     pred_pos = transpose(pred_pos, [0, 2, 1, 3])
537     # Compute MSE loss between predicted motion trajectories and
538     # ground-truth motion trajectories of skeleton sequences
539     t_recon_loss = MSE_loss(pred_pos, gt_pos) / batch_size
540     return t_recon_loss
541
542 # Spatial-temporal Skeleton Context Reconstruction (SSCR) loss for
543 # Prompter learning
544 def Prompt_loss():
545     return alpha * skeleton_recon_loss(h, spatial_mask, gt_pos) + (1 -
546         alpha) * trajectory_recon_loss(traj_h, temporal_mask, gt_pos)

```

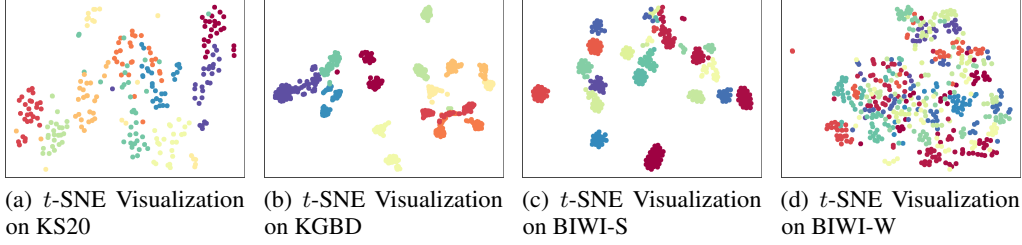


Figure 9:  $t$ -SNE visualization for the first ten classes of testing sets (probe sets): KS20 (a), KGBD (b), BIWI-S (c), and BIWI-W (d). Different colors indicate representations of different classes.

## C SUPPLEMENTARY VISUALIZATION AND ANALYSIS

### C.1 VISUALIZATION OF TRAINING PROCESS

In Fig. 4, we visualize the Prompter loss  $\mathcal{L}_{SSCR}$  when applying it to the TranSG model. The results suggest that the Prompter learning can converge very fast in the first 50 optimization epochs, while the downstream loss ( $\mathcal{L}_{GPC}$  (Rao & Miao, 2023)) curve shows similar learning effects with  $\mathcal{L}_{SSCR}$ , as presented in Fig. 5. This validates our intuition that the skeleton semantics learning of Prompter and the discriminative feature learning of downstream objective 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 (Rao & Miao, 2022) to estimate the *mean intra-class tightness* ( $mACT$ ) and *mean inter-class looseness* ( $mRCL$ ) of the learned skeleton representations *w.r.t.* the ground-truth classes. The  $mACT$  and  $mRCL$  can serve as effective evaluation metrics of the skeleton representation learning and identity-associated semantics learning<sup>2</sup>. As shown in Fig. 6 and 7, the training of our approach progressively and significantly improves both  $mACT$  and  $mRCL$  of the learned skeleton representations on different datasets, which demonstrates that co-training TranSG with Prompter 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.

### C.2 SKELETON REPRESENTATION VISUALIZATION

We conduct the  $t$ -SNE visualization (Van der Maaten & Hinton, 2008) of skeleton representations learned from TranSG using Prompter on different datasets. The results in Fig. 8 show that the learned skeleton representations on the training sets can achieve similar inter-class separation on IAS and BIWI datasets and higher inter-class distance on KS20 and KGBD datasets compared with the original TranSG model. This suggests that applying the proposed Prompter may help capture more discriminative features for person re-ID. We also visualize the skeleton representations on different testing sets (see Fig. 9). It is observed that the representations of different classes on BIWI-W have more confusion (*i.e.*, vaguer class margins) than other testing sets, which is consistent with the performance results shown in our paper.

### C.3 CONFUSION MATRIX VISUALIZATION

As shown in Fig. 10, we visualize the confusion matrices of the TranSG model using Prompter 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. 10 (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. 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. 10 (c))

<sup>2</sup>According to the assumption and criterion in (Rao & Miao, 2022), good skeleton representations should satisfy: The same-class representations are gathered closer (higher  $mACT$ ) while different-class representations possess larger distances (higher  $mRCL$ ).

