Camera-LiDAR Cross-modality Gait Recognition (Supplementary Material)

Wenxuan Guo[1*], Yingping Liang[2*], Zhiyu Pan[1], Ziheng Xi[1], Jianjiang Feng[1†], and Jie Zhou[1]

[1] Department of Automation, BNRist, Tsinghua University
[2] Beijing Institute of Technology
{gwx22, pzy20, xizh21}@mails.tsinghua.edu.cn liangyingping@bit.edu.cn
{jfeng, jzhou}@tsinghua.edu.cn

We provide more details in this supplementary material, including: 1) Experiments on the impact of modality on gait recognition. 2) Feature visualization about our contrastive silhouette-point pre-training strategy (CSPP). 3) Examples of generated multimodal gait data for contrastive pre-training.

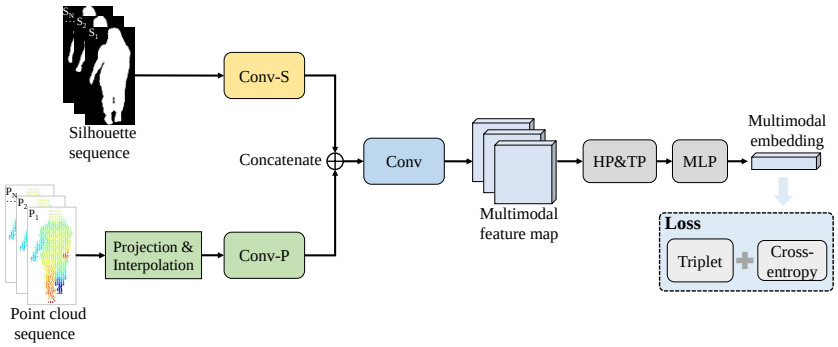


Fig. 1: The overview of CL-Gait-F. We modify the cross-modality network of CL-Gait to obtain CL-Gait-F for camera-LiDAR multi-modality gait recognition. The input to CL-Gait-F consists of synchronized sequences of silhouettes and point clouds.

A Impact of Modality

To investigate the impact of different modalities on gait recognition tasks, we compare the performance of single-modality, multi-modality, and cross-modality gait recognition approaches. For single-modality methods, we compare with the state-of-the-art methods on the SUSTech1K dataset [9]. Because there are no existing camera-LiDAR multi-modality methods for gait recognition, we modify the cross-modality network of CL-Gait to obtain CL-Gait-F, as illustrated in Fig. 1. The input to CL-Gait-F consists of synchronized sequences of silhouettes

* Equal contribution. † Corresponding author.

from camera and point clouds from LiDAR. The quantitative results are shown in Tab. 1, from which the following observations can be obtained: 1) The multi-modality method, CL-Gait-F, surpasses all single-modality methods, indicating the complementarity of the two modalities and also proving that the network of CL-Gait can effectively extract distinguishable features from both modalities. 2) Cross-modality gait recognition performs worse than other methods because it needs to handle the significant modality discrepancy between different modalities. It is a valuable but challenging task that still requires further research.

Table 1: Performance of state-of-the-art methods under different modalities on the SUSTech1K dataset. '-' indicates that the result is not reported in the paper.

Method	Modality	Rank-1	Rank-3	Rank-5
GaitSet [1]	Camera	65.04	-	84.76
GaitPart [3]		59.19	-	80.79
GaitGL [5]		63.14	-	82.82
GaitBase [2]		75.98	86.22	89.59
PointNet [7]	LiDAR	31.33	-	59.75
PointNet++ [8]		50.78	-	82.38
PointTransformer [10]		44.37	-	76.70
SimpleView [4]		64.83	-	85.77
LidarGait [9]		86.66	94.10	95.92
CL-Gait-F (ours)	Camera and LiDAR	**90.06**	**95.97**	**97.31**
CL-Gait (ours)	LiDAR to Camera	53.29	69.54	75.59
	Camera to LiDAR	55.12	71.23	77.31

