GaitMorph: Transforming Gait by Optimally Transporting Discrete Codes

Adrian Cosma, Emilian Rădoi

University Politehnica of Bucharest, Bucharest, Romania

`cosma.i.adrian@gmail.com, emilian.radoi@upb.ro`

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 [23], sports [38], and surveillance[11]. Most notably, in recent years, it has been successfully used as a unique biometric fingerprint to accurately identify individuals from a distance [8]. The biggest challenge in gait analysis [19] 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. Con-

sequently, deploying highly accurate gait recognition systems in real-world unconstrained scenarios remains a difficult problem.

Previous works [8, 7] have shown that self-supervised pretraining is a promising new direction, but is still not enough to achieve high performance modelling. Self-supervised pre-training exposes the backbone model to a large variety of walking registers, increasing the robustness in downstream tasks. However, contrastive pre-training requires high degree of variation in the data [5, 47], which is often hard to obtain automatically for gait. Fine-tuning is still necessary for effective gait recognition [7], especially in specific and uncommon environments. Currently, data variation for gait modelling systems is obtained by using data augmentation techniques [8], which have the goal to distort the gait sequence while preserving the person identity. However, 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 [56], viewpoint variation is induced through learned methods such as approaches in human pose transfer [39].

We propose GaitMorph, a novel method that is able to modify skeleton gait sequences to synthesize novel views. We use a high-compression model for gait sequences that leverages large amounts of in-the-wild and unlabelled data to construct a discrete and interpretable latent space for skeleton gait sequences. Our model is based on the vector-quantized variational autoencoder (VQ-VAE) [48], and achieves a high degree of compression for skeleton sequences (up to $500\times$ lower storage demands). We show that the model is able to reconstruct gait sequences with a high degree of fidelity, without losing identity-related features.

Furthermore, 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 [49] to learn transport maps between walking variations, allowing morphing gait sequences into a desired variation or viewpoint.

This work makes the following contributions:

1. We demonstrate that compression of gait sequences into a discrete latent space is feasible, and can be achieved while preserving identity-related information of the underlying walker. We achieve a high degree (500×) of compression without deteriorating downstream gait recognition performance (maximum of 3% accuracy loss for normal walking).

2. We propose a novel method to morph gait sequences between variations. Using optimal transport theory, we learn transport maps between variations that generate realistic and novel views for a walk. Our experiments show that the distribution of morphed sequences is similar to the real walk distribution, potentially making our method useful for data augmentation.

3. We perform extensive experiments on the core aspect of our proposed method: the VQ-VAE dictionary size. We show that, while a small dictionary size obtains good reconstruction error at a high compression level, the latent space is not sufficiently disentangled to allow easy morphing.

2. Related Work

Works in motion sythetisation are predominantly directed towards generating controllable, general actions for use in animation [42, 27, 36, 3]. Yan et al. [54] 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 [34] employed a transformer VAE model conditioned on the action.

Li et. al [27] proposed a method for performing motion "in-betweening" using physically plausible constraints. Raab et al. [36] perform motion in-betweening by using diffusion models. Wang et al. [51] constructed a method for generating movement animations which also takes the target environment into account.

Some works tackle the problem of motion prediction [17, 31, 33]. Ma et al. [31] used a graph-convolutional network for motion prediction of skeleton sequences. Zhang et al. [63] 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 [29, 53]. Liu et al. [29] used a cross-modal approach for audio to sign pose sequence generation using a GRU-based model. Xie et al. [53] used a VQ-VAE to generate sign pose sequences, using a discrete diffusion prior model. Zhang et al. [61] propose a Motion VQ-VAE for text-conditioned action generation, and demonstrate that a simple VQ-VAE recipe [37] 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[5, 16, 47] point out that the high quality data augmentation is crucial for learning good representations. Tian et al.[47] argues that optimal views for self-supervised contrastive learning are task-dependent. For instance, in gait analysis, Yu et al. [58] 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 identification and not generation in itself. Yao et al. [6] 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

The use of a VQ-VAE [48] for learning a latent walking representation for skeletons is motivated by the discrete nature of the latent embeddings, which simplifies the constraint optimization for morphing between walking variations. While the VQ-VAE is widely used in generative modelling [61, 53, 29], other algorithms such as diffusion models [62] might offer higher quality reconstructions. However, our aim is not to generate walking sequences, but to manipulate the latent space to change existing walks into desired variations.

In this section, we describe the main components of GaitMorph: we describe the pretraining dataset used for training the VQ-VAE, the architecture and training procedure for the VQ-VAE, and the morphing algorithm based on optimal transport between the latent codes.

3.1. Pretraining Dataset

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 [8] and GREW [66] 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. In Table 1 we showcase the existing datasets that comprises our pretraining dataset. We used **DenseGait** [8] and **GREW** [66], 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** [65], 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 [59], as we use them for downstream evaluation.