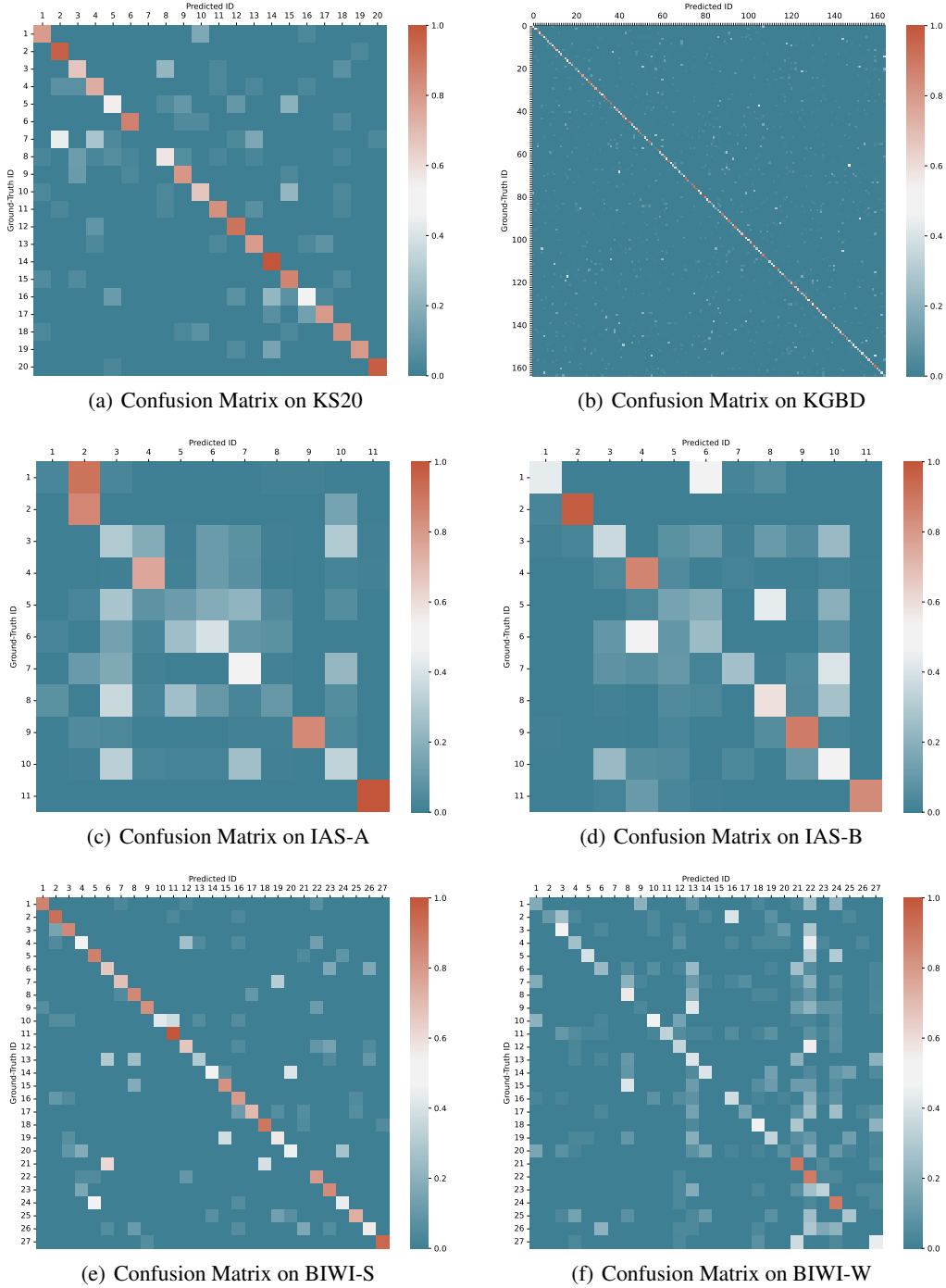


Figure 10: Visualization of confusion matrices on KS20 (a), KGBD (b), IAS-A (c), IAS-B (d), BIWI-S (e), and BIWI-W (f) when using the Rank-1 matching. Note that abscissa and ordinate denote the predicted and ground-truth identities, respectively. The position in the  $a^{th}$  row and  $b^{th}$  column 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 probe set.

and BIWI-Walking (see Fig. 10 (f)) imply that the TranSG model using Prompter tends to confuse skeleton sequences of more different identities on these datasets. These results are consistent with the performance results shown in the paper.

