GAITPOINT: A GAIT RECOGNITION NETWORK BASED ON POINT CLOUD ANALYSIS

Jiajing Chen, Huantao Ren, Frank (Sicong) Chen, Senem Velipasalar, Vir V. Phoha*

Dept. of Electrical Engineering and Computer Science
Syracuse University, Syracuse, NY, USA
{jchen152, hren11, schen154, svelipas, vvphoha}@syr.edu

2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897558

ABSTRACT

We propose a novel gait recognition method that combines convolutional features with features of human pose key points obtained by a point cloud analysis model. Currently, most state-of-the-art works on gait recognition rely on only images and are purely based on convolutional neural networks. Most of these methods are very sensitive to small variations in the appearance of a walking person. For instance, if a person wears a coat or carries a bag, the accuracy of these methods may drop significantly. To address this problem, we propose to treat a sequence of human key points as a point cloud and combine human key point features and convolution feature map for final prediction. The experimental results show the promise of this approach, which outperforms three state-of-the-art baselines in all walking scenarios, including the ones involving heavy clothing or carried items.

Index Terms— Gait recognition, Point cloud, Human key points, Convolution feature map

1. INTRODUCTION

Gait is a type of behavioral biometric trait that describes the way people walk. Compared to other biometrics, such as iris, face and fingerprint, gait analysis allows identifying a person from a far distance without direct contact. Thus, gait recognition can be applied in different areas, including security, crime prevention, and video surveillance.

Performance of gait recognition is usually affected by different factors, including variations in appearance of individuals. Wearing outerwear, such as a coat, or carrying a bag can change the appearance of individuals. Different camera viewpoints [23] and occlusions by an object or by a part of an individual's own body in certain camera views [21] can also lead to appearance variations.

To address these issues, many deep learning-based methods have been developed providing state-of-the-art (SOTA) performance. Most current approaches use silhouettes [17, 3, 5, 9, 1, 26] to represent human body. Silhouettes can be extracted by using background subtraction followed by binarization. A sequence of silhouettes can represent useful gait features, such as gait cycle time, leg length, speed, and stride length. Yet, silhouettes are sensitive to appearance changes, such as different clothing, body type, and hairstyle. Hence, skeleton-based methods [20, 12] have been proposed to address this problem. Skeletons can be obtained by high-accuracy depth sensors or pose estimation algorithms. Although skeletons can be extracted robustly, body shape and an individual's appearance still provide useful information for gait recognition, and thus, they should not be entirely omitted. Silhouette and skeleton data represent different but useful and complementary gait information.

Inspired by the success of PointNet [16] in 3D point cloud processing, we propose GaitPoint, a novel and general approach, wherein we adopt PointNet as an auxiliary module for feature extraction from skeleton key points and then combine silhouette and skeleton features to provide a more exhaustive and richer feature set for gait recognition. The main contributions of this work include the following:

- We show that skeleton key points can be regarded as a 3D point cloud and use a point cloud processing approach, namely PointNet, to extract features from these key points.
- We provide extensive experimental results and comparisons with three SOTA baselines on the CASIA-B dataset [25], and show that our approach consistently improves the gait recognition performance of the purely image- and convolution-based approaches by treating skeleton points as 3D point clouds, and utilizing the point cloud analysis model as an auxiliary module.

2. RELATED WORK

Gait Recognition: Cameras can capture gait videos from different angles. However, different camera viewpoints can greatly affect the accuracy of gait recognition. Several studies [4, 11, 6] normalized the gait features from different view points into a specific view angle to address this problem. Other works [7, 24] built a View Transformation Model (VTM) to solve this multi-view problem. In recent years, inspired by the success of Convolutional Neural Networks (CNNs) in different applications, researchers used convolutional layers and pooling layers to extract the most distinguishing features from videos for gait recognition [3, 1, 10].

*This work was funded in part by National Science Foundation under Grant 1816732 and by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0000940.

