Optimal Transport Maps between Walking Variations

Adrian Cosma[*] **and Emilian Radoi**

University Politehnica of Bucharest, Bucharest, Romania

Abstract. Gait, the manner of walking, has been proven to be a reliable biometric with uses in surveillance, marketing and security. A promising new direction for the field is training gait recognition systems without explicit human annotations, through self-supervised learning approaches. Such methods are heavily reliant on strong augmentations for the same walking sequence to induce more data variability and to simulate additional walking variations. Current data augmentation schemes are heuristic and cannot provide the necessary data variation as they are only able to provide simple temporal and spatial distortions. In this work, we propose GaitMorph, a novel method to modify the walking variation for an input gait sequence. Our method entails the training of a high-compression model for gait skeleton sequences that leverages unlabelled data to construct a discrete and interpretable latent space, which preserves identity-related features. Furthermore, we propose a method based on optimal transport theory to learn latent transport maps on the discrete codebook that morph gait sequences between variations. We perform extensive experiments and show that our method is suitable to synthesize additional views for an input sequence.

1 Introduction

The way people walk, also known as gait, is a crucial biometric trait that has numerous applications in medicine [13], sports [22], and surveillance[7]. Most notably, in recent years, it has been successfully used as a unique biometric fingerprint to accurately identify individuals from a distance [5]. The biggest challenge in gait analysis [11] is disentangling confounding factors which significantly affect and obfuscate gait, such as the individual clothing, footwear, walking speed, injury, state of mind, and social environment. Moreover the extrinsic characteristics of gait sensors (such as camera viewpoint, distance and resolution) severely affect the quality of the captured gait. Developing a robust model, able to ignore these factors and represent the essential gait characteristics is still an open problem. Previous works [5, 4] have shown that self-supervised pretraining is a promising new direction, but is still not enough to achieve high performance modelling. However, contrastive pre-training requires high degree of variation in the data [2, 27], which is often hard to obtain automatically for gait. Heuristical augmentation procedures are not able to reliably produce novel viewpoints for a gait sequence, or to seamlessly change the walking variation as they only provide simple temporal and spatial distortions. For other similar tasks such as person re-identification [34], viewpoint variation is induced through learned methods such as approaches in human pose transfer [23].

We propose GaitMorph, a novel method that is able to modify skeleton gait sequences to synthesize novel views. Our model is based on the vector-quantized variational autoencoder (VQ-VAE) [28]. Compressing gait sequences in a discrete latent space enables easy manipulation of codebook entries between walking variations. We propose to make use of optimal transport [29] to learn transport maps between walking variations, allowing morphing gait sequences into a desired variation or viewpoint.

2 Related Work

Works in motion sythetisation are predominantly directed towards generating controllable, general actions for use in animation [14, 20]. Yan et al. [32] proposed a convolutional architecture named Convolutional Sequence Generation Network (CSGN) for generating skeleton sequences for action recognition. The authors employed spatial graph downsampling and temporal downsampling to generate the whole sequence in a single pass, using latent vectors sampled from gaussian processes. Petrovich [19] employed a transformer VAE model conditioned on the action.

Li et. al [14] proposed a method for performing motion "in-betweening" using physically plausible constraints. Raab et al. [20] perform motion in-betweening by using diffusion models. Wang et al. [30] constructed a method for generating movement animations which also takes the target environment into account.

Some works tackle the problem of motion prediction [17, 38]. Ma et al. [17] used a graph-convolutional network for motion prediction of skeleton sequences. Zhang et al. [38] generate unbounded motion sequences conditioned only on a single starting skeleton. The authors employ an RNN-based architecture to procedurally generate skeletons.

Motion generation techniques have also been used for sign language generation [15, 31]. Liu et al. [15] used a cross-modal approach for audio to sign pose sequence generation using a GRU-based model. Xie et al. [31] used a VQ-VAE to generate sign pose sequences, using a discrete diffusion prior model. Zhang et al. [37] propose a Motion VQ-VAE for text-conditioned action generation, and demonstrate that a simple VQ-VAE recipe [21] can have very good performance for this data modality without any major bells and whistles.