## D BROADER IMPACTS

Prompter could be applied to more tasks (*e.g.*, skeleton-based action recognition) and be potentially generalized to semantics learning of different fields (*e.g.*, masked context reconstruction of 3D point clouds). Moreover, as shown in (Rao & Miao, 2022) that *masked* contrastive representation learning could encourage learning more intra-sequence features and high-level motion semantics, the idea of probabilistic spatial-temporal context masking can be further applied into (*i.e.*, instead of combining) different downstream objectives (*e.g.*, GPC (Rao & Miao, 2023), SPC (Rao & Miao, 2022), MPC (Rao & Miao, 2022), PoseGait (Liao et al., 2020)) to sample more random subsequences to enhance skeleton representation learning. It can also serve as an effective representation-level augmentation strategy to combine with model-level augmentations such as Dropout algorithms (Baldi & Sadowski, 2014; 2013) to help reduce model over-fitting and improve their robustness against random perturbations. On the other hand, we hope the proposed first SSL generality assessment framework SUCT can inspire researchers to explore more useful SSL tasks and its broader application for different pattern recognition tasks including the emerging skeleton-based person re-ID and the aforementioned areas.

## E ADDITIONAL EXPERIMENTAL RESULTS

### E.1 SUPPLEMENTARY RESULTS FOR REVIEWER # IRXH

Please see Table 9 and Table 10: (1) We provide an additional comparison of key differences and similarity between our method (*i.e.*, skeleton-based person re-ID) and skeleton-based gait recognition methods; (2) We compare the performance of our method with different state-of-the-art gait recognition methods (SkeletonGait, GaitTR, GPGait) on all datasets.

Table 9: Comparison of key differences and similarity between our approach and gait recognition methods (*e.g.*, (Fan et al., 2024; Fu et al., 2023; Huang et al., 2023)).

Method	Prompter (Ours)	Skeleton-Based Gait Recognition Methods
<b>Task</b>	3D Skeleton Based Person Re-Identification	2D/3D Skeleton Based Gait Recognition
<b>Focused Problem</b>	Generally Matching and Retrieving; Classification	Generally Matching and Retrieving; Classification
<b>Input Skeleton Type</b>	Generally 3D Skeletons	Generally 2D Skeletons
<b>Application Scenarios</b>	Generally Sensor-based skeletons; Can be applied to Model-estimated skeletons (Explored)	Generally Model-estimated skeletons; Can be applied to Sensor-based skeletons (Unexplored)
<b>Base Architectures</b>	Can be flexibly applied to different models ( <i>e.g.</i> , Transformer, GAT, MLP)	Generally a specific model ( <i>e.g.</i> , GCN, CNN, Transformer)
<b>Datasets</b>	Generally 3D skeleton datasets (sensor-based); Datasets with estimated skeletons	22 non-skeleton datasets; 2 datasets with estimated skeletons
<b>Learning Scenarios</b>	Support supervised and unsupervised	Only Supervised
<b>Input Skeletal Topology</b>	Support different-level/type input skeleton data with varying nodes/topologies ( <i>e.g.</i> , 25, 20, or 14 joints, 10 (part-level) or 5 nodes (body-level))	Generally unified input skeleton data with same topology ( <i>e.g.</i> , COCO2017 format)

### E.2 SUPPLEMENTARY RESULTS FOR REVIEWER # DVW6

Please see Table 11, Fig. 11, Fig. 12, and Fig. 13: (1) We provide a performance comparison of different SSL tasks (DR, MIC, STPR, Prompter) under different skeleton levels (Joint-Level, Part-Level, Body-Level) on different datasets; (2) We offer qualitative examples and analyses for the cross-domain person re-ID performance, including confusion matrices and *t*-SNE feature visualization.

### E.3 SUPPLEMENTARY RESULTS FOR REVIEWER # V2ZJ

Table 10: Performance and efficiency comparison between our approach (best setting with TranSG), and SkeletonGait Fan et al. (2024) (skeletons used), GPGait Fu et al. (2023), and GaitTR Zhang et al. (2023) with optimal parameter settings. We adopt the same evaluation protocol (i.e., same datasets, skeletal topology, joint number, limb/bone partition, etc.) used in our work for a fair comparison. Considering that the inherent gaps (please refer to our responses) between these methods and our method might influence their performance on a different task to cause unfair comparison, the provided additional experimental results under default optimal parameter settings are only for a performance reference.