Point Cloud Analysis: 3D point cloud data is obtained by sensors, such as a depth camera and Lidar. Unlike 2D images, 3D point cloud data is unstructured. Thus, traditional CNN-based methods cannot be readily applied to point clouds to extract features. PointNet [16] is a pioneer work performing deep learning-based point cloud analysis by using a max-pooling layer to obtain permutation invariant features for the downstream tasks, e.g., classification and segmentation. PointNet++ [15], developed based on PointNet, uses Farthest Point Sampling to obtain a set of points in different resolutions for final prediction. DGCNN [22] uses a dynamic graph CNN to aggregate each point's non-local neighbor features in each layer. Instead of aggregating k-nearest neighbors' point features, other methods [13, 8, 27] used 3D convolution to extract features for classification or segmentation.

3. PROPOSED METHOD

We propose a novel, one-stage gait recognition model, which combines convolutional features with features from human pose key points obtained from a skeleton. We provide the motivation for our work below before describing the details.

3.1. Motivation

In vision-based gait recognition, an individual's (subject's) gait video V (the input) is matched to an existing/known individual's gait video $V^{'}$ in a database to recognize the person. However, the appearance of the unknown subject might be different from the videos in the database, e.g. the person can be wearing a coat or carrying an item in the input video. This presents a challenge to the existing image-based gait recognition approaches. Although CNNs have been shown to be effective in extracting features from images and videos, they are prone to overfitting to a person's contour/shape rather than focusing on the actual walking motion. Thus, changes in a subject's contour, e.g. due to outfit or carried items, greatly affect the performance of the CNN-based approaches.

Table 1 shows performances of several SOTA image- and CNN-based gait recognition algorithms for different appearances walking statuses. These results have been obtained by running the authors' codes on a local machine with only one GPU. Thus, the results can be slightly different from the ones reported in the original papers. In Table 1, 'Normal' refers to walking without carrying anything and without wearing heavier outfits. It can be seen that, for all the models, the highest accuracy is achieved for 'Normal' walking. If a person carries a bag, the models' performances degrade considerably, and all the models have the worst performance when a person wears a coat. To further analyze the reasons behind this phenomenon, we visualize the same individual's walking silhouettes from different view angles and with varying appearances in Fig. 1. As can be seen, when carrying a backpack, the contour of the individual is similar to that with normal walking, especially for front and back views (0° and 180°). However, when wearing a coat, the individual's upper body looks larger, regardless of the viewing angle. This can explain why the models in Table 1 have a better performance for carrying a bag than wearing a coat. Purely image- and convolution-based models rely highly on an individual's contour features, causing them not to learn the walking features well. For this reason, changes in the appearances of individuals can have significant negative effects on recognition accuracy.

	Normal	Carrying Bag	Wearing Coat
GaitSet [1]	93.87%	87.15%	72.75%
GaitPart [3]	94.48%	89.29%	74.71%
GaitGL [10]	95.39%	91.94%	79.54%

Table 1. Performance of SOTA CNN-based gait recognition methods for different walking scenarios with varying appearances.

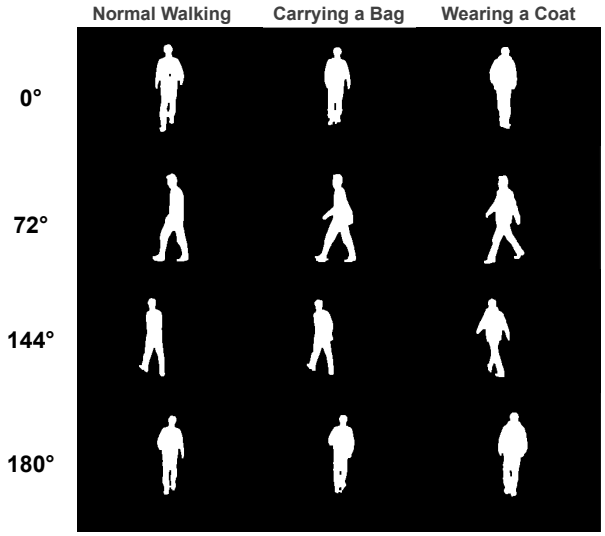

Fig. 1. An individual's silhouette from different angles when walking normal, carrying a bag and wearing a coat.