All walking sequences in this dataset are 2D skeletons in COCO pose format. We chose 2D poses to unify all datasets, as every dataset is providing 2D poses by default, while only some are also providing silhouettes or body meshes. Even though many gait processing models have good results using and appearance-based approach with silhouettes [4, 28], pose sequences only encode movement and abstract away any appearance information, preserving the privacy of walking individuals [45, 8]. Skeleton sequences are a more interpretable and a plethora of models employ them for motion synthesis [42, 61, 46, 35].

Skeleton sequences from each datasets are pre-processed in the same way. We filtered out skeletons that have too small or too large joint variance, which corresponds to static or erratic movement, respectively. We found that this procedure ensures that only properly moving skeletons are kept in the dataset. Furthermore, skeleton sequences are normalized and aligned at the pelvis, using the following formulae, considering that each of the $J = 18$ joints have (x, y) coordinates:

$$\hat{x}_{joint} = \frac{x_{joint} - x_{pelvis}}{|x_{R.shoulder} - x_{L.shoulder}|}$$

$$\hat{y}_{joint} = \frac{y_{joint} - y_{pelvis}}{|y_{neck} - y_{pelvis}|}$$

Dataset	Split	# Sequences	Duration (hr.)
DenseGait [8]	Train	217,954	614.75
	Validation	10,733	36.53
GREW [66]	Train	102,888	175.92
	Test	24,000	65.37
	Distractor	226,588	154.82
OU-ISIR [1]	Train	133,872	57.82
	Test	134,199	57.92
Gait3D [65]	Train	18,940	42.70
	Test	6,369	14.23
Total		**875,543**	**1220.06**

Table 1. Datasets that make up our pretraining dataset. We combined all the major in-the-wild and controlled datasets (including all splits) into a single, large-scale and diverse dataset. The dataset contains gait samples from a diverse set of walking registers, environments and camera angles.

In this manner, every skeleton sequence is aligned spatially and the differences in height and width of individuals are essentially eliminated. Consequently, only movement is encoded irrespective of the screen coordinates, distance to camera or appearance cues. We employ minimal augmentations to the skeleton sequences, adopting only random temporal cropping and walking pace modifications [26, 50, 8]. We crop each skeleton sequence to be $T = 64$ frames long.

3.2. Learning the Walking Codebook

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) [48]. The VQ-VAE model has been shown to be effective for a range of tasks, including image compression and generation [10, 37], and speech recognition [48]. 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 [30] 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 [55, 40] by a large margin. Graph convolutional models are well established in the field of skeleton sequence processing [18] and were developed to properly handle spatial and temporal variation of the skeleton graph.

3.2.1 MS-G3D Encoder-Decoder

The encoder and decoder models for the skeleton autoencoder are both based on the MS-G3D architecture [30]. For simplicity, we did not perform any graph subsampling [54], and only used temporal pooling to compress the skeleton sequence. We follow the official model implementation [30], and adapt it for gait processing. Specifically, we changed all activations to GeLU [20], we removed the initial data batch-normalization since the 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. A MS-G3D model is composed of multiple Spatial-Temporal Graph Convolution (ST-GC) blocks. Each block consists of a Multi-scale Graph Convolution block (MS-GCN) and two Multi-Scale Temporal Convolutional blocks [30]. We used the default 6 G3D scales and 13 GCN scales for both the encoder and decoder models.

For the MS-G3D encoder $E(\cdot)$, we used 20 ST-GC encoder blocks. We used a feature map size of 64 for the first 5
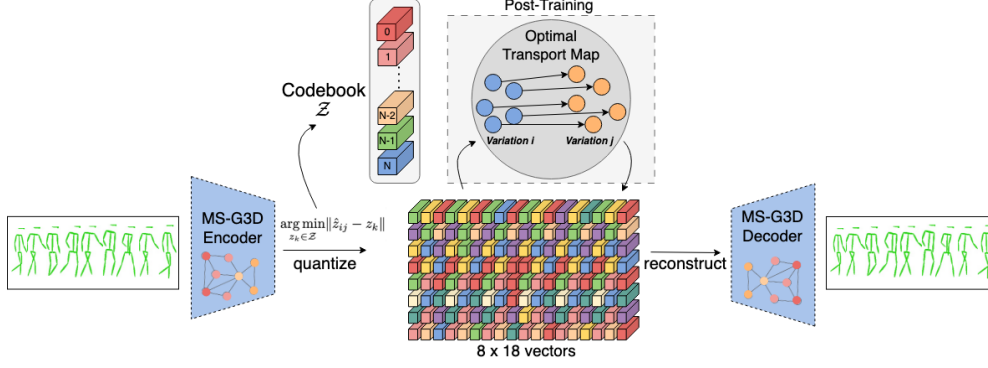
Figure 1. Overall architecture of GaitMorph. We train a MS-G3D encoder-decoder to quantize gait representations into a learned fixed-size codebook. After training, we can manipulate the discrete latent space and morph a walking variation into another using a transport map learned on the training set of a controlled walking dataset.

blocks, 128 for the next 5 and 256 for the final 10. Temporal pooling is performed every 5 blocks. Therefore, for an initial skeleton sequence $x \in \mathbb{R}^{T \times J \times 2}$ consisting of $T = 64$ skeletons (i.e. frames) with $J = 18$ joints, the sequence is temporally downsampled to $\hat{z} \in \mathbb{R}^{\frac{T}{4} \times J \times n_{\hat{z}}}$, where, in our case, $n_{\hat{z}} = 256$ the encoder embedding size.