Method	# Paras	GFLOPs	KS20		BIWI-W		BIWI-S		IAS-A		IAS-B		KGBD	
			mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
SkeletonGait (Skeleton)	11.11M	1592.60	14.5	22.2	11.2	10.8	9.3	15.1	25.0	31.4	21.5	31.5	1.1	1.7
GaitTR	0.49M	18.20	25.5	52.4	19.8	21.9	18.8	46.3	31.1	44.2	34.2	49.4	13.5	51.6
GPGait	1.30M	49.62	41.1	71.4	25.5	29.0	23.5	54.1	<b>34.6</b>	<b>50.9</b>	43.1	60.1	17.8	53.9
Prompter (Ours)	0.41M	20.20	<b>48.3</b>	<b>74.2</b>	<b>27.3</b>	<b>34.6</b>	<b>30.3</b>	<b>66.8</b>	34.1	49.5	<b>43.8</b>	<b>60.4</b>	<b>21.3</b>	<b>59.5</b>

Table 11: Performance comparison of different SSL tasks using different skeleton levels. Each level is trained under the same base model TranSG, and we follow (Rao et al., 2021a; Rao & Miao, 2023) to construct part-level and body-level skeleton representations. The original skeletons, part-level skeletons, and body-level skeletons are shown in Fig. 2 and 3.

Method	BIWI-W			KS20		
	Joint-Level	Part-Level	Body-Level	Joint-Level	Part-Level	Body-Level
DR	33.8	17.6	13.2	73.2	48.4	39.3
MIC	34.5	19.1	12.2	72.3	48.4	40.8
STPR	32.7	20.0	17.0	73.6	48.1	40.8
Prompter (Ours)	34.6	20.1	16.3	74.2	49.4	41.9

Please see Table 12 and Sec. E.3.1: We provide an overview for state-of-the-art skeleton semantics learning (SSL) tasks and their source methods; We offer a detailed comparison between our masking method and existing state-of-the-art masking strategies.

Table 12: Overview of existing state-of-the-art skeleton semantics learning (SSL) tasks and their source methods, learning types.

ID	SSL Task	Source Method	Method Type
1	DR	—	—
2	AR	AGE (Rao et al., 2020)	Self-Supervised
3	AR + AC	SGELA (Rao et al., 2021b)	Self-Supervised
4	MSSP	MG-SCR (Rao et al., 2021c)	Supervised
5	MSR	SM-SGE (Rao et al., 2021a)	Supervised / Self-Supervised
6	MIC	SimMC (Rao & Miao, 2022)	Unsupervised
7	STPR	TranSG (Rao & Miao, 2023)	Supervised
8	Prompter (Ours)	—	—

### E.3.1 COMPARISON WITH EXISTING STATE-OF-THE-ART MASKING STRATEGIES

Firstly, compared with existing spatial-temporal masking strategies such as SkeletonMAE Wu et al. (2023) and MS2L Lin et al. (2020), we hope to clarify that the key novelty of the proposed Probabilistic Spatial Context Masking (PSCM) and Probabilistic Temporal Context Masking (PTCM) is that they are devised at an independent level of body structural locations and motion trajectory positions based on independent and identically distributed (IID) Bernoulli random masks. It possesses higher generality than previous methods (detailed in Sec. 3.2 of our paper) and can be probabilistically generalized to different existing masking mechanisms for more effective skeleton semantics learning (please see TT property and Line 367-377 of our paper). By contrast, existing masking strategies such as MS2L Lin et al. (2020) directly masks the later consecutive skeletons (*i.e.