CarGait: Cross-Attention based Re-ranking for Gait recognition

Gavriel Habib Noa Barzilay Or Shimshi Rami Ben-Ari Nir Darshan
OriginAI, Israel
{gavrielh, noab, ors, ramib, nir}@originai.co

Abstract

*Gait recognition is a computer vision task that identifies individuals based on their walking patterns. Its performance is commonly evaluated by ranking a gallery of candidates and measuring the identification accuracy at Rank-K. Existing models are typically single-staged, searching for the probe's nearest neighbors in a gallery, using a global feature representation. While these models can excel at retrieving the correct identity within the top-K predictions, they often struggle when hard negatives are among the top shortlist, leading to relatively low performance at the highest ranks (*e.g.*, Rank-1). In this paper, we introduce CarGait, a Re-ranking (re-ordering the top-K list) method for gait recognition, leveraging the fine-grained correlations between pairs of gait sequences, through cross-attention between gait strips. This re-ranking scheme can be adapted to existing single-stage models to enhance their final results. We demonstrate the capabilities of CarGait by extensive experiments on three common gait datasets, Gait3D, GREW, and OU-MVLP, and seven different gait models, showing consistent gains in Rank-1,5 accuracy, while outperforming existing re-ranking approaches, and a strong baseline.*

1. Introduction

Gait recognition (GR) identifies individuals based on their walking patterns, utilizing features like body shape, stride length, gait cycle dynamics, and limb movements. GR is applicable in various fields, including healthcare [30], criminal investigations [3], surveillance [33], and sport [47]. The complexity in GR arises from several factors, *e.g.* diverse camera views, occlusions, changing in clothing, or carrying of bags, that alter the shape and the observed gait dynamics.

Gait recognition is typically framed as a retrieval task, where there is a gallery of encoded gait sequences from various individuals, each associated with an identity. Given a new probe sequence, the objective is to retrieve the sequences from the gallery that match the probe, based on their gait embeddings.

The performance of gait recognition models is evaluated using Rank-K accuracy, where the gallery is ranked in ascending distance order (descending similarity) from the probe. Ideally, the probe's matching identity should be at the top of the list (high Rank-1 accuracy). In real-world applications like security and surveillance, accurately identifying an individual at Rank-1 is important for fast and precise decision-making, minimizing the need for additional verification steps. It further facilitates user experience and better demonstrates the model's robustness to varying conditions, like changes in lighting, clothing, or walking styles.

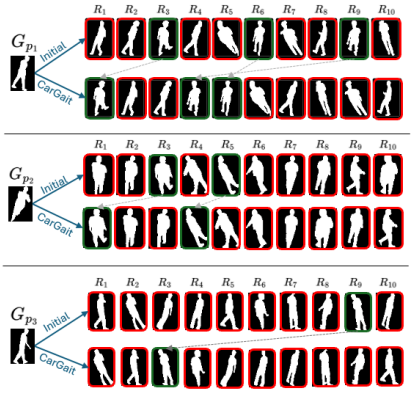

Figure 1. Examples of top-10 retrievals (from left to right) of probe sequences from the *Gait3D* [53] and *OU-MVLP* [38] datasets, before and after applying our CarGait re-ranking method. To simplify, each gait sequence is shown by a single image. Top row: without re-ranking (Initial), Bottom row: with CarGait re-ranker. Green rectangles indicate correct identity (true-positives), while the red are incorrect recognition (false-positives). CarGait improves Rank-1 and Rank-5 by initial list re-ordering.

Current gait recognition models [4, 10, 14, 15] operate in a single stage, encoding gait sequences into a single and often referred to as a *global* feature, which allows for efficient search in scale and ranking within a large dataset. In recent years, significant efforts have been made to enhance global gait representation [11, 12, 15, 19, 48], by changing architectures to transformers [11] or combining different modalities [15], yet overlooking a two-stage approach commonly employed in various domains, *e.g.* image

retrieval [36], visual place recognition [42, 50, 56], and person re-identification [2, 54, 55]. Two-stage methods incorporate a second stage, following the *initial* stage, referred as *re-ranking*. Re-ranking is the process of re-ordering the top K predictions, which are initially retrieved in a global stage from a large dataset and ranked based on their similarity to the probe. A re-ranker often leverages additional information [5, 54] and can afford higher computational costs, since it operates on a relatively short list.