For the MS-G3D decoder $G(\cdot)$, we opted for a slightly smaller model, since we experimentally observed that the encoder size is more negatively correlated to the final reconstruction error than the decoder size. Moreover, having a smaller decoder is more computationally efficient, and enables faster reconstruction of latent codes. The overall constituent decoder blocks are identical to the encoder blocks, but we replaced the strided convolution with a strided deconvolutional block for the temporal upsampling. We used 16 ST-GC blocks, with feature maps of size 32 for the first 4 blocks, 16 for the next 4 and 8 for the final 8. Temporal upsampling was performed every 4 blocks.

3.2.2 Skeleton Vector Quantization

Instead of utilizing a continuous latent space to encode the skeleton sequences, we quantize each latent embedding into a fixed-length learnable codebook $\mathcal{Z} = \{z_k\}_{k=1}^{K} \subset \mathbb{R}^{n_z}$. Any skeleton sequence $x \in \mathbb{R}^{T \times J \times 2}$ is encoded using the MS-G3D encoder described above into a temporally-compressed representation $\hat{z} \in \mathbb{R}^{\frac{T}{4} \times J \times n_{\hat{z}}}$, which is then quantized into $z_{\mathbf{q}} \in \mathbb{R}^{\frac{T}{4} \times J \times n_z}$, where $n_{\mathbf{z}}$ is the codebook dimensionality, not necessarily equal to the encoder embedding size. Each $\hat{z}$ is encoded using a nearest neighbor search in the codebook (see Eq. 1).

$$z_{\mathbf{q}} = \mathbf{q}(\hat{z}) \coloneqq \underset{z_k \in \mathcal{Z}}{\arg\min} \|\hat{z}_{ij} - z_k\| \in \mathbb{R}^{T \times J \times n_z} \quad (1)$$

After quantization, skeletons are reconstructed using the MS-G3D decoder: $\hat{x} = G(z_{\mathbf{q}}) = G(\mathbf{q}(E(x)))$. The model

is trained end-to-end using a stop-gradient operation (see Eq. 2) since the dictionary look-up is not differentiable. For more details regarding training, readers are referred to the work of Van Den Oord et al. [48].

$$\mathcal{L}_{VQ}(E, G, \mathcal{Z}) = \|x - \hat{x}\| + \|\text{sg}[E(x)] - z_{\mathbf{q}}\|_2^2 \\ + \|\text{sg}[z_{\mathbf{q}}] - E(x)\|_2^2 \quad (2)$$

In practice, instead of the l_2 loss for the reconstruction error, we employed a Smooth l_1 with $\beta = 0.25$ [14] to further penalize small reconstruction errors. VQ-VAE models are notoriously hard to train [25], primarily due to the dictionary collapse problem, in which most of the codebook entries are not utilized in reconstruction, yielding poor performance. To deal with this problem, we employed a standard array of "bag-of-tricks" to increase codebook usage. We used K-means initialization of the codebook [60], we used a lower codebook dimensionality of $n_z = 16$ by linearly projecting down the encoder embedding, we used cosine similarity for codebook search [57], expiring stale codes [60] and orthogonal regularization [41] of the codebook vectors to encourage linear independence. The codebook is learned using an exponential moving average approach with a decay rate of $\gamma = 0.9$. Autoencoder warm-up [13] was not necessary. We experimented with using a separate codebook for each limb, similar to Xie et al. [53], but did not observe a substantial improvement.

Training was performed on a single NVIDIA RTX 3060, using mixed-precision training, with a batch size of 48. The network was updated for 50k steps, using AdamW [24] optimizer using a cyclical learning rate schedule [43] which varies the learning rate between 0.0025 and 0.0075. The model has 4.8M non-embedding parameters. The training duration for the VQVAE is approximately 15 hours.

3.3. 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 [49]. 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:

$$\gamma^* = \arg\min_{\gamma \in \mathbb{R}_+^{m \times n}} \sum_{i,j} \gamma_{i,j} C_{i,j} \tag{3}$$
$$\text{s.t.} \gamma 1 = a; \gamma^T 1 = b; \gamma \geq 0$$

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. In practice, we use the algorithm proposed by Bonneel et al. [2] implemented in the PyOT [12] library.

In practical scenarios where the gait variation is not known beforehand, domain-expert models such as pedestrian attribute identification models [52] can be used to estimate particular walking attributes, similar to the approach of Cosma and Radoi [8], to inform the morphing target. This method can also be used as data augmentation to generate multiple views for the same walking sequence, for use in contrastive self-supervised training [22, 8].