*, 150 frames) for temporal predictions while failing to learn effective spatial relations (*i.e.*, performance degrades) under the used spatial masking. In Wu et al. (2023), the structural positions are masked conditioned on the temporally-masked frames. Such direct frame-level or conditioned masking has several limitations, such as they cannot explicitly and individually model effective spatial semantics, nor can it feasibly evaluate the performance contribution of spatial masking (please refer to STE property defined in our

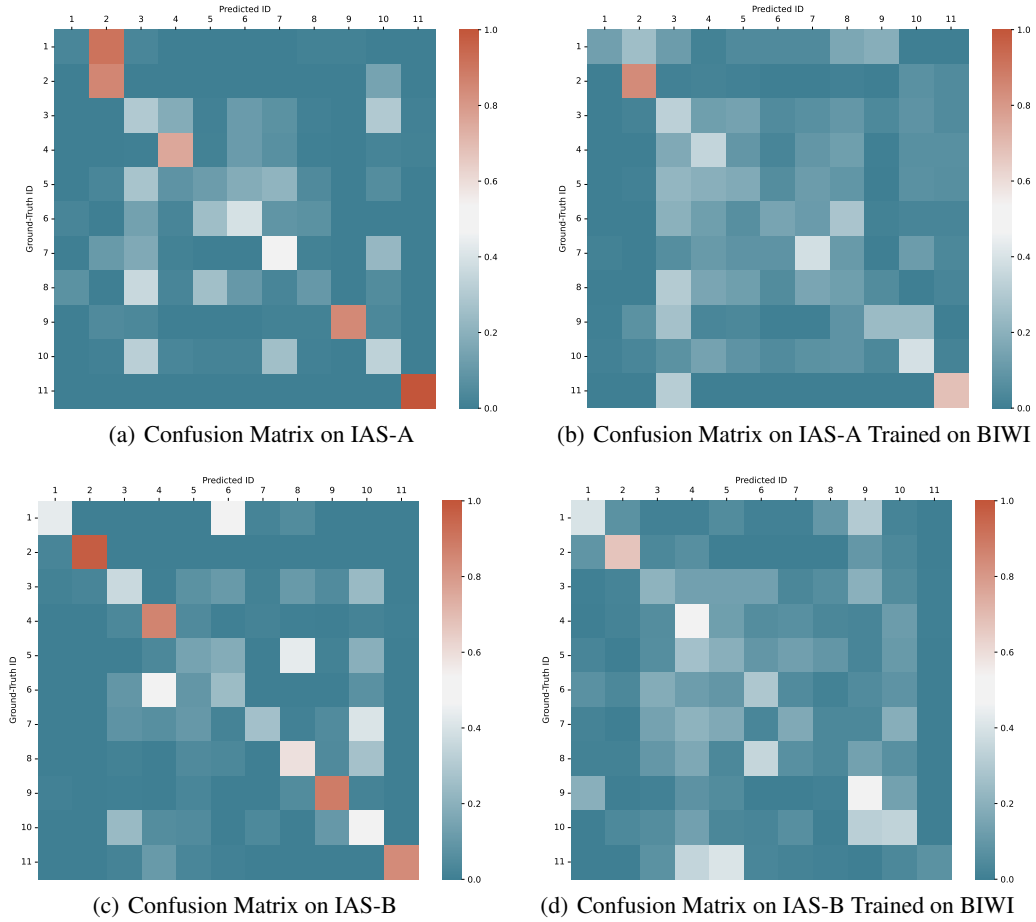


Figure 11: Visualization of confusion matrices on IAS-A and IAS-B when training on original dataset ((a) and (c)) and training on BIWI dataset ((b) and (d)) (i.e., generalized performance across datasets) using the Rank-1 matching. Note that abscissa and ordinate denote the predicted and ground-truth 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 probe set. Note that here ID =  $i$  corresponds class =  $i-1$  in Fig. 12 and Fig. 13.

work), while our method has solved these challenges with a focus of more generalizable skeleton context reconstruction. Secondly, in our experiments, we also systematically compare our method with state-of-the-art SSL tasks using different masking strategies: Direct temporal masking (MIC), random masking with fixed-number masks (STPR), and the baseline without masking (DR). The experimental results demonstrate the higher effectiveness of our approach that adopts independent and finer-grained spatial-temporal masking. Moreover, we hope to highlight another novel contribution of our work is to propose a systematic SSL generality assessment framework (SCUT) to explore the multi-faceted performance and bottlenecks of existing SSL tasks under varying models and scenarios (please see Line 58-76 of our paper). Motivated by the identified key properties of SCUT, we focus on devising a general solution (Prompter) that can be flexibly applied to different state-of-the-art skeleton-based person re-ID models (e.g., graph transformer, GAT, Siamese encoders). This is fundamentally different from MS2L Wu et al. (2023) and Lin et al. (2020) that rely on a certain action recognition backbone or model (i.e., GRU or STTFormer) to design effective masking strategies. Therefore, our method could be more general and scalable than these methods. Prompter can also be flexibly applied to RGB-estimated skeletons, unsupervised scenarios, different graph modeling, and cross-domain person re-ID tasks (see Further Analyses in our paper).