While recent GR models generally perform well across various datasets [38, 53, 57], they often face challenges in achieving high Rank-1 accuracy. This can be attributed to distortions in the model's typical inputs, such as silhouettes or skeletons, as well as the presence of hard-negatives in the gallery (identities with gait patterns similar to the probe). Moreover, the model's reliance on a single global representation limits its discrimination capability. There is often a large gap in retrieving the correct identity at the first rank than within the top-5. For example, GaitPart model [10] Rank-1 accuracy on the Gait3D dataset [53] is 28.2%, while its Rank-5 is 47.6% (see more examples in Tab. 1, under "initial" column). This underscores the potential to enhance Rank-1 accuracy by effectively re-ordering the top results, even within the top-5 shortlist. This paper addresses this re-ranking problem.

Gait strips are spatio-temporal aggregated units (weakly) associated with spatial parts of the human body [10, 45]. The distance between two gait features is often computed as an average over corresponding strip distances [4, 10, 11]. This implies that strips carry gait information in a fine-grained manner. Since the top-K results at the global stage is populated with hard-negative cases, it is natural to assume that a fine-grained comparison can improve over the initial global feature based ranking.

In this paper, we present a re-ranking method for gait recognition that can be integrated with existing single-stage gait models. Building on a pre-trained gait recognition model, we propose CarGait, a Cross-Attention based Re-ranking approach designed to enhance identity recognition by capturing fine-grained correlations between gait sequences. This is achieved through cross-attention between different gait strips of two sequences, the probe and each of its top-K ranked candidates. Leveraging a metric learning approach, we map the original embedding space into a new one (of the same dimension), where probe-candidate distances are adjusted to improve Rank-1 and Rank-5 accuracy (see Fig. 1). This enhancement stems from learning fine-grained and meaningful interactions between body strips, allowing for better differentiation between subtle variations of the same identity (positives) and hard negatives appearing at the top of the ranked list.

We demonstrate the capabilities of CarGait through extensive evaluations on three common gait datasets, Gait3D, GREW, and OU-MVLP, and seven gait models, with additional baseline, showing consistent improvements in Rank-1 and 5 accuracy. Figure 5 depicts the Rank-1 improvements by CarGait.

In summary, our key contributions are as follows:

- Targeting the underexplored task of re-ranking in gait recognition, we introduce a tailored metric learning to learn new discriminative embeddings.
- Proposing a novel re-ranking approach that leverages pairwise probe-candidate gait-strip correlations to learn conditioned representations, enabling a fine-grained pairwise similarity refinement.
- CarGait consistently outperforms state-of-the-art re-ranking methods in person re-identification, and a strong baseline, across diverse models and challenging datasets.

2. Related Work

Gait recognition models are commonly classified into two categories: model-based and appearance-based approaches. Model-based approaches [16, 18, 23, 40, 41, 49] try to recognize walking patterns using the estimated structure of the human body, such as 2D or 3D pose, while appearance-based approaches [4, 10, 14] extract gait features directly from RGB images or binary silhouette sequences. Although model-based approaches are theoretically robust to changes in clothing and carrying objects, they tend to under-perform appearance-based approaches on in-the-wild benchmarks, *e.g.* Gait3D [53] and GREW [57]. This disparity is likely due to the challenges in estimating body parameters in low-resolution videos. In this work, we focus on appearance-based approaches due to their demonstrated superiority.

2.1. CNNs and Transformers

Although transformers have demonstrated superior performance in numerous computer vision tasks, the majority of common gait recognition models are still based on CNNs [4, 10, 14, 15, 24]. However, some recent methods leverage the capabilities of Transformers for gait recognition [11, 29, 31]. Our method is compatible with both architectures and can be learned on top of each. We specifically demonstrate this capability on the following CNN [4, 10, 14, 15, 24] and Transformer architectures [11]. The self-attention mechanism [8, 43] is widely used in gait recognition [11, 29, 31] to emphasize key spatial areas or time slots within a single sequence. In contrast, we have created a cross-attention component specifically designed to learn the relationships between two distinct gait sequences.