4. Experiments and Results

For our experiments, we used CASIA-B [59] and Front-View Gait (FVG) [64] to evaluate the performance of our proposed method. CASIA-B is a popular gait recognition dataset, widely used to test the robustness of gait analysis model across multiple viewpoints and walking variations. It contains gait sequences from 124 subjects, captured in 11 different viewpoints, under three walking variations: normal walking (NM), clothing walking (CL) and walking with

---

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. 3
    $\Gamma_j \leftarrow \gamma^*$
  end for
  return Γ

---

a bag (BG). For a walker, 6 sessions are captured under the normal walking variation, 2 sessions with clothing change and 2 sessions under the bag carrying variation. Each walk has its variation / viewpoint known. We use the standard [59] training / testing split, consisting of the first 62 training subjects, with all available walking sessions. FVG is another popular dataset for gait recognition that features walks from only the front facing viewpoint, which is considered the most challenging due to the reduced perceived movement variation. It is comprised of 226 subjects captured in 6 walking variations: normal walk (NM), walk speed (WS), change in clothing (CL), carrying bag (CB), cluttered background (CBG) and ALL. In our experiments, we omit the "ALL" variation to properly isolate confounding factors. We used the first 136 subjects as the training split, and the rest for testing. It is important to note that the VQ-VAE model is trained on the dataset described in Section 3.1, and remains frozen throughout the rest of the experiments. Moreover, the transport maps are learned only on the training split of each dataset (CASIA-B / FVG) and are utilized as-is on the testing split.

4.1. The effect of dictionary size on the reconstructed gait sequences

We trained several VQ-VAE models with increasingly larger dictionary sizes ($|\mathcal{Z}| \in \{2, 8, 16, 32, 128, 512, 2048\}$) to gauge the effect on the reconstructed sequences. In Figure 3, we showcase the reconstruction error for each model on CASIA-B evaluation set, for each walking variation. For $|\mathcal{Z}| = 2$, the results are significantly worse than other dictionary sizes,
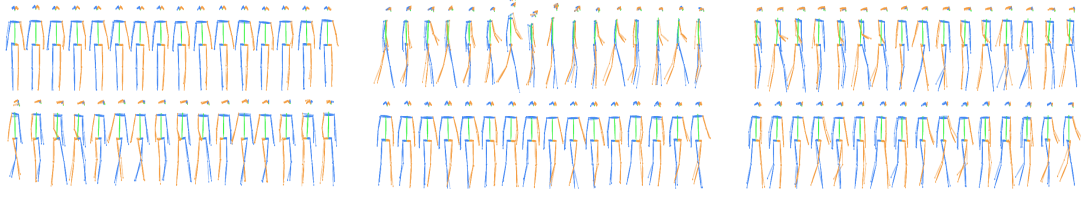
Figure 2. Overlapped original and reconstructed gait sequences from CASIA-B validation set using a VQ-VAE with $|\mathcal{Z}| = 8$. We differentiate left and right laterals with appropriate colors. Original skeletons are transparent, while reconstructed skeletons are opaque. The model can properly reconstruct sequences, and acts as a low-pass filter on the skeleton sequence, dampening exaggerated movements caused by inaccurate pose extraction (middle-top). Best viewed in color.
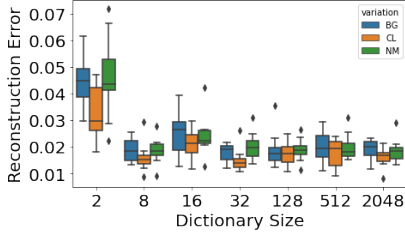

Figure 3. Boxplots for the across viewpoint distribution of reconstruction errors for different dictionary sizes.
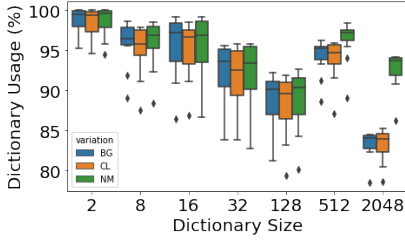

Figure 4. Dictionary usage for each of the trained VQ-VAE models. Increasing the dictionary size is correlated with lower dictionary usage and codebook under-utilization.

due to the extreme compression. The model with $|\mathcal{Z}| = 8$ achieves the best overall performance. We showcase qualitative reconstruction samples in Figure 2 - the model can reliably reconstruct skeleton sequences and acts as a low-pass filter which dampens exaggerated movements caused by inaccurate pose extractions.

Our model achieves a high degree of compression for skeleton sequences - a skeleton sequence represented as a float32 sequence of 2304 numbers is equivalent to storing 73728 bits of information, but using a VQ-VAE approach, the storage space is reduced to only 144 bits for $|\mathcal{Z}| = 2$ and 432 bits for $|\mathcal{Z}| = 8$. This compression level potentially allows on-device storage of massive amounts of skeleton sequences.

Figure 4 showcases the dictionary usage for each dictionary size. Increasing the dictionary size slightly decreases dictionary usage, which implies that some tokens are under-utilized by the model. This effect is more pronounced for $|\mathcal{Z}| = 2048$, especially for non-normal walking variations.