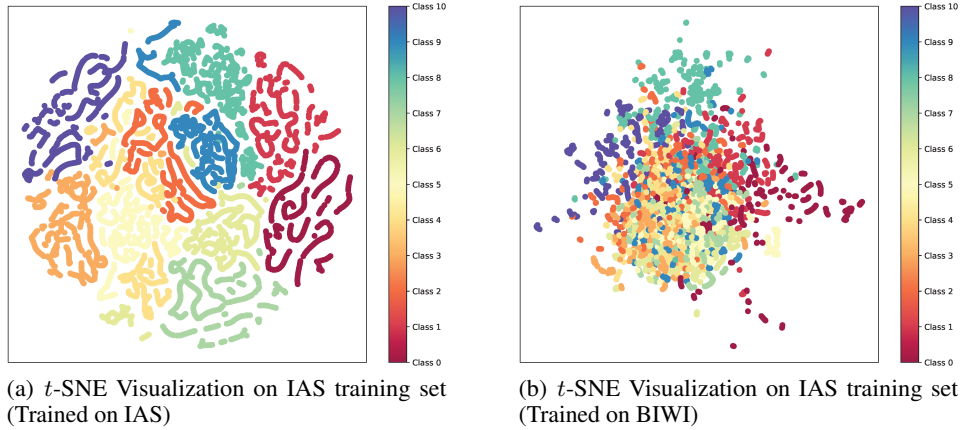


Figure 12:  $t$ -SNE feature visualization for for all classes of IAS training sets when training our method on the original IAS dataset (a) and BIWI dataset (b) (i.e., generalized performance across datasets). Different colors indicates representations of different classes. Note that here class =  $i$  corresponds ID =  $i+1$  in Fig. 11.

Apart from the key differences (i.e., independent masking of body structural locations and motion trajectory positions based on independent and identically distributed (IID) Bernoulli random masks) from existing masking mechanisms, the main novelty of Prompter include (1) *Explicit* effective temporal and spatial modeling (i.e., implement spatial-temporal effectiveness (STE)) in terms of body structure and motion trajectory (we have verified the effectiveness of each part in Ablation study), unlike previous SSL tasks used in skeleton-based person re-ID that do not distinguish these two parts for reconstruction (see Sec. 3.3 of our paper); (2) Fully exploit varying valuable *context* information (e.g., temporal context of trajectory) of *fine-grained* skeleton representations to capture richer skeleton semantics (please see Line 299-309 of our paper), which is achieved by combining multiple skeleton context based learning sub-objectives (i.e., establish Task Transformability (TT)), while existing SSL tasks typically utilize a fixed reconstruction objective (e.g., direct reconstruction or with fixed masks). Moreover, it is also a general SSL task that does not rely on any specific model architectures or feature representations, which is inspired and designed by the crucial properties/principles of SSL identified by SCUT (please see Line 302-309 of our paper). This also suggests the potential value of the proposed SCUT framework to devise more general SSL tasks for different scenarios. It can also be potentially modeled as a model regularization method like Dropout (please see theoretical assumptions and analyses in Appendix II).

#### E.4 SUPPLEMENTARY RESULTS FOR REVIEWER # DUN6

Please see Table 13: We provide an additional evaluation of the proposed Prompter on the representative state-of-the-art gait recognition method GPGait Fu et al. (2023), and the results demonstrate the effectiveness of Prompter (i.e., the proposed spatial-temporal skeleton semantics learning) to improve the performance of gait recognition method GPGait on all datasets.

Table 13: Performance (mAP) evaluation of our method when applied to the representative state-of-the-art gait recognition method GPGait (Fu et al., 2023) on different datasets. The **bold numbers** indicate higher performance than the base model *without* using SSL.