2.2. Cross-Attention

Cross attention has been used in various applications [1, 22, 25]. Specifically, in gait recognition, Cui *et al.* [6] used cross-attention to combine features of a gait sequence from different modalities, such as silhouettes and skeletons for

global single-stage ranking. In our study, we introduce a cross-attention method specifically designed for re-ranking.

While previous models utilized the corresponding gait strips to optimize the feature space to rank the entire gallery directly [4, 10, 11, 14, 15, 24], CarGait takes a different approach. It maps the global feature maps from a pre-trained single-stage model into a new feature space, by applying cross-attention across *all strips* of the probe and its top candidates. This transformation produces refined feature representations that enhance the re-ranking process.

2.3. Re-ranking

Re-ranking is the process of refining or re-ordering an initial list of results to enhance the accuracy and relevance of the final ranked list. It is widely used in various applications, such as text retrieval [32, 37] and different applications of image retrieval [36, 39, 56]. However, these methods are often not designed to address spatio-temporal "instance" matching, as in gait recognition. In this context, Gordo *et al*. [17] suggested the "query expansion method", which was originally designed for one-stage image retrieval and was later used by [34] for re-ranking. Our method is different as it conducts a pairwise interaction and impacts not only the probe representation but the candidate as well.

A closer domain to gait recognition in re-ranking is person re-identification (reID), which focuses on matching images of the same individual across different viewpoints using visual cues like clothing. Notable approaches include k-reciprocal encoding (KR) [54], ECN [35], and a specialized feature-learning method [51] that leverages the relative structure of the top-ranked list in the feature space. In this context, Bai *et al*. [2] introduced a re-ranking approach for object retrieval and reID by incorporating metric fusion and capturing the geometric structure of multiple data manifolds. LBR [27] formulates data as a graph to optimize group-wise similarities, while GCR [52] employs a graph-based strategy that refines the global feature representations by aligning them across similar samples. Instead, CarGait introduces a *fine-grained* and *pairwise* comparison between the probe and each individual candidate. This distinct approach is better suited for gait recognition and performs better than common re-rankers, particularly in scenarios where positives are rare in the gallery [53, 57], even when there is only a single positive sample (as demonstrated in Tab. 3).

Unlike re-ranking methods that often focus on refining the similarity matrix of the top-K samples [27, 35, 54], CarGait modifies the gait representations at the re-ranking stage, using detailed comparisons through cross-attention.

Despite their potential to boost accuracy, re-ranking methods are rarely used in gait recognition. In this regard, Chen *et al*. [5] proposed a re-ranking method based on *engineered features* such as Gait Energy Image (GEI) or Active Energy Image (AEI). They enhance robustness against external factors like clothing changes, view angle, and walking speed by transferring the GEI into a new feature space through sparse coding. In contrast, CarGait leverages a *deep learning* architecture that integrates feature learning into the re-ranking process. Our approach employs *cross-attention* to capture fine-grained interactions between pairs of probe and candidate feature maps, refining the ranking process.

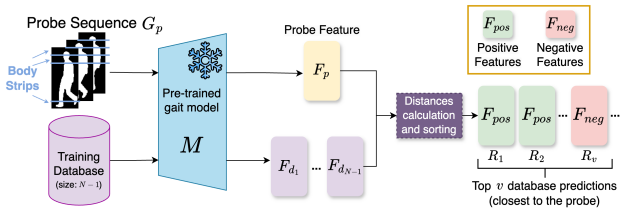

Figure 2. **Train set generation:** a pre-trained gait model M is used as a feature extractor. A training dataset of gait probe feature maps F_p, with their nearest v candidates, is constructed. Within this top-v list, some features may be positives (*i.e*., sharing the same identity as the probe, shown in green), while others may be negatives (shown in red).