This is most likely due to the fact that the pretraining dataset for the VQ-VAE mostly contains in-the-wild walks, which make the BG and CL variations easier to reconstruct. The reconstructed gait sequences are not detrimental to downstream gait recognition models. To gauge the faithfulness of the reconstructed skeletons to the real walking skeletons, in Table 2 we showcase gait recognition results for CASIA-B using reconstructed skeletons as training data. The performance loss by using reconstructed skeletons is marginal, and even beneficial in some cases. This result can be attributed to the fact that the VQ-VAE acts as a low-pass filter and can slightly improve data quality across training. Furthermore, performance on gait recognition is not correlated with the reconstruction error of the VQ-VAE: the model with $|\mathcal{Z}| = 2$ achieves comparable results with the baseline method using real skeletons.

	NM	BG	CL
Baseline	0.79 ± 0.06	0.46 ± 0.08	0.24 ± 0.06
$|\mathcal{Z}| = 2$	0.76 ± 0.09	0.46 ± 0.09	0.24 ± 0.07
$|\mathcal{Z}| = 8$	0.75 ± 0.13	0.44 ± 0.08	0.21 ± 0.08
$|\mathcal{Z}| = 16$	0.75 ± 0.12	0.44 ± 0.08	0.22 ± 0.08
$|\mathcal{Z}| = 32$	0.76 ± 0.12	0.47 ± 0.1	0.23 ± 0.09
$|\mathcal{Z}| = 128$	0.78 ± 0.1	0.46 ± 0.09	0.22 ± 0.07
$|\mathcal{Z}| = 512$	0.76 ± 0.12	0.46 ± 0.09	0.22 ± 0.09
$|\mathcal{Z}| = 2048$	0.78 ± 0.11	0.47 ± 0.11	0.24 ± 0.1

Table 2. Accuracy on CASIA-B for GaitFormer [8] trained with reconstructed skeletons. We report the mean and standard deviation of recognition accuracy across 4 distinct runs and all viewpoints.

4.2. Evaluation of morphed gait sequences

We first present a qualitative evaluation for gait morphing. Figure 5 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. Transport maps from VQ-VAE models with a larger dictionary size exhibit more token changes, but the earth movers distance between variations is comparatively smaller. This implies that many
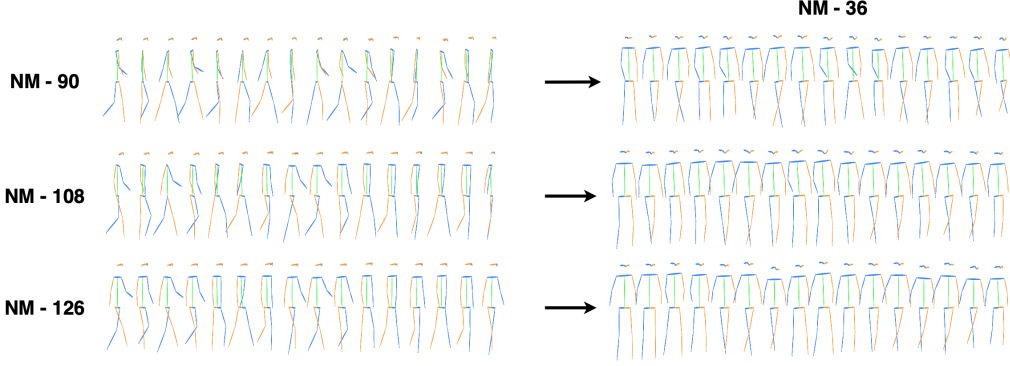
Figure 5. Examples of modified skeleton sequences using optimal transport maps. We differentiate left and right laterals with appropriate colors. The model is able to successfully change the walking viewpoint to a normal walk under viewpoint $36°$ (NM-36). For this example, we chose a VQ-VAE with $|\mathcal{Z}| = 512$. Best viewed in color.
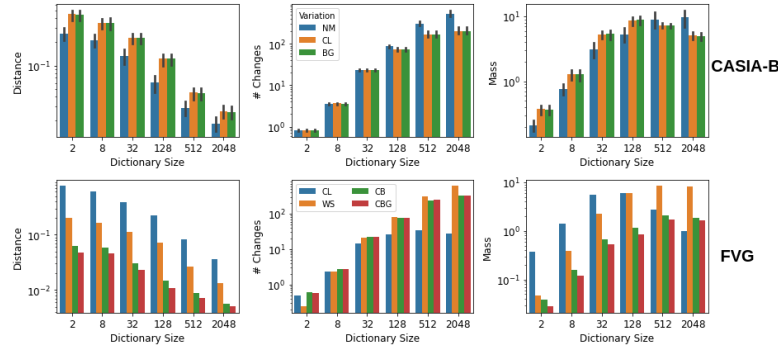

Figure 6. The average moved distance, number of changes and total mass moved between variations for CASIA-B and FVG. In the case of CASIA-B, the error bars represent standard deviation across viewpoints.

smaller changes are performed with same effect. In Figure 6 we showcase the average moved distance, the number of changes and the total mass moved across the walking variations for both CASIA-B and FVG. We define mass as the number of changes multiplied by the average change cost. The number of changes is larger when the dictionary size is larger, but the total mass moved remains constant after the $|\mathcal{Z}| = 128$. This implies that the tokens from the model with $|\mathcal{Z}| = 2048$ are more disentangled, since less mass is moved to achieve the same outcome.