B Feature Visualization

To visually demonstrate the effectiveness of our proposed contrastive silhouette-point pre-training strategy (CSPP), we use t-SNE to visualize the feature distributions of the first 100 individuals in the SUSTech1K test set extracted by CL-Gait with and without CSPP, as shown in Fig. 2 and Fig. 3. We refer to CL-Gait without pre-training as CL-Gait-B. We can observe that the feature distributions extracted by CL-Gait and CL-Gait-B share certain similarities. However, for cross-modality retrieval tasks, CL-Gait demonstrates better discriminative ability. In Fig. 2, features of the same modality for some individuals cluster together and are distant from features of another modality, as emphasized by the colored elliptical circles. This is primarily due to the significant modality discrepancy between 2D silhouettes and 3D point clouds. Conversely, in Fig. 3, the clustering phenomenon within the same modality is significantly reduced, making cross-modality retrieval more accurate, as indicated by the gray boxes. We attribute this to the potent influence of our proposed CSPP, which effectively mitigates modality discrepancy.

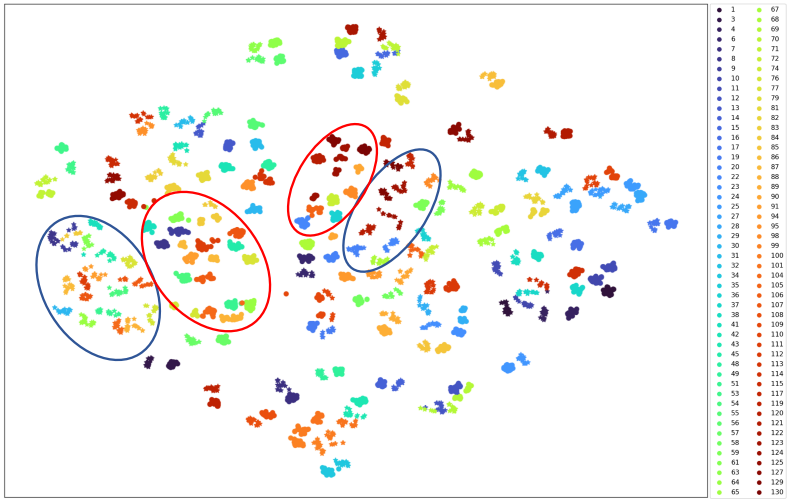


Fig. 2: The feature distribution of the first 100 individuals in the SUSTech1K test set extracted by CL-Gait without pre-training (CL-Gait-B). Stars and points respectively represent the features of silhouette sequences and point cloud sequences, and distinct colors indicate different individuals. Features of the same modality for some individuals cluster together and are distant from features of another modality, as indicated by the elliptical circles. This is primarily due to the significant modality discrepancy between 2D silhouettes and 3D point clouds.

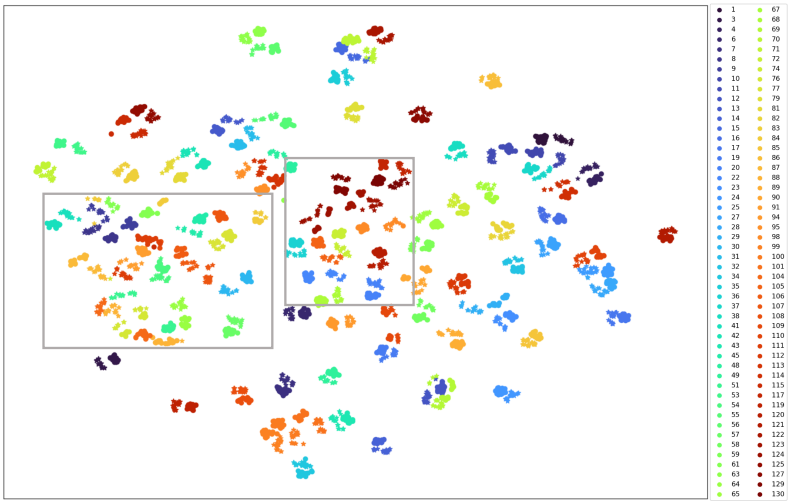


Fig. 3: The feature distribution of the first 100 individuals in the SUSTech1K test set extracted by CL-Gait with contrastive pre-training. The individuals within gray boxes correspond to these within the elliptical circles in Fig. 2. The clustering phenomenon within the same modality has been greatly reduced, making cross-modality retrieval more accurate. This can be attributed to the effectiveness of contrastive pre-training in mitigating modality discrepancy.