In summary, most existing re-ranking methods are designed for image retrieval or person re-identification, focusing on refining rankings based on the spatial or relative arrangement of samples in the initial embedding space. In contrast, CarGait generates a new embedding space specifically tailored for gait spatio-temporal re-ranking. To highlight the advantages of CarGait, we conduct extensive comparisons with various approaches from these domains, KR [54], LBR [27], and GCR [52], demonstrating its effectiveness in re-ranking for gait recognition.

3. Method

In this section, we introduce CarGait, a novel re-ranking method for gait recognition. Figure 3 presents an overview of CarGait with its training phase as well as the inference scheme. Our re-ranking method is based on cross-attention between the probe and each candidate in its top-K global ranking results. Let us start by denoting a pair of gait sequences, G_p and G_c, as the *probe* and *candidate* sequence respectively, each consisting of a set of silhouette frames and optionally also skeletons. Figure 2 illustrates the process of generating the training data for the re-ranker.

We now define the gait feature maps derived from G_p and G_c after being processed by a pre-trained model M, yielding $M(G_p, G_c) = \{F_p, F_c\}$. Here, $F_p, F_c \in \mathbb{R}^{s \times d}$ are the extracted feature maps, where s represents the number of horizontal body *strips*, and d denotes the feature dimension. These feature maps are obtained after temporal aggregation within the backbone.

Next, we compute multi-head cross-attention between F_p and F_c (see Fig. 3, top-view), where F_p serves as the

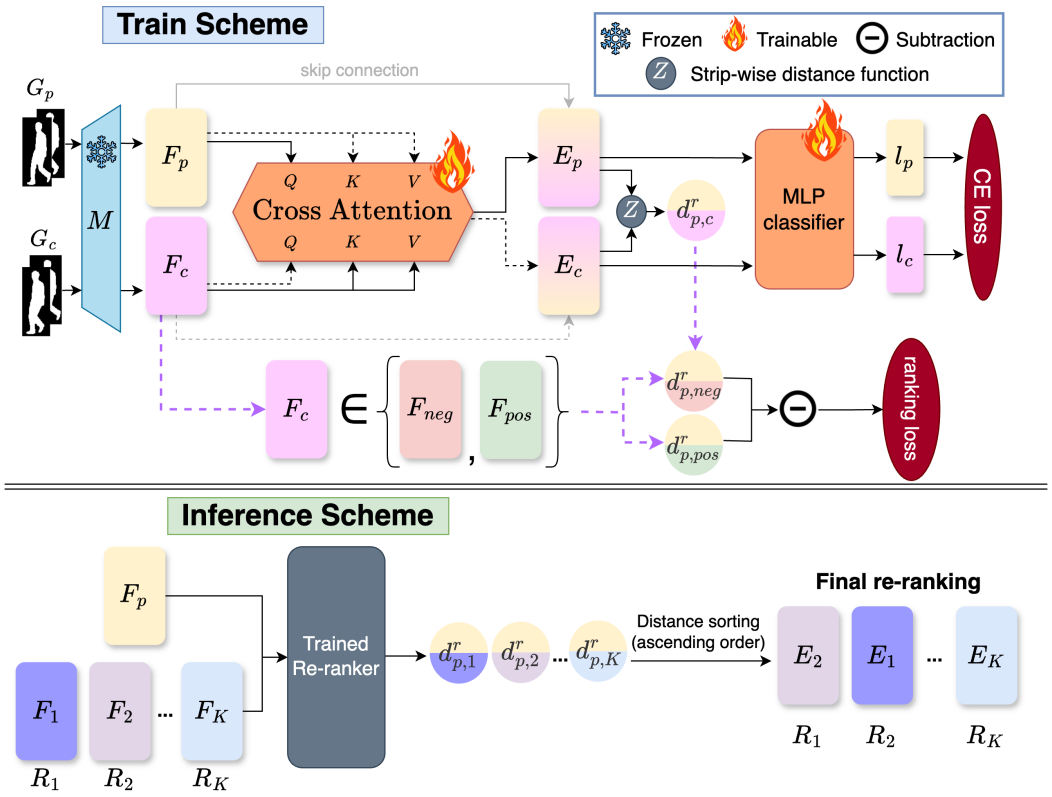

Figure 3. An overview of CarGait method. **Train Scheme:** a strip-wise multi-head cross-attention based re-ranker is trained with ranking and cross-entropy losses, learning the part relations between pairs of gait sequences. Practically, the cross-attention module is applied twice for each probe-candidate pair (illustrated by the solid and dashed lines). **Inference Scheme:** re-ranking is achieved by sorting the probe's top K candidate predictions in ascending order by their new distances to the probe, as determined by the trained re-ranker with $d^r_{p,c}$.