In the area of gait recognition, synthesising walks has been only briefly studied in the past, partially due to the lack of large-scale datasets, and the unique constraints of this settings. Works in self-supervised for images[2, 9, 27] point out that the high quality data augmentation is crucial for learning good representations. Tian et al.[27] argues that optimal views for self-supervised contrastive learning are task-dependent. For instance, in gait analysis, Yu et al. [35] train a generative adversarial network to generate silhouette sequences that are invariant to walking confounding factors such as viewpoint and clothing change. However, the goal was downstream

[*] Corresponding Author. Email: ioan_adrian.cosma@upb.ro

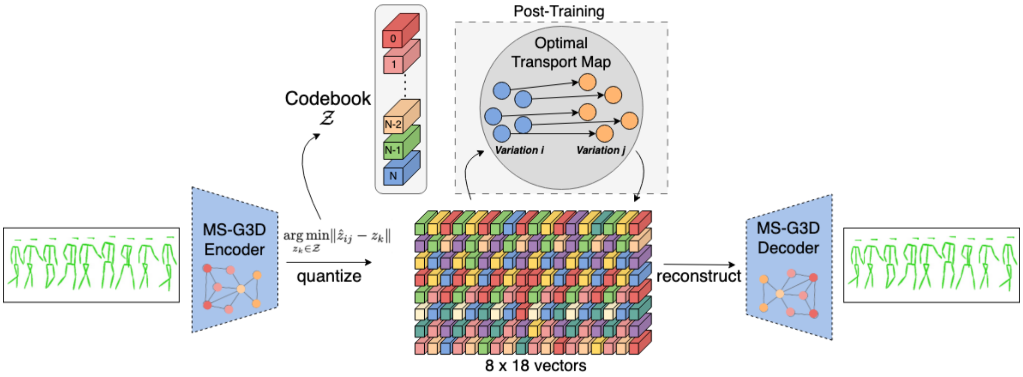

Figure 1. Overall architecture of GaitMorph. We train a MS-G3D encoder-decoder to quantize gait representations into a learned fixed-size codebook. After training, we can manipulate the discrete latent space and morph a walking variation into another using a transport map learned on the training set of a controlled walking dataset.

identification and not generation in itself. Yao et al. [3] propose a framework for walking synthetisation based on an autoencoder and a parametric body model, but their experiments are mainly based on silhouette-based identification models. Different from previous works, we are interested in manipulating the walking variation and viewpoint of existing walks.

3 Method

3.1 Learning a Discrete Latent Space

In order to train a sufficiently large and general autoencoder model, we assess that current gait datasets are too small. Even though datasets such as DenseGait [5] and GREW [40] are collected "in-the-wild" outdoor environments using surveillance cameras, they nonetheless lack some walking registers such as treadmill walking, more aggressive camera angles and indoor environments. However, by combining the major large-scale gait datasets into a single dataset, we can ensure more diversity of walking registers. We used **DenseGait** [5] and **GREW** [40], two similar in-the-wild datasets for their diverse walking sequences in outdoor environments, **OU-ISIR** [1] for more controlled walking in indoor and treadmill registers, and **Gait3D** [39], and indoor "in-the-wild" dataset collected in a supermarket setting. After concatenation of all skeleton sequences from the datasets, we obtain 875,543 walking sequences, totalling 1220.06 hours. To increase the size as much as possible, we also included the testing / distractor splits of each dataset whenever possible. We purposely did not include controlled, small scale datasets such as CASIA-B [36], as we use them for downstream evaluation.

In order to learn an informative and context-rich walking codebook, we leverage the expressive power of a Vector Quantized Variational AutoEncoder model (VQ-VAE) [28]. The VQ-VAE model has been shown to be effective for a range of tasks, including image compression and generation [6, 21], and speech recognition [28]. It is particularly useful in situations where the input data has a high degree of variability, and where traditional continuous latent space models may struggle to capture the underlying structure of the data. Furthermore, a discrete latent space enables a high degree of data compression, and allows the input data to be further processed as a sequence of discrete tokens.

To properly encode skeleton sequences, we construct a skeleton autoencoder based on the MS-G3D [16] model. Figure 1 showcases the overall architecture of our method. MS-G3D is a powerful graph convolutional model that has state-of-the-art results in skeleton action recognition, surpassing other graph-based methods [33, 24] by a large margin. Graph convolutional models are well established in the field of skeleton sequence processing [10] and were developed to properly handle spatial and temporal variation of the skeleton graph. For simplicity, we did not perform any graph subsampling [32], and only used temporal pooling to compress the skeleton sequence. We follow the official model implementation [16], and adapt it for gait processing. Specifically, we changed all activations to GeLU [12], we removed the initial data batch-normalization since skeletons were already normalized. Initial experiments showed that the default model was not large enough to reconstruct sequences other than the mean skeleton. Consequently, we doubled each convolution - batch normalization - activation block to increase model capacity.