C Generation of Pre-training Gait Data

C.1 Comparison with Real Gait Data

To demonstrate the effectiveness and realism of our proposed method of multi-modal gait data generation, we present several generated examples and compare them with the real data, as shown in Fig. 4. The examples are from SUSTech1K dataset, because it includes RGB images, real point clouds and corresponding depth images, for the comparison with our generated data. From Fig. 4, we can observe that the depths estimated by our method are realistic. The generated point clouds and the depth maps obtained from point clouds are very close to the real data. Furthermore, as shown in Fig. 4b, even in low-light environments, our method can still obtain accurate depth estimation and generate gait data that closely matches real-world conditions. The realism of the generated point clouds and depth maps, along with the consistency of the paired gait data, ensures the effective implementation of contrastive pre-training to mitigate modality discrepancy in cross-modality gait recognition tasks.

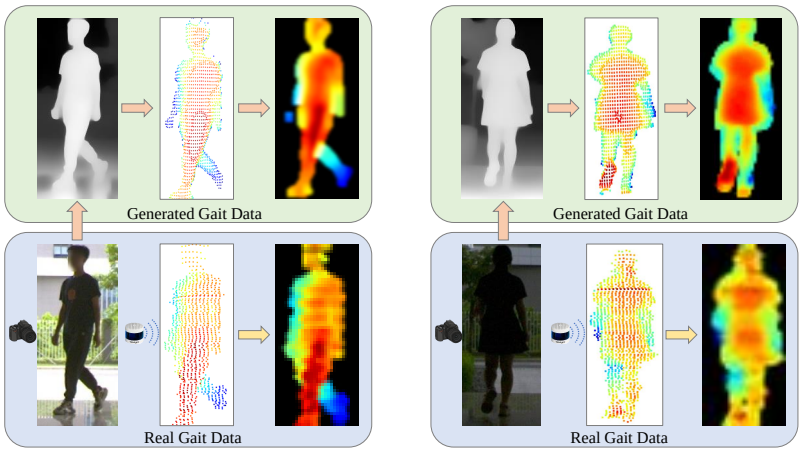
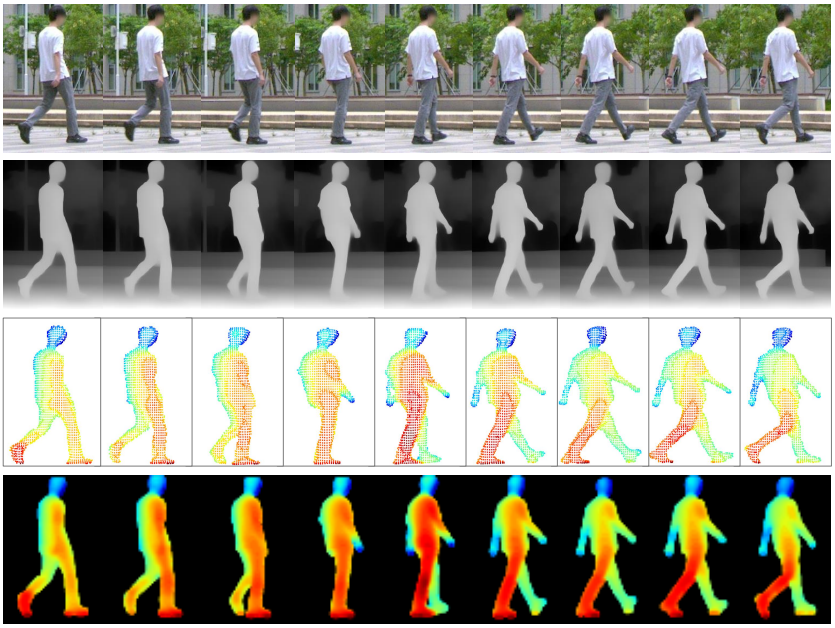