Query and F_c as both the Key and Value (solid lines), producing a new feature map $E_p \in \mathbb{R}^{s \times d}$. Similarly, we obtain $E_c \in \mathbb{R}^{s \times d}$ by performing the reverse cross-attention, using F_c as the Query and F_p as the Key and Value (dashed lines).

To preserve information from the pre-trained model, we incorporate a residual (skip) connection between F_p and E_p, as well as between F_c and E_c. This ensures that the re-ranker is initialized with the pre-trained feature space while refining the representations. In practice, the cross-attention module takes two *distinct* feature maps as input and generates two *conditioned* representations. Each strip in E_p is now influenced by its attention relationships with all the strips in the candidate.

We now derive a new metric space, with modified probe and candidate representations, for re-ranking. The distance is computed as the average euclidean distance between corresponding strip features *after* cross-attention, namely, $d^r_{p,c} = \mathcal{Z}(E_p, E_c)$, where the function $\mathcal{Z}$ from [4, 10, 11] maps two given representations to a distance. We train our re-ranker with two loss objectives, ranking and classification (see Eq. (2)). For the classification loss, we add a trainable MLP, applied on top of E_p and E_c (see Fig. 3 top-view). The classification loss functions as a regularization term, preserving identity information within the learned representations and contributing to improved performance.

Figure 4 illustrates the impact of our cross-attention module. We compute the cosine similarity matrix between strips of a probe and a positive candidate, *i.e* from the same identity. For comparison, we show the strip correlations from the initial global features F_p, F_c, namely $CS(F)$ vs. $CS(E)$, based on E_p and E_c strips after our learnable cross-attention module for re-ranking, indicating CarGait. In both cases, matched body strips have higher similarity (diagonals), while in CarGait new interactions are learned between different body strips with the cross-attention module (brighter colors on off-diagonal blocks).

We train our re-ranker for each global model M, keeping M's layers frozen. First, we generate a dataset $\mathcal{D}$ comprising the feature map representations (F_p, F_c) of all training samples along with their top-v closest predictions, as illustrated in Fig. 2. Then we train CarGait modules (see Fig. 3) with an objective to improve the ranked list.

Loss: We optimize our re-ranking module using two loss

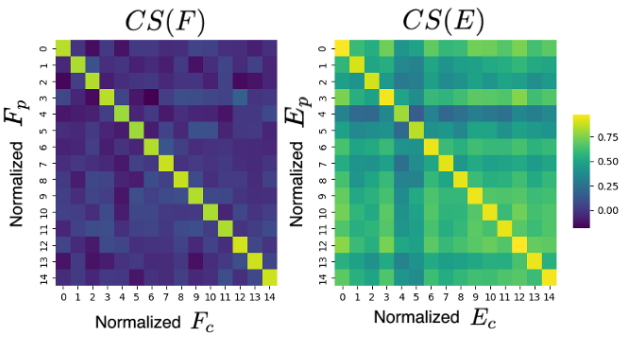

Figure 4. Cosine similarities (CS) between L2-normalized features of s body strips (here, $s = 15$) in two gait sequences of the same identity, in the initial global space F and the re-ranker space E. Each row/column represents a strip. In $CS(F)$, the features are distinct, indicating low correlation between different body strips (blue colors off-diagonal). In the re-ranker space $CS(E)$, that leverages cross-attention, higher correlation between strips are observed (green/yellow colors), indicating cross-strip interactions learned through the cross-attention. Samples are from the *Gait3D* dataset [53].

objectives: ranking loss and classification loss. The ranking loss is a metric-based loss that penalizes the model whenever a negative sample is positioned closer to the probe than a positive one, in Euclidean space. The penalty is proportional to the relative distance difference between the probe-negative and probe-positive pairs. The classification loss can be viewed as a regularization term ensuring that identity information is effectively preserved within the learned representations. This helps maintain discriminative features while refining the ranking process.