In terms of numeric evaluation, our goal is to compare the morphed walks to the real walks of a particular variation. In our experiments, our comparisons are made with regard to NM-36 variation for CASIA-B and NM for FVG. The most straightforward comparison is to use mean squared error between skeletons, but we have no guarantees of sequence alignment between variations in either dataset. As such, we propose a metric between walking distributions, similar to the FID [21] distance. The Frechet Inception Distance (FID) was introduced by Heusel et al. [21] to measure the generation quality of GANs compared to real images. The FID score is based on the Frechet Distance [9], and measures the distance $d(\cdot)$ between two gaussian dis-

tributions $\phi = (\mathbf{m}, \mathbf{C})$ and $\phi_w = (\mathbf{m}_w, \mathbf{C}_w)$, corresponding to a distribution of real and synthetic samples, respectively: $d^2(\phi, \phi_w) = ||\mathbf{m} - \mathbf{m}_w||^2 + Tr(\mathbf{C} + \mathbf{C}_w - 2(\mathbf{CC}_w)^{\frac{1}{2}})$. The means and standard deviations of the Gaussians are, for images, the means and standard deviations of a set of embedding vectors of an Inception network [44] pretrained on ImageNet. The metric captures levels of perceived disturbance between real and synthetic samples [21].

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 Gait-Former network on DenseGait [8]. 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 [15, 42, 32], but a specialized variant for gait has not yet been adopted.

In Tables 3 and 4, we present our results for gait morphing for CASIA-B and FVG, 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 and NM for FVG). For CASIA-B we focus our evaluation in terms of viewpoint,

		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
	$|\mathcal{Z}| = 2$	0.754251	0.320832	0.315892	0.195995	0.339743	0.74698
	$|\mathcal{Z}| = 8$	0.379271	0.200036	0.104054	0.207928	0.673472	0.549586
	$|\mathcal{Z}| = 16$	0.136429	0.11086	0.22199	0.434574	0.530231	0.128465
	$|\mathcal{Z}| = 32$	0.091645	0.293709	0.400562	0.557834	0.63797	0.20792
	$|\mathcal{Z}| = 128$	0.08184	0.465142	0.597904	0.766841	0.697738	0.268043
	$|\mathcal{Z}| = 512$	0.074983	0.11027	0.117966	0.110944	0.140733	**0.107217**
	$|\mathcal{Z}| = 2048$	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
	$|\mathcal{Z}| = 2$	0.716497	0.304378	0.243575	0.193439	0.599968	0.743501
	$|\mathcal{Z}| = 8$	0.320088	0.184071	0.177406	0.190533	0.653437	0.639077
	$|\mathcal{Z}| = 16$	0.169745	0.129582	0.232889	0.455946	0.536296	0.185364
	$|\mathcal{Z}| = 32$	0.088956	0.341674	0.407447	0.589189	0.635381	0.203955
	$|\mathcal{Z}| = 128$	0.07735	0.343908	0.514205	0.592576	0.527332	0.307418
	$|\mathcal{Z}| = 512$	0.080196	**0.062556**	**0.070378**	**0.094266**	**0.115324**	**0.136357**
	$|\mathcal{Z}| = 2048$	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
	$|\mathcal{Z}| = 2$	0.726999	0.312846	0.30105	0.376383	0.338582	0.670051
	$|\mathcal{Z}| = 8$	0.261182	0.191699	0.129651	0.219145	0.656104	0.521065
	$|\mathcal{Z}| = 16$	0.142326	0.194611	0.273543	0.50944	0.566114	0.177739
	$|\mathcal{Z}| = 32$	0.07656	0.316372	0.418504	0.572725	0.589525	0.217498
	$|\mathcal{Z}| = 128$	**0.064639**	0.380276	0.514381	0.531558	0.4364	0.284238
	$|\mathcal{Z}| = 512$	0.084125	**0.057824**	**0.063853**	**0.095734**	**0.128801**	**0.147653**
	$|\mathcal{Z}| = 2048$	0.075194	0.096743	0.128168	0.148594	0.159419	0.192654

Table 3. 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 <u>underline</u> the second smallest distance.

	WS	CB	CL	CBG
Baseline (vs real NM)	*0.001754*	*0.039509*	*0.014785*	*0.001582*
Heuristic Aug. (vs real NM-36)	0.002493	0.042682	0.01474	0.001812
$|\mathcal{Z}| = 2$	0.051189	0.081181	0.189054	0.077997
$|\mathcal{Z}| = 8$	0.055022	0.100748	0.107133	0.082222
$|\mathcal{Z}| = 16$	0.032046	0.04954	0.067667	0.053239
$|\mathcal{Z}| = 32$	0.031136	0.058868	0.059492	0.04834
$|\mathcal{Z}| = 128$	0.032589	0.059443	0.056675	0.034228
$|\mathcal{Z}| = 512$	0.03081	0.050809	0.035253	0.028172
$|\mathcal{Z}| = 2048$	<u>0.022087</u>	<u>0.043005</u>	<u>0.019479</u>	<u>0.025608</u>

Table 4. FGD values between the morphed gait to the NM variation and the real NM for FVG validation set. Baseline values corresponds to the FGD between the real unmodified gait and NM. The morphed walk is similar to the real NM variations, but the effect is not pronounced due to the same underlying viewpoint for all variations. We denote with **bold** the smallest distance and with <u>underline</u> the second smallest distance.