Using skeleton points or incorporating them into gait recognition can remedy this problem. GaitGraph [20] is a skeleton-based gait recognition method, which uses HRNet [19] to obtain the human pose key points in each frame. Then, it employs ResGCN [18] to extract human motion features based on those key points. However, since most human contour information is filtered out during pose key point detection, the performance of GaitGraph was lower than most image- and convolution-based methods [1, 3, 10, 5]. Developed based on ResGCN, MSGG [14] extracts human motion information from human pose key points obtained by HRNet. Along with the human's contour features obtained by GaitPart [3], MSGG achieved better performance than the original GaitPart. However, the training of MSGG [14] is computationally expensive. The parts for learning human contour features and skeleton key point features are pre-trained separately and then combined for final fine-tuning. In addition, the complexity of the key point processing method in [14] may limit its generalizability and widespread use.

To address all the aforementioned issues, our goals are

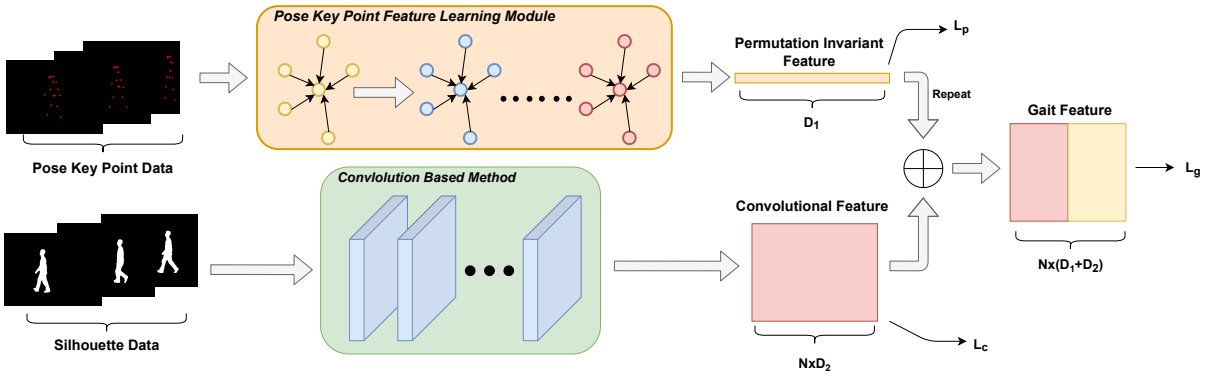

Fig. 2. The network structure of the proposed GaitPoint. L_p, L_c and L_g are the triplet losses for the permutation invariant, convolutional and gait features, respectively.

to design a one-stage, end-to-end gait recognition method, and to utilize human pose key points as auxiliary information by combining them with convolutional features. Since our proposed gait recognition model is one-stage, its human pose key point learning module should satisfy the following conditions: (i) The key point learning module should be compatible with the learning strategy of the most image- and convolution-based methods. For instance, in GaitGraph, ResGCN is trained with contrastive learning [2], followed by fine-tuning. This training strategy is very different from the ones used with convolution-based methods, wherein, walking feature distance is minimized for the same person, while it is maximized for different people; (ii) The key point learning module should be lightweight. Since the image- and convolution-based methods already require considerable amount of computational resources, it is preferable that the key point learning module is faster and has lighter weight.

3.2. Proposed GaitPoint

To address the issues of purely image- and convolution-based methods, and to consider other features that are less sensitive to appearance changes, we propose a new gate recognition network, referred to as the GaitPoint, which incorporates features obtained from skeleton key points into the gait recognition pipeline. The structure of GaitPoint is shown in Fig. 2. For the lower branch, the pipeline is similar to the image- and convolution-based gait recognition methods [3, 1, 10]. Taking the silhouette data as input, the model learns the most distinguishing features. It performs the pooling operation on feature maps multiple times, to obtain features in different receptive fields, and outputs the convolutional features $C \in \mathbf{R}^{N \times D_2}$, where N is the number of output features, and D_2 is the dimension of each feature. For the upper branch, the input is a sequence of 2D pose key point data, which are obtained by HRNet [19] from each frame. We first convert this key point data into a 3D point cloud as follows:

$$\{(x_{it}, y_{it}, t) \mid i \in \{1, ..., K\}, t \in \{0, .., T-1\}\}, \quad (1)$$

where K refers to the number of key points in each frame and T is the number of frames. The Pose Key Point Feature Learning Module, shown in Fig. 2, can be any point cloud analysis method in general. In our work, we employ PointNet, for point cloud analysis, for multiple reasons. In order to analyze which model is a better fit as the human key point processing module, we compared ResGCN [18] with PointNet [16] in terms of model size, speed, and accuracy. Both networks are trained and evaluated in the same way as the CNN-based methods [3, 1, 10] on the CASIA-B dataset [25]. The results, summarized in Table 2, show that PointNet has smaller memory requirement, is faster, and provides better accuracy at the same time compared to ResGCN. In addition, PointNet is more robust to appearance variations having a lower standard deviation in accuracy for three different walking scenarios. Thus, in our method, we employ PointNet for pose key point learning. Permutation invariant features $F \in \mathbf{R}^{D_1}$, obtained from PointNet, contain motion/walking information that is more robust to appearance changes. The vector F is replicated N times, to obtain an $N \times D_1$ vector, which is then concatenated with convolutional features C to obtain the overall gait features $G \in \mathbf{R}^{N \times (D_1 + D_2)}$. We employ the loss function $L = L_p + L_c + L_g$, where L_p, L_c and L_g are the triplet losses for F, C, and G, respectively. The triplet loss $L_{tp} = max(D(A,P) - D(A,N) + M, 0)$, where $D(A,P)$ and $D(A,N)$ are distances of the anchor to a positive sample and a negative sample, respectively. M is the margin value set by user, which controls the trade-off between faster convergence of training and ease of separating positive and negative samples . With the loss L, the distance between the features from the videos of the same person is decreased, while the distance between features from the videos of different people is increased.

4. EXPERIMENTAL RESULTS

For evaluation, we employ the CASIA-B dataset [25], which is a commonly-used dataset for gait recognition. CASIA-B dataset contains videos of 124 people. Each person has multiple videos while walking in normal case, carrying a bag, or wearing a coat. For different walking scenarios, the view angle ranges from 0° to 180°, with 18° increments. We use the

Model	Model Attribute			Model Accuracy			
	Forw./backw. pass size (MB)	Estim. Total Size(MB)	Running Time/ Batch (ms)	Normal Walking	Carrying Bag	Wearing Coat	SD
PointNet [16]	**2541.75**	**2546.52**	**19.07**	**64.28%**	**56.18%**	**51.92%**	**5.13%**
ResGCN [18]	3604.76	3607.71	34.11	60.85%	51.07%	40.95%	8.12%

Table 2. Comparison of ResGCN and PointNet in terms of size, running speed and accuracy. SD is the standard deviation of the accuracy for different walking scenarios.