3.2 Learning Optimal Transport Mappings

In order to exploit the expressive power of the learned gait tokens, we posit that only specific tokens from a tokenized gait sequence are responsible for encoding the gait viewpoint and variation. Therefore, for a set of walks from a particular variation $\mathcal{T}$, we can learn a set of transport maps $\Gamma = \{\gamma_j^* | j \in 1 \ldots (\frac{T}{4} \times J)\}$, for each encoded position j, that transform the target quantized gait representation into a quantized representation of a baseline walk $\mathcal{B}$. The transformed walk $\mathcal{T}$ is then decoded by the generator: $\mathcal{T}^* = G(\Gamma(\mathbf{q}(E(\mathcal{T}))))$. The walks $\mathcal{B}$ and $\mathcal{T}^*$ should be from the same walking variation. We propose to learn the transport maps Γ by utilizing optimal transport theory [29]. We learn a transport map γ_j^* by minimizing the Earth Mover's Distance (EMD) between the histograms of two quantized gaits. EMD assumes there is a cost for moving one quantity to another, which is encoded into a cost matrix C. In general, EMD is defined as:

$$\begin{aligned} \gamma^* &= \arg\min_{\gamma \in \mathbb{R}_+^{m \times n}} \sum_{i,j} \gamma_{i,j} C_{i,j} \\ &\text{s.t.} \gamma 1 = a; \gamma^T 1 = b; \gamma \geq 0 \end{aligned} \tag{1}$$

In our case, a and b are histograms of the token occurrences in each gait sequence, and the cost matrix C is given by the pairwise distances between the token embeddings. To account for multiple occurrence of the same token in a quantized gait sequence, we scale the corresponding vector embedding by the number of occurrences. We describe our method in Algorithm 1. The algorithm is an instance of an assignment problem for each token position, and is similar to finding the minimum flow between the two token distributions.

Algorithm 1 Finding the optimal transport maps between walking variations.

```
Require:
    $E$ - Trained MS-G3D gait encoder
    $\mathcal{B} \in \mathbb{R}^{B^{(b)} \times T \times J \times 2}$ - baseline variation walks
    $\mathcal{T} \in \mathbb{R}^{B^{(t)} \times T \times J \times 2}$ - target walks
    $\mathcal{Z}$ - learned codebook vectors
    $s$ - token sequence length

    $k^{(b)} \leftarrow \arg(\mathbf{q}(E(\mathcal{B})))$                    ▷ Baseline token indices.
    $k^{(t)} \leftarrow \arg(\mathbf{q}(E(\mathcal{T})))$                    ▷ Target token indices.
    $\Gamma \leftarrow \emptyset$
    for $j \leftarrow 1 \ldots s$ do
    ▷ Count occurrences of each baseline and target tokens.
        $c^{(b)} \leftarrow \{\sum_l^{B^{(b)}} \mathbb{1}[k_{l,j}^{(b)} = r] | r \in 1 \ldots |\mathcal{Z}|\}$
        $c^{(t)} \leftarrow \{\sum_l^{B^{(t)}} \mathbb{1}[k_{l,j}^{(t)} = r] | r \in 1 \ldots |\mathcal{Z}|\}$
    ▷ Increase codebook embedding magnitude.
        $C^{(b)} \leftarrow \mathcal{Z} \odot c^{(b)}$
        $C^{(t)} \leftarrow \mathcal{Z} \odot c^{(t)}$
    ▷ Compute cost matrix as pairwise distances between scaled token embeddings.
        $C \leftarrow C^{(b)} \cdot (C^{(t)})^\top$
    ▷ Find optimal transport map for position j
        $\gamma^* \leftarrow \arg\min_\gamma \sum \gamma C$                    ▷ Eq. 1
        $\Gamma_j \leftarrow \gamma^*$
    end for
    return $\Gamma$
```

4 Results

For gait synthetisation, we propose a specialized variant of the FID score, which we name "*Frechet Gait Distance (FGD)*", in which walks are processed by a pretrained GaitFormer network on DenseGait [5]. FGD stands as a automatic measure of walking "naturalness", by measuring the similarity to a given real gait distribution. Variants have been proposed for measuring motion naturalness and are geared towards general action synthesis [8, 25, 18], but a specialized variant for gait has not yet been adopted.