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[8, 45]: 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 represent a good way to augment existing walks to synthesize novel views. Since all the walks in FVG are from the same viewpoint, the differences between walking variations are not as evident. Consequently, the distance between distributions is comparatively smaller than in CASIA-B.
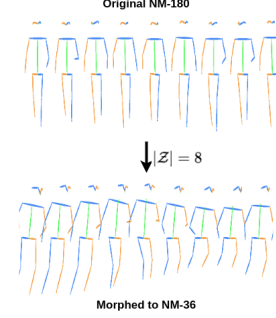

Figure 7. Failure case for $|\mathcal{Z}| = 8$ when morphing a normal walk from CASIA-B from viewpoint $180°$ to viewpoint $36°$. The latent space is not sufficiently disentangled to learn a general transport map without severely distorting the resulting gait sequence.

It is clear from results in Tables 3 and 4 that 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 7 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 manipulability of the latent codes: larger dictionary sizes have more disentangled representations which allow for more informed changes at the expense of lower data compression.

5. Conclusions

In this work, we presented GaitMorph, a novel method for modifying gait sequences into new walking variations. Our proposed approach entails firstly training a discrete latent model (in our case, a VQ-VAE) that compresses the walking sequences into a sequence of interpretable tokens, and learning an optimal latent transport map across variations. Our extensive experiments show that the trained VQ-VAE model preserves the walker's identity, achieving a marginal loss in performance when utilizing reconstructed sequences in gait recognition scenarios. Furthermore, we showed that the distribution of morphed sequences is similar to the real walk distribution. Our approach has the potential to be applied to self-supervised learning scenarios for gait recognition [8], which are heavily reliant [5, 47] on multiple strong augmentations / views for the same input.

References

[1] W. An, S. Yu, Y. Makihara, X. Wu, C. Xu, Y. Yu, R. Liao, and Y. 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] N. Bonneel, M. Van De Panne, S. Paris, and W. Heidrich. Displacement interpolation using lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, pages 1–12, 2011.

[3] Y. Cai, Y. Wang, Y. Zhu, T.-J. Cham, J. Cai, J. Yuan, J. Liu, C. Zheng, S. Yan, H. Ding, et al. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11645–11655, 2021.

[4] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng. Gaitset: Cross-view gait recognition through utilizing gait as a deep set. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[6] Y. Cheng, G. Zhang, S. Huang, Z. Wang, X. Cheng, and J. Lin. Synthesizing 3d gait data with personalized walking style and appearance. *Applied Sciences*, 13(4), 2023.

[7] A. Cosma, A. Catruna, and E. Radoi. Exploring self-supervised vision transformers for gait recognition in the wild. *Sensors*, 23(5), 2023.

[8] A. Cosma and E. Radoi. Learning gait representations with noisy multi-task learning. *Sensors*, 22(18), 2022.

[9] D. Dowson and B. Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[10] P. Esser, R. Rombach, and B. Ommer. Taming transformers for high-resolution image synthesis. in 2021 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12868–12878, 2020.

[11] C. Filipi Gonçalves dos Santos, D. d. S. Oliveira, L. A. Passos, R. Gonçalves Pires, D. Felipe Silva Santos, L. Pascotti Valem, T. P. Moreira, M. Cleison S. Santana, M. Roder, J. Paulo Papa, and D. Colombo. Gait recognition based on deep learning: A survey. *ACM Comput. Surv.*, 55(2), jan 2022.

[12] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[13] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 240–250, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[15] D. Gopinath and J. Won. fairmotion - tools to load, process and visualize motion capture data. Github, 2020.

[16] T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang, and R. Ding. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 762–770, 2022.

[17] W. Guo, Y. Du, X. Shen, V. Lepetit, X. Alameda-Pineda, and F. Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4809–4819, 2023.

[18] P. Gupta, A. Thatipelli, A. Aggarwal, S. Maheshwari, N. Trivedi, S. Das, and R. K. Sarvadevabhatla. Quo vadis, skeleton action recognition? *International Journal of Computer Vision*, 129(7):2097–2112, 2021.

[19] E. J. Harris, I.-H. Khoo, and E. Demircan. A survey of human gait-based artificial intelligence applications. *Frontiers in Robotics and AI*, 8, 2022.

[20] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[22] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

[23] S. Jana, N. Das, S. Basu, and M. 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.