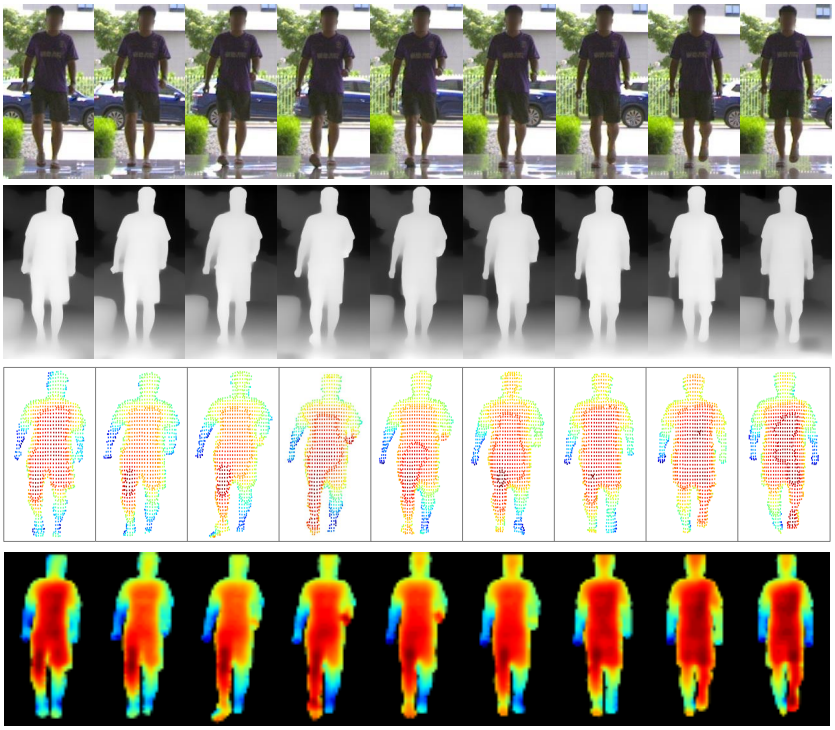
Fig. 4: Examples of generated multimodal gait data and corresponding real data. In each example, the green box contains the estimated and generated gait data, and the blue box contains the real data. Real data acquisition costs are high. In contrast, our synthetic data can be easily accessed at scale.

C.2 Visualization of Generated Samples

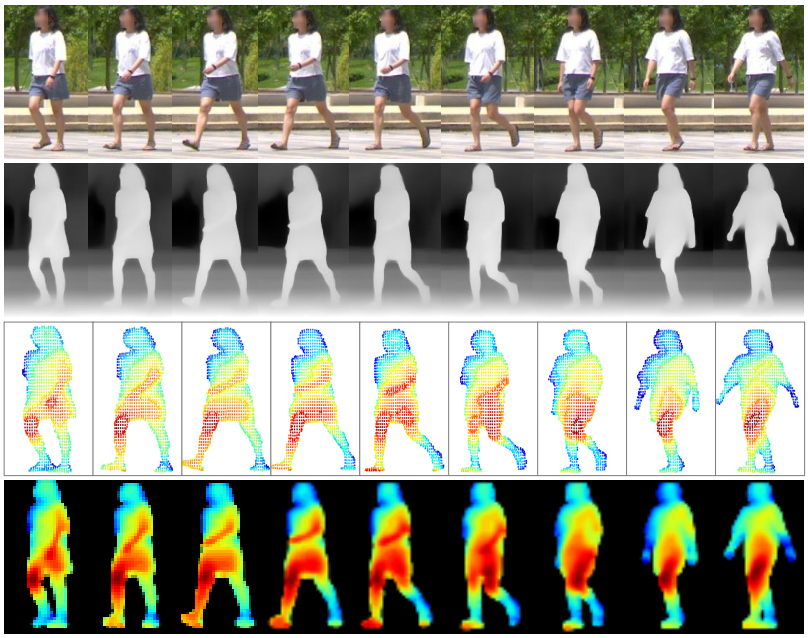
Our CL-Gait can automatically generate training pairs from RGB images or videos without human involvement. Thus, huge amount of RGB images can be used to generate training pairs for contrastive pre-training, which helps the supervised networks generalize to various scenes. Figure 6 shows some generated samples from real-world video sequences that cover individuals from different datasets, scenarios, and views.