		0°	**72°**	**90°**	**126°**	**162°**	**180°**
NM	*Baseline (vs real NM-36)*	***0.045532***	*0.070282*	*0.111757*	*0.138415*	*0.19525*	*0.265378*
	Heuristic Aug. (vs real NM-36)	0.047659	0.076971	0.115972	0.138536	0.195943	0.27324
	$\lvert\mathcal{Z}\rvert = 2048$	0.046048	**0.060231**	**0.082002**	**0.102774**	**0.104749**	0.135883
BG	*Baseline (vs real NM-36)*	***0.05295***	*0.074746*	*0.114694*	*0.150358*	*0.211948*	*0.274384*
	Heuristic Aug. (vs real NM-36)	0.055826	0.083356	0.119362	0.152413	0.209289	0.283982
	$\lvert\mathcal{Z}\rvert = 2048$	0.056126	0.081991	0.106456	0.131166	0.137103	0.161214
CL	*Baseline (vs real NM-36)*	*0.110895*	*0.140185*	*0.189128*	*0.230226*	*0.320092*	*0.411968*
	Heuristic Aug. (vs real NM-36)	0.120972	0.147726	0.197784	0.235666	0.318236	0.420584
	$\lvert\mathcal{Z}\rvert = 2048$	0.075194	0.096743	0.128168	0.148594	0.159419	0.192654

Table 1. FGD values between the morphed gait to the NM-36 variation and the real NM-36 for CASIA-B validation set. Baseline values corresponds to the FGD between the real unmodified gait and NM-36. In most variations, the morphed walk is much closer to the real NM-36 than the unmodified walk, especially for extreme viewpoints. We denote with **bold** the smallest distance and with underline the second smallest distance.

In Table 1, we present our results for gait morphing for CASIA-B, respectively. We utilized the proposed FGD metric to compare the distance between the distribution of the morphed walks to the real baseline walking variation (NM-36 for CASIA-B). For CASIA-B we focus our evaluation in terms of viewpoint, since it is the principal confounding factor, especially for 2D poses. Results show that the morphed walks are properly generated and are closer to the real NM-36 walking variation compared to the unmodified walk and for more extreme viewpoints, the effect is larger. Results are more correlated with the dictionary usage for each dictionary size, rather than reconstruction error (which is low for every dictionary size). Additionally, we compared morphed gaits with standard array of heuristic skeleton augmentations present in other works[5, 26]: random pace with a time multiplier sampled from $\{0.5, 0.75, 1, 1.25, 1.5, 1.75, 2.0\}$, joint and point noise with standard deviation of 0.001, random mirroring and reversing the walk. While heuristic augmentations provide some variation in the vicinity of the original walk, the FGD across views are similar to the non-augmented walks. These results show that the morphed walks with our method are a reasonable way to augment existing walks to synthesize novel views.

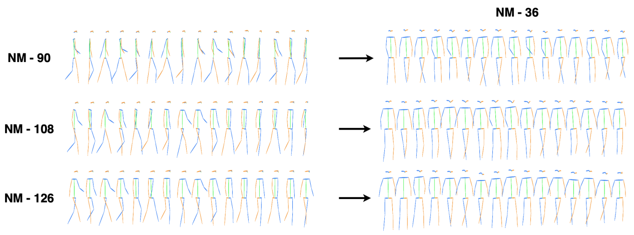

Figure 2. Examples of modified skeleton sequences using optimal transport maps. We differentiate left and right laterals with appropriate colors. The model is able to successfully change the walking viewpoint to a normal walk under viewpoint 36° (NM-36). For this example, we chose a VQ-VAE with $|\mathcal{Z}| = 512$. Best viewed electronically, zoomed-in and in color.

Figure 2 showcases selected gait sequences from three different viewpoints morphed to a common NM-36 variation. The model is able to morph sequences into the baseline sequence, properly handling limb switching (left and right limbs are properly swapped when the viewpoint is from behind the walker). For similar baseline / target pairs, the transport maps exhibit fewer changes.