Considering a triplet of gait feature maps for probe (p) and corresponding positive (pos - same identity) and negative (neg) samples, denoted by $\{F_{p_i}, F_{pos_i}, F_{neg_i}\}$ sampled from $\mathcal{D}$, the ranking loss is given by:

$$\begin{aligned} \mathcal{L}_i^* &= -log[\sigma(d^r_{p_i,neg_i} - d^r_{p_i,pos_i})] \\ \mathcal{L}_i &= \begin{cases} \beta\mathcal{L}_i^* & \text{if } d^r_{p_i,neg_i} \geq d^r_{p_i,pos_i} \\ \mathcal{L}_i^* & \text{otherwise} \end{cases} \\ \mathcal{L}_{ranking} &= \sum_i \mathcal{L}_i \end{aligned} \tag{1}$$

where σ is a sigmoid function and $d^r_{\cdot,\cdot}$ indicates distance in the new re-ranking space. For sake of effectiveness, during CarGait training, we downscale the loss of the triplets that are correctly ranked (*i.e.* $d^r_{p_i,neg_i} \geq d^r_{p_i,pos_i}$) by a scale factor $\beta < 1$.

The classification loss ($\mathcal{L}_{CE}$) is a standard multi-class cross-entropy loss [7], applied to all training samples. To compute this loss, the attended representation E_i is fed into an MLP classifier, producing a logits vector l_i with a size matching the number of training classes C. Eventually, the losses are linearly combined:

$$\mathcal{L} = \mathcal{L}_{ranking} + \alpha\mathcal{L}_{CE} \tag{2}$$

where α is a standard weighting hyper-parameter.

Train Stopping Criteria: We use $\mathcal{D}_{val}$, a dataset constructed in the same manner as $\mathcal{D}$ but derived from the validation set, for stopping criteria. Specifically, the ranking loss described in Eq. (1) is calculated each T_{val} iterations during the training, and eventually the checkpoint with the minimum loss is chosen.

Inference: For a given probe sequence G_p, the top-K nearest sequences are first retrieved using a given pre-trained model. The re-ranker is then applied to all pairs (F_p, F_c) within the top-K list, to compute the updated distances ($d^r_{p,1}$, ..., $d^r_{p,K}$). The re-ranking is then achieved by re-ordering the top-K gallery sequences based on their new distances to the probe sequence, in ascending order.

4. Evaluation

4.1. Gait recognition models

We evaluate CarGait on six silhouette based gait recognition models with a wide range of performance levels: SwinGait-3D, DeepGaitV2-P3D [11], GaitBase [14], GaitSet [4], GaitGL [24], and GaitPart [10], and on a combined silhouette-skeleton SoTA model of SkeletonGait++ [15]. These models are then tested across three different benchmarks[1]. For each model, we train a re-ranker, adjusting the input size according to the feature map dimensions obtained from the pre-trained model. All other hyper-parameters, such as the number of attention heads, remain unchanged.

`SwinGait-3D` [11] combines convolutional layers followed by transformer blocks on shifted windows.

`DeepGaitV2-P3D (DGV2-P3D)` [11] and `GaitBase` [14] are based on deep ResNet, aiming to generalize better for in-the-wild scenarios.

`GaitSet` [4] is trained on a set of silhouettes, `GaitPart` [10] emphasizes the importance of body parts, and `GaitGL` [24] fuses global and local information. All three are CNN-based.

`SkeletonGait++ (SG++)` [15] combines silhouette and skeleton features in a multi-branch architecture.

4.2. Datasets

We evaluate CarGait on three well-known gait datasets, *Gait3D* [53], *GREW* [57], and *OU-MVLP* [38]. Our experiments strictly adhere to the official evaluation protocols.

Gait3D [53] contains 25,309 sequences of 4,000 identities recorded from 39 cameras in a supermarket. The training set includes 3,000 subjects, while the test set includes the remaining 1,000 subjects.

[1]For comparison, we chose the models that offer publicly available code and checkpoints.