Method	KS20	BIWI-W	BIWI-S	IAS-A	IAS-B	KGBD
GPGait	41.1	25.5	23.5	34.6	43.1	17.8
+ Prompter (Ours)	<b>43.3</b>	<b>27.2</b>	<b>24.7</b>	<b>37.5</b>	<b>45.6</b>	<b>18.6</b>

#### E.5 SUPPLEMENTARY RESULTS FOR REVIEWER # BHKC

Please see Table 14: We provide an additional evaluation of the proposed Prompter when applied to the representative state-of-the-art skeleton-based person re-ID method TranSG (Rao & Miao,

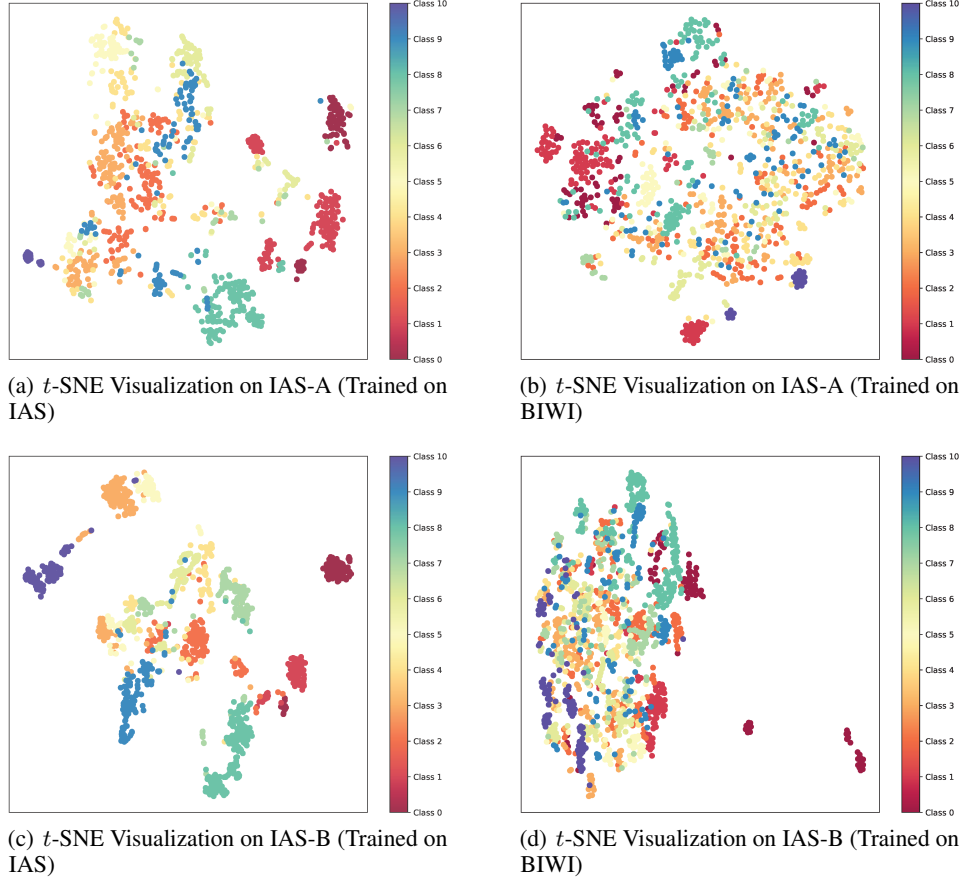


Figure 13:  $t$ -SNE feature visualization for for all classes of IAS-A and IAS-B testing sets (probe sets) when training our method on the original IAS dataset ((a) and (c)) and BIWI dataset ((b) and (d)) (i.e., generalized performance across datasets). Different colors indicates representations of different classes. Note that here class =  $i$  corresponds ID =  $i+1$  in Fig. 11.

2023), gait recognition method GPGait (Fu et al., 2023), and action recognition method ST-GCN (Yan et al., 2018). The results demonstrate the effectiveness and generality of Prompter (i.e., the proposed spatial-temporal skeleton semantics learning) when applied to different architectures from varying research communities (e.g., gait recognition method and action recognition) to improve their performance (Rank-1 accuracy) in most cases under the same evaluation setting.

Table 14: Performance (Rank-1 accuracy) evaluation of our method when applied to the representative state-of-the-art skeleton-based person re-ID method TranSG (Rao & Miao, 2023), gait recognition method GPGait (Fu et al., 2023) and action recognition method ST-GCN (Yan et al., 2018) on different datasets. The **bold numbers** indicate higher performance than the base model *without* using SSL.