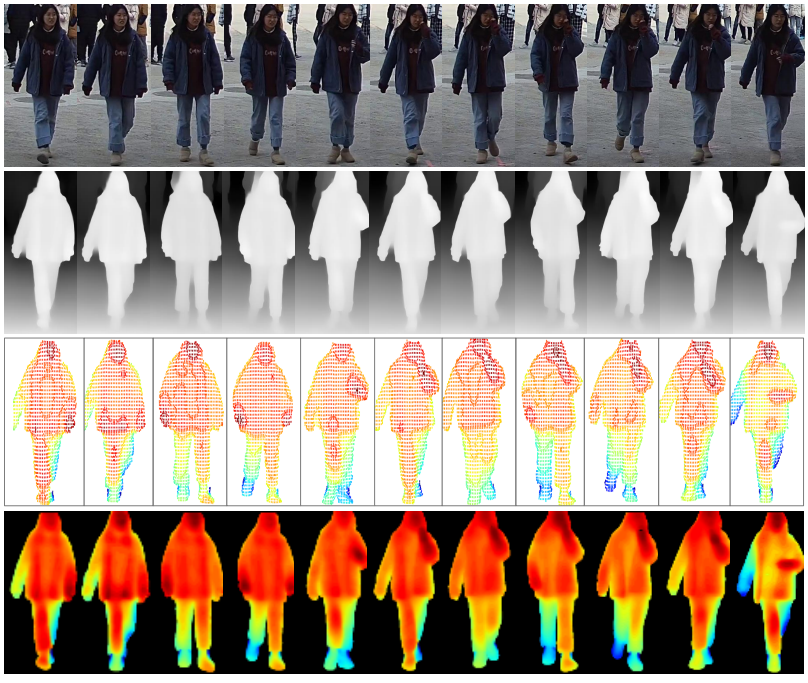
(a) SUSTech1K [9], scene 1, view 225°.



(b) SUSTech1K [9], scene 2, view 0°.



(c) SUSTech1K [9], scene 3, view 45°.



(d) HITSZ-VCM [6].

Fig. 6: Examples of our multimodal gait data generation method. The examples are sequence data, covering individuals from different datasets, scenarios, and views.

References

1. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8126–8133 (2019)
2. Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., Yu, S.: Opengait: Revisiting gait recognition towards better practicality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9707–9716 (2023)
3. Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., He, Z.: Gaitpart: Temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14225–14233 (2020)
4. Goyal, A., Law, H., Liu, B., Newell, A., Deng, J.: Revisiting point cloud shape classification with a simple and effective baseline. In: International Conference on Machine Learning. pp. 3809–3820. PMLR (2021)
5. Lin, B., Zhang, S., Wang, M., Li, L., Yu, X.: Gaitgl: Learning discriminative global-local feature representations for gait recognition. arXiv preprint arXiv:2208.01380 (2022)
6. Ling, Y., Zhong, Z., Luo, Z., Rota, P., Li, S., Sebe, N.: Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 889–897 (2020)
7. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660 (2017)
8. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems **30** (2017)
9. Shen, C., Fan, C., Wu, W., Wang, R., Huang, G.Q., Yu, S.: Lidargait: Benchmarking 3D gait recognition with point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1054–1063 (2023)
10. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16259–16268 (2021)