Models operating with a low dictionary size are not appropriate to be used for morphing. This is most likely due to the latent embeddings being severely entangled. Figure 3 showcases a selected failure case for morphing a NM-180 walk from CASIA-B into NM-36 using a VQ-VAE with $|\mathcal{Z}| = 8$. The generated walk has severe artifacts and cannot be considered appropriate for downstream model training. Inherently, there is a trade-off between dictionary size and the malleability of the latent codes: larger dictionary sizes have more disentangled representations which allow for more informed changes at the expense of lower data compression.

References

[1] Weizhi An, Shiqi Yu, Yasushi Makihara, Xinhui Wu, Chi Xu, Yang Yu, Rijun Liao, and Yasushi Yagi, 'Performance evaluation of model-based gait on multi-view very large population database with pose sequences', *IEEE Trans. on Biometrics, Behavior, and Identity Science*, (2020).

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, 'A simple framework for contrastive learning of visual representations', *arXiv preprint arXiv:2002.05709*, (2020).

[3] Yao Cheng, Guichao Zhang, Sifei Huang, Zexi Wang, Xuan Cheng, and Juncong Lin, 'Synthesizing 3d gait data with personalized walking style and appearance', *Applied Sciences*, **13**(4), (2023).

[4] Adrian Cosma, Andy Catruna, and Emilian Radoi, 'Exploring self-supervised vision transformers for gait recognition in the wild', *Sensors*, **23**(5), (2023).

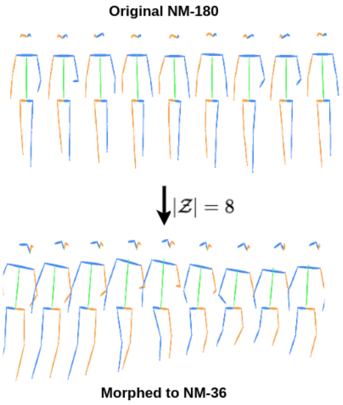

Figure 3. Failure case for $|\mathcal{Z}| = 8$ when morphing a normal walk from CASIA-B from viewpoint 180° to viewpoint 36°. The latent space is not sufficiently disentangled to learn a general transport map without severely distorting the resulting gait sequence.

[5] Adrian Cosma and Emilian Radoi, ‘Learning gait representations with noisy multi-task learning’, *Sensors*, **22**(18), (2022).

[6] Patrick Esser, Robin Rombach, and Björn Ommer, ‘Taming transformers for high-resolution image synthesis. in 2021 ieee’, in *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12868–12878, (2020).

[7] Claudio Filipi Gonçalves dos Santos, Diego de Souza Oliveira, Leandro A. Passos, Rafael Gonçalves Pires, Daniel Felipe Silva Santos, Lucas Pascotti Valem, Thierry P. Moreira, Marcos Cleison S. Santana, Mateus Roder, Jo Paulo Papa, and Danilo Colombo, ‘Gait recognition based on deep learning: A survey’, *ACM Comput. Surv.*, **55**(2), (jan 2022).

[8] Deepak Gopinath and Jungdam Won. fairmotion - tools to load, process and visualize motion capture data. Github, 2020.

[9] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding, ‘Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition’, in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 762–770, (2022).

[10] Pranay Gupta, Anirudh Thatipelli, Aditya Aggarwal, Shubh Maheshwari, Neel Trivedi, Sourav Das, and Ravi Kiran Sarvadevabhatla, ‘Quo vadis, skeleton action recognition?’, *International Journal of Computer Vision*, **129**(7), 2097–2112, (2021).

[11] Elsa J. Harris, I-Hung Khoo, and Emel Demircan, ‘A survey of human gait-based artificial intelligence applications’, *Frontiers in Robotics and AI*, **8**, (2022).

[12] Dan Hendrycks and Kevin Gimpel, ‘Gaussian error linear units (gelus)’, *arXiv preprint arXiv:1606.08415*, (2016).

[13] Shantanu Jana, Nibaran Das, Subhadip Basu, and Mita Nasipuri, ‘Survey of human gait analysis and recognition for medical and forensic applications’, *International Journal of Digital Crime and Forensics*, **13**, 1–20, (11 2021).

[14] Yunhao Li, Zhenbo Yu, Yucheng Zhu, Bingbing Ni, Guangtao Zhai, and Wei Shen, ‘Skeleton2humanoid: Animating simulated characters for physically-plausible motion in-betweening’, in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1493–1502, (2022).