Method Source (Research Community)	Method	KS20	IAS-A	IAS-B	KGBD
Person Re-Identification	TranSG	71.3	48.0	56.1	57.0
	+ Prompter (Ours)	<b>74.2</b>	<b>49.5</b>	<b>60.4</b>	<b>59.5</b>
Gait Recognition	GPGait	71.4	50.9	60.1	53.6
	+ Prompter (Ours)	<b>72.7</b>	<b>55.3</b>	<b>61.7</b>	53.4
Action Recognition	ST-GCN	60.4	43.6	49.1	57.7
	+ Prompter (Ours)	<b>65.6</b>	<b>53.4</b>	<b>58.8</b>	<b>59.0</b>

## REFERENCES

- Pierre Baldi and Peter Sadowski. The dropout learning algorithm. *Artificial intelligence*, 210:78–122, 2014.
- Pierre Baldi and Peter J Sadowski. Understanding dropout. *Advances in Neural Information Processing Systems (NeurIPS)*, 26, 2013.
- Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2019.
- Ching-Hang Chen and Deva Ramanan. 3D human pose estimation= 2D pose estimation+ matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7035–7043, 2017.
- Chao Fan, Jingzhe Ma, Dongyang Jin, Chuanfu Shen, and Shiqi Yu. Skeletongait: Gait recognition using skeleton maps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1662–1669, 2024.
- Yang Fu, Shibe Meng, Saihui Hou, Xuecai Hu, and Yongzhen Huang. Gpgait: Generalized pose-based gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19595–19604, 2023.
- Xiaohu Huang, Xinggang Wang, Zhidianqiu Jin, Bo Yang, Botao He, Bin Feng, and Wenyu Liu. Condition-adaptive graph convolution learning for skeleton-based gait recognition. *IEEE Transactions on Image Processing*, 2023.
- Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020.
- Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2490–2498, 2020.
- Zheng Liu, Zhaoxiang Zhang, Qiang Wu, and Yunhong Wang. Enhancing person re-identification by integrating gait biometric. *Neurocomputing*, 168:1144–1156, 2015.
- Lutz Prechelt. Early stopping-but when? In *Advances in Neural Information Processing Systems (NeurIPS) Workshop*, pp. 55–69, 1998.
- Haocong Rao and Chunyan 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.
- Haocong Rao and Chunyan Miao. Skeleton prototype contrastive learning with multi-level graph relation modeling for unsupervised person re-identification. *arXiv preprint arXiv:2208.11814*, 2022.
- Haocong Rao and Chunyan Miao. TranSG: Transformer-based skeleton graph prototype contrastive learning with structure-trajectory prompted reconstruction for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- Haocong Rao, Siqi Wang, Xiping Hu, Mingkui Tan, Huang Da, Jun Cheng, and Bin Hu. Self-supervised gait encoding with locality-aware attention for person re-identification. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 1, pp. 898–905, 2020.
- Haocong Rao, Xiping Hu, Jun Cheng, and Bin 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, 2021a.

- Haocong Rao, Siqi Wang, Xiping Hu, Mingkui Tan, Yi Guo, Jun Cheng, Xinwang Liu, and Bin 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*, 44(10): 6649–6666, 2021b.
- Haocong Rao, Shihao Xu, Xiping Hu, Jun Cheng, and Bin 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, 2021c.
- Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark J Finocchio, Richard Moore, Alex Abenathar Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1297–1304, 2011.
- Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.
- Wenhan Wu, Yilei Hua, Ce Zheng, Shiqian Wu, Chen Chen, and Aidong Lu. Skeletonmae: Spatial-temporal masked autoencoders for self-supervised skeleton action recognition. In *2023 IEEE international conference on multimedia and expo workshops (ICMEW)*, pp. 224–229. IEEE, 2023.
- 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 (AAAI)*, pp. 7444–7452, 2018.
- Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *International Conference on Pattern Recognition (ICPR)*, volume 4, pp. 441–444. IEEE, 2006.
- Cun Zhang, Xing-Peng Chen, Guo-Qiang Han, and Xiang-Jie Liu. Spatial transformer network on skeleton-based gait recognition. *Expert Systems*, 40(6):e13244, 2023.
- Rui Zhao, Kang Wang, Hui Su, and Qiang Ji. Bayesian graph convolution lstm for skeleton based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6882–6892, 2019.
- Liang Zheng, Liye Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1116–1124, 2015.
- Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1318–1327, 2017.