GREW [57] is a large-scale dataset containing 128,671 sequences of 26,345 subjects captured from 882 cameras in the wild. The training set contains 20,000 subjects, while the test set comprises 6,000 subjects.

OU-MVLP [38] is a large indoor dataset captured in a controlled environment from 14 viewing angles. It includes 10,307 subjects, with 5,153 used for training and the remaining 5,154 used for testing.

We use the `typewriter` font for model names and *italic* font for dataset names for better distinction.

4.3. Implementation details

The implementations of the gait models and checkpoints for academic datasets are publicly available in the OpenGait codebase [13]. In each dataset, the input is a silhouette sequence where the silhouettes are resized to 64×44. In SkeletonGait++ model [15], additional skeleton-based information of size 2×64×44 is added. The training set is divided into separate training and validation sets by designating the last 10% training identities as the validation set.

We adopt the AdamW optimizer [26] with a learning rate of 1e-5 and a weight decay of 1e-2. In each training iteration, a batch size of 32×4 was used[2]. The re-ranker was trained using the loss function described in Eq. (2) with $\alpha = 0.01$, and with $\beta = 0.1$ in the ranking loss (see Eq. (1)). The multi-head cross-attention module was designed as a single block with 8 attention heads and with a hidden dimension of $256D$. For creating the dataset $\mathcal{D}$, the top $v = 30$ predictions were considered. The validation loss was calculated every T_{val} = 10,000 training iterations, and the total number of iterations was set to 100,000. During inference, the re-ranking process was applied to the top $K = 10$ predictions of each probe sequence. Our experiments were conducted on four NVIDIA A100 GPUs for training, each with 40 GB of memory. The average training time per experiment was 16 hours, with inference time of approximately 6.5 milliseconds per probe on a single GPU. Further runtime analysis is shown in the appendix.

4.4. Results

We show the re-ranker results applied on top of various gait recognition models for multiple datasets in Tab. 1. We compare the results obtained by CarGait applied on the top-10 list, with those of the original pre-trained models before re-ranking (referred as *initial* results). A spider-plot visualization of Rank-1 results is shown in Fig. 5, along with an illustration of the re-ranking depicted in Fig. 1. CarGait demonstrates consistent enhancements in Rank-1 accuracy across various models and datasets, highlighting the effectiveness of our approach in improving global single-stage models in gait recognition. For example, in `GaitBase` on *GREW* dataset, the initial Rank-1 of 60.1% is improved to 67.2%. Note that since our re-ranker is designed to re-arrange a given list of top-10 results, it also improves Rank-5 accuracy. For instance, in `SwinGait-3D` on *Gait3D* dataset, the initial Rank-5 of 86.7% is improved to 88.6%.

[2]For DeepGaitV2-P3D the batch size was reduced to 16×4, and for SwinGait-3D and SkeletonGait++ to 8×4, due to memory issues.

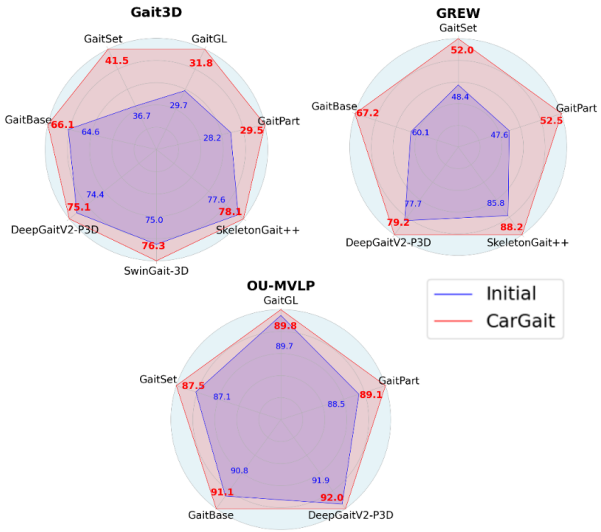

Figure 5. Comparison of Rank-1 accuracy of different models before (in blue) and after CarGait re-ranker (in red), across three different benchmarks.