		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	N	Mean
NM	GaitSet[1]	89.80%	97.80%	98.30%	95.30%	91.00%	90.90%	93.60%	97.30%	97.00%	96.10%	85.50%	/	93.87%
	GaitSet+PointNet	**94.04%**	96.87%	96.80%	**96.06%**	**95.41%**	**95.00%**	**94.54%**	96.20%	96.70%	**96.60%**	**89.80%**	7	95.27% (↑1.4%)
	Gaitpart[3]	90.30%	97.40%	98.40%	97.10%	93.70%	91.00%	93.30%	97.40%	98.20%	95.40%	87.10%	/	94.48%
	GaitPart+PointNet	**95.76%**	95.05%	95.60%	95.46%	**95.71%**	**95.65%**	**95.46%**	96.70%	96.60%	**96.50%**	**91.90%**	6	95.49% (↑1.01%)
	GaitGL[10]	93.80%	96.80%	98.50%	97.20%	94.90%	90.50%	94.60%	97.80%	98.20%	96.80%	90.20%	/	95.39%
	GaitGL+PointNet	**96.16%**	96.06%	95.30%	95.66%	**95.00%**	**97.18%**	**96.70%**	97.60%	95.00%	96.70%	**93.00%**	5	95.85% (↑0.46%)
BG	GaitSet[1]	84.30%	90.51%	93.94%	90.61%	83.80%	79.00%	83.90%	90.50%	92.10%	90.41%	79.60%	/	87.15%
	GaitSet+PointNet	**89.70%**	**92.04%**	92.45%	**92.19%**	**90.52%**	**88.82%**	**89.19%**	**93.30%**	**95.60%**	**93.94%**	**86.60%**	10	91.30% (↑4.15%)
	Gaitpart[3]	84.60%	93.54%	95.25%	93.88%	88.60%	82.80%	87.80%	92.40%	93.30%	90.30%	79.70%	/	89.29%
	GaitPart+PointNet	**91.62%**	90.10%	92.25%	92.50%	**90.93%**	**88.82%**	**90.92%**	92.30%	**94.10%**	**93.84%**	**87.10%**	7	91.32% (↑2.03%)
	GaitGL[10]	91.30%	94.24%	96.26%	94.80%	90.50%	84.70%	87.90%	94.30%	96.10%	94.55%	86.70%	/	91.94%
	GaitGL+PointNet	**93.84%**	93.78%	93.16%	94.48%	**94.43%**	**92.90%**	**92.45%**	**94.60%**	93.30%	**95.86%**	**92.00%**	7	93.71% (↑1.77%)
CT	GaitSet[1]	64.00%	77.70%	78.60%	76.90%	72.90%	71.60%	72.70%	73.80%	77.80%	75.00%	59.20%	/	72.75%
	GaitSet+PointNet	**79.90%**	**85.70%**	**83.70%**	**82.14%**	**81.04%**	**78.75%**	**80.83%**	**82.80%**	**83.80%**	**84.10%**	**76.40%**	11	81.74% (↑8.99%)
	Gaitpart[3]	68.00%	81.50%	83.50%	79.30%	74.00%	70.30%	74.40%	78.40%	78.00%	74.40%	60.00%	/	74.71%
	GaitPart+PointNet	**82.00%**	**84.70%**	**86.40%**	**84.90%**	**83.33%**	**81.71%**	**83.75%**	**85.70%**	**84.60%**	**84.90%**	**81.30%**	11	83.94% (↑9.23%)
	GaitGL[10]	71.40%	86.70%	89.50%	85.60%	78.60%	70.40%	77.90%	83.50%	85.60%	79.00%	66.70%	/	79.54%
	GaitGL+PointNet	**85.20%**	**87.60%**	87.50%	**90.00%**	**85.31%**	**84.09%**	**87.40%**	**89.20%**	**89.60%**	**87.60%**	**81.10%**	10	86.78% (↑7.24%)

Table 3. Results on the CASIA-B dataset. For all benchmarks, our method increases the mean accuracy for all walking scenarios (NM: normal walking, BG: carrying a bag, CT: wearing a coat). N is the number of viewing angles, for which the accuracy is improved.

videos of the first 70 people and videos of the remaining 54 people for training and testing, respectively. At each training step, 4 people (out of 70) are randomly chosen, and 8 videos are randomly picked from each person's video pool, resulting in a total of 32 videos for each step. By using the triplet loss, the distance between the features from the videos of the same person is decreased while distances between features belonging to different people are increased. During testing, for each of the 54 people, 4 out of 6 normal walking videos are put into the gallery set, while the remaining 2 normal walking videos and the videos of carrying a bag and wearing a coat are put into the query set. If a video in the query set is correctly matched with the same person's video in the gallery set, it is counted as successful recognition.