[24] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] A. Łańcucki, J. Chorowski, G. Sanchez, R. Marxer, N. Chen, H. J. Dolfing, S. Khurana, T. Alumäe, and A. Laurent. Robust training of vector quantized bottleneck models. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[26] N. Li and X. Zhao. A strong and robust skeleton-based gait recognition method with gait periodicity priors. *IEEE Transactions on Multimedia*, pages 1–1, 2022.

[27] Y. Li, Z. Yu, Y. Zhu, B. Ni, G. Zhai, and W. Shen. Skeleton2humanoid: Animating simulated characters for physically-plausible motion in-betweening. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1493–1502, 2022.

[28] B. Lin, S. Zhang, M. Wang, L. Li, and X. Yu. Gaitgl: Learning discriminative global-local feature representations for gait recognition. *arXiv preprint arXiv:2208.01380*, 2022.

[29] X. Liu, Q. Wu, H. Zhou, Y. Xu, R. Qian, X. Lin, X. Zhou, W. Wu, B. Dai, and B. Zhou. Learning hierarchical cross-

modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022.

[30] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 143–152, 2020.

[31] T. Ma, Y. Nie, C. Long, Q. Zhang, and G. 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*, pages 6437–6446, 2022.

[32] A. Maiorca, Y. Yoon, and T. Dutoit. Evaluating the quality of a synthesized motion with the fréchet motion distance. In *ACM SIGGRAPH 2022 Posters*, pages 1–2. 2022.

[33] W. Mao, M. Liu, M. Salzmann, and H. Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019.

[34] M. Petrovich, M. J. Black, and G. Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021.

[35] M. Petrovich, M. J. Black, and G. Varol. Temos: Generating diverse human motions from textual descriptions. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision – ECCV 2022*, pages 480–497, Cham, 2022. Springer Nature Switzerland.

[36] S. Raab, I. Leibovitch, G. Tevet, M. Arar, A. H. Bermano, and D. Cohen-Or. Single motion diffusion. *arXiv preprint arXiv:2302.05905*, 2023.

[37] A. Razavi, A. Van den Oord, and O. Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[38] E. Ross, A. Milian, M. Ferlic, S. Reed, and A. S. Lepley. A data-driven approach to running gait assessment using inertial measurement units. *Video Journal of Sports Medicine*, 2(5):26350254221102464, 2022.

[39] S. Sanyal, A. Vorobiov, T. Bolkart, M. Loper, B. Mohler, L. S. Davis, J. Romero, and M. 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*, pages 11138–11147, 2021.

[40] L. Shi, Y. Zhang, J. Cheng, and H. 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*, pages 12026–12035, 2019.

[41] W. Shin, G. Lee, J. Lee, J. Lee, and E. Choi. Translation-equivariant image quantizer for bi-directional image-text generation. *arXiv preprint arXiv:2112.00384*, 2021.

[42] L. Siyao, W. Yu, T. Gu, C. Lin, Q. Wang, C. Qian, C. C. Loy, and Z. Liu. Bailando: 3d dance generation via actor-critic gpt with choreographic memory. In *CVPR*, 2022.

[43] L. N. Smith. No more pesky learning rate guessing games. *CoRR*, abs/1506.01186, 2015.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[45] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll. GaitGraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2314–2318, 2021.

[46] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022.

[47] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

[48] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

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

[50] J. Wang, J. Jiao, and Y.-H. Liu. Self-supervised video representation learning by pace prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 504–521. Springer, 2020.

[51] J. Wang, Y. Rong, J. Liu, S. Yan, D. Lin, and B. Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20460–20469, 2022.

[52] X. Wang, S. Zheng, R. Yang, A. Zheng, Z. Chen, J. Tang, and B. Luo. Pedestrian attribute recognition: A survey. *Pattern Recognition*, 121:108220, 2022.

[53] P. Xie, Q. Zhang, Z. Li, H. Tang, Y. Du, and X. Hu. Vector quantized diffusion model with codeunet for text-to-sign pose sequences generation. *arXiv preprint arXiv:2208.09141*, 2022.

[54] S. Yan, Z. Li, Y. Xiong, H. Yan, and D. Lin. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4394–4402, 2019.

[55] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[56] 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.

[57] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

[58] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh. Gaitgan: Invariant gait feature extraction using generative adversarial networks. In *2017 IEEE Conference on Computer Vision and*

Pattern Recognition Workshops (CVPRW), pages 532–539, 2017.

[59] S. Yu, D. Tan, and T. 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, pages 441–444. IEEE, 2006.

[60] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi. Soundstream: An end-to-end neural audio codec, 2021.

[61] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052*, 2023.

[62] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[63] Y. Zhang, M. J. Black, and S. Tang. Perpetual motion: Generating unbounded human motion. *arXiv preprint arXiv:2007.13886*, 2020.

[64] Z. Zhang, L. Tran, X. Yin, Y. Atoum, J. Wan, N. Wang, and X. Liu. Gait recognition via disentangled representation learning. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Long Beach, CA, June 2019.

[65] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[66] Z. Zhu, X. Guo, T. Yang, J. Huang, J. Deng, G. Huang, D. Du, J. Lu, and J. Zhou. Gait recognition in the wild: A benchmark. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.