[15] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou, ‘Learning hierarchical cross-modal association for co-speech gesture generation’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10462–10472, (2022).

[16] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang, ‘Disentangling and unifying graph convolutions for skeleton-based action recognition’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 143–152, (2020).

[17] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li, ‘Progressively generating better initial guesses towards next stages for high-quality human motion prediction’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6437–6446, (2022).

[18] Antoine Maiorca, Youngwoo Yoon, and Thierry Dutoit, ‘Evaluating the quality of a synthesized motion with the fréchet motion distance’, in *ACM SIGGRAPH 2022 Posters*, 1–2, (2022).

[19] Mathis Petrovich, Michael J Black, and Gül Varol, ‘Action-conditioned 3d human motion synthesis with transformer vae’, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10985–10995, (2021).

[20] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit H Bermano, and Daniel Cohen-Or, ‘Single motion diffusion’, *arXiv preprint arXiv:2302.05905*, (2023).

[21] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals, ‘Generating diverse high-fidelity images with vq-vae-2’, *Advances in neural information processing systems*, **32**, (2019).

[22] Erin Ross, Anthony Milian, Mason Ferlic, Samuel Reed, and Adam S. Lepley, ‘A data-driven approach to running gait assessment using inertial measurement units’, *Video Journal of Sports Medicine*, **2**(5), 26350254221102464, (2022).

[23] Soubhik Sanyal, Alex Vorobiov, Timo Bolkart, Matthew Loper, Betty Mohler, Larry S Davis, Javier Romero, and Michael J Black, ‘Learning realistic human reposing using cyclic self-supervision with 3d shape, pose, and appearance consistency’, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11138–11147, (2021).

[24] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu, ‘Two-stream adaptive graph convolutional networks for skeleton-based action recognition’, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12026–12035, (2019).

[25] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu, ‘Bailando: 3d dance generation via actor-critic gpt with choreographic memory’, in *CVPR*, (2022).

[26] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll, ‘GaitGraph: Graph convolutional network for skeleton-based gait recognition’, in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 2314–2318, (2021).

[27] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola, ‘What makes for good views for contrastive learning?’, *Advances in neural information processing systems*, **33**, 6827–6839, (2020).

[28] Aaron Van Den Oord, Oriol Vinyals, et al., ‘Neural discrete representation learning’, *Advances in neural information processing systems*, **30**, (2017).

[29] Cédric Villani et al., *Optimal transport: old and new*, volume 338, Springer, 2009.

[30] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai, ‘Towards diverse and natural scene-aware 3d human motion synthesis’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20460–20469, (2022).

[31] Pan Xie, Qipeng Zhang, Zexian Li, Hao Tang, Yao Du, and Xiaohui Hu, ‘Vector quantized diffusion model with codeunet for text-to-sign pose sequences generation’, *arXiv preprint arXiv:2208.09141*, (2022).

[32] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin, ‘Convolutional sequence generation for skeleton-based action synthesis’, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4394–4402, (2019).

[33] 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*, volume 32, (2018).

[34] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. H. Hoi, ‘Deep learning for person re-identification: A survey and outlook’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**(06), 2872–2893, (jun 2022).

[35] Shiqi Yu, Haifeng Chen, Edel B. García Reyes, and Norman Poh, ‘Gaitgan: Invariant gait feature extraction using generative adversarial networks’, in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 532–539, (2017).

[36] Shiqi Yu, Daoliang Tan, and Tieniu Tan, ‘A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition’, in *18th International Conference on Pattern Recognition (ICPR’06)*, volume 4, pp. 441–444. IEEE, (2006).

[37] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen, ‘T2m-gpt: Generating human motion from textual descriptions with discrete representations’, *arXiv preprint arXiv:2301.06052*, (2023).

[38] Yan Zhang, Michael J Black, and Siyu Tang, ‘Perpetual motion: Generating unbounded human motion’, *arXiv preprint arXiv:2007.13886*, (2020).

[39] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei, ‘Gait recognition in the wild with dense 3d representations and a benchmark’, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022).

[40] Zheng Zhu, Xianda Guo, Tian Yang, Junjie Huang, Jiankang Deng, Guan Huang, Dalong Du, Jiwen Lu, and Jie Zhou, ‘Gait recognition in the wild: A benchmark’, in *IEEE International Conference on Computer Vision (ICCV)*, (2021).