For the lower branch of our method, we use different image-based approaches, namely GaitSet, Gaitpart and GaitGL, and compare the performance of our proposed approach (incorporating skeleton point features via PointNet), with the performance of the image- and CNN-based approaches. For a commensurate comparison, all the models are run on our local machine, with exactly the same training and testing environment, e.g., the same optimizer, batch size, learning rate etc. We set K, M and T as 17, 0.2 and 30, respectively, in our experiments. Table 3 shows the results, including the accuracy comparison for different view angles and walking scenarios. In the table, our approach is listed as Baseline + PointNet, and it improves the performance of the corresponding image- and CNN-based baseline for all walking scenarios in terms of mean accuracy over different view angles. The overall best performance is achieved by GaitGL+PointNet for all walking scenarios. Based on these results, we can draw the following conclusions: (i) the features extracted by the point cloud analysis model improve the performance of all the image- and CNN-based gait recognition networks with varying degrees for all walking scenarios; (ii) improvement is positively correlated with how different contours of the query and gallery images are. All gallery images in the dataset belong to the normal walking scenario. For all benchmarks, our approach provides the most improvement for the wearing coat scenario, since the contour of a person wearing a coat is more different than that of a normal case, as explained in Sec. 3.1; (iii) features from the point cloud analysis model not only alleviate the problem of CNNs overfitting to person contours but also increase the robustness to different view angles. Our method improves the accuracy for more than half of the view angles in most cases, indicating that the point cloud analysis model can learn general motion features from human key points under different view angles.

5. CONCLUSION

We have proposed a new gait recognition network, GaitPoint, to address the limitations of the only image- and CNN-based methods and to increase the robustness against the variations in appearance due to different viewing angles or clothing or carried items. By incorporating features from human pose key points, via point cloud analysis, our approach improves the performance of different image- and CNN-based benchmarks, which shows the generalizability of our method.

References

[1] H. Chao, Y. He, J. Zhang, and J. Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, volume 33, pages 8126–8133, 2019.

[2] T. Chen et al. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[3] C. Fan et al. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14233, 2020.

[4] M. Goffredo et al. Self-calibrating view-invariant gait biometrics. *IEEE Trans. on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(4):997–1008, 2009.

[5] S. Hou et al. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *ECCV*, pages 382–398. Springer, 2020.

[6] F. Jean et al. Towards view-invariant gait modeling: Computing view-normalized body part trajectories. *Pattern Recognition*, 42(11):2936–2949, 2009.

[7] W. Kusakunniran et al. Support vector regression for multi-view gait recognition based on local motion feature selection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 974–981. IEEE, 2010.

[8] T. Le and Y. Duan. Pointgrid: A deep network for 3d shape understanding. In *CVPR*, pages 9204–9214, 2018.

[9] X. Li et al. End-to-end model-based gait recognition. In *ACCV*, 2020.

[10] B. Lin et al. Gait recognition via effective global-local feature representation and local temporal aggregation. In *ICCV*, pages 14648–14656, 2021.

[11] N. Liu et al. Joint subspace learning for view-invariant gait recognition. *IEEE Signal processing letters*, 18(7):431–434, 2011.

[12] M. Mao and Y. Song. Gait recognition based on 3d skeleton data and graph convolutional network. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2020.

[13] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IROS*, pages 922–928. IEEE, 2015.

[14] Y. Peng et al. Learning rich features for gait recognition by integrating skeletons and silhouettes. *arXiv preprint arXiv:2110.13408*, 2021.

[15] C. R. Qi et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*, 2017.

[16] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017.

[17] A. Sepas-Moghaddam and A. Etemad. View-invariant gait recognition with attentive recurrent learning of partial representations. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(1):124–137, 2020.

[18] Y.-F. Song et al. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In *ACM International Conference on Multimedia*, pages 1625–1633, 2020.

[19] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019.

[20] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. *arXiv preprint arXiv:2101.11228*, 2021.

[21] M. Z. Uddin et al. Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion. *IPSJ Trans. on Computer Vision and Applications*, 11(1):1–18, 2019.

[22] Y. Wang et al. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019.

[23] Z. Wu et al. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):209–226, 2016.

[24] X. Xing et al. Complete canonical correlation analysis with application to multi-view gait recognition. *Pattern Recognition*, 50:107–117, 2016.

[25] S. Yu et al. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *CVPR*, volume 4, pages 441–444, 2006.

[26] Y. Zhang et al. Cross-view gait recognition by discriminative feature learning. *IEEE Transactions on Image Processing*, 29:1001–1015, 2019.

[27] X. Zhu et al. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *CVPR*, pages 9939–9948, 2021.