Notably, CarGait shows greater improvements when applied to *Gait3D* and *GREW* datasets, compared to *OU-MVLP*. We suggest that these differences are due to the initial results achieved on each dataset. While *Gait3D* and *GREW* are challenging datasets captured in-the-wild, *OU-MVLP* is an indoor dataset collected in a controlled environment. As a result, the Rank-1 performances on *OU-MVLP* appears to be saturated even before applying CarGait, leaving less potential for improvement.

Next, we compare CarGait to three existing re-rankers [27, 52, 54] across different benchmarks, as presented in Tab. 2, 3 (and in Tab. 1 in the appendix). CarGait consistently outperforms other re-rankers [27, 52, 54] across all benchmarks. We attribute this to its modeling, addressing gait recognition and by capturing probe-candidate internal relationships through strip-wise cross-attention, rather than merely re-ordering the top-K gallery predictions based on global feature similarities. In some scenarios, particularly when positives are rare in the gallery, other methods may even degrade the initial single-stage results, whereas CarGait continues to enhance them (see Tab. 3).

Table 4 presents the runtime analysis. While CarGait adds some complexity and requires training, it is significantly faster than other re-ranking methods at inference.

Method	Publication	*Gait3D* R1 Initial	*Gait3D* R1 CG	*Gait3D* R5 Initial	*Gait3D* R5 CG	*Gait3D* mAP Initial	*Gait3D* mAP CG	*GREW* R1 Initial	*GREW* R1 CG	*GREW* R5 Initial	*GREW* R5 CG	*OU-MVLP* R1 Initial	*OU-MVLP* R1 CG
GLN [20]	ECCV 20	31.4	*	52.9	*	24.74	*	-	-	-	-	89.2	*
CSTL [21]	ICCV 21	11.7	*	19.2	*	5.59	*	-	-	-	-	90.2	*
GaitGCI [9]	CVPR 23	50.3	*	68.5	*	39.50	*	68.5	*	80.8	*	92.1	*
DANet [28]	CVPR 23	48.0	*	69.7	*	-	*	-	-	-	-	90.7	*
HSTL [44]	ICCV 23	61.3	*	76.3	*	55.48	*	62.7	*	76.6	*	92.4	-
DyGait [46]	ICCV 23	66.3	*	80.8	*	56.40	*	71.4	*	83.2	*	-	-
VPNet [29]	CVPR 24	75.4	*	87.1	*	-	*	80.0	*	89.4	*	92.4	*
GaitPart [10]	CVPR 20	28.2	**29.5**	47.6	**48.5**	21.58	**22.73**	47.6	**52.5**	60.7	**67.5**	88.5	**89.1**
GaitGL [24]	ICCV 21	29.7	**31.8**	48.5	**51.0**	22.29	**23.55**	47.3	*	63.6	*	89.7	**89.8**
GaitSet [4]	AAAI 19	36.7	**41.5**	58.3	**62.1**	30.01	**32.97**	48.4	**52.0**	63.6	**68.0**	87.1	**87.5**
GaitBase [14]	CVPR 23	64.6	**66.1**	81.5	**82.8**	55.29	**57.66**	60.1	**67.2**	75.5	**78.5**	90.8	**91.1**
DGV2-P3D [11]	ArXiv 23	74.4	**75.1**	**88.0**	87.5	65.76	**66.89**	77.7	**79.2**	87.9	**88.7**	91.9	**92.0**
SwinGait3D [11]	ArXiv 23	75.0	**76.3**	86.7	**88.6**	66.69	**67.59**	79.3	*	88.9	*	-	-
SG++ [15]	AAAI 24	77.6	**78.1**	89.4	**90.4**	70.30	**70.86**	85.8	**88.2**	92.6	**94.6**	-	-

Table 1. Improvement in Rank-K accuracy [%] and mAP with CarGait re-ranker, over different methods and multiple *datasets*. Initial - indicates the global result w/o re-ranking, while CG indicates results after applying CarGait. CarGait was applied on top of various gait recognition models, with wide range of performance levels and with publicly available code and checkpoints. Best results are in bold. (*) denotes unavailable code or checkpoint. (-) indicates no result for the corresponding setting.

Method	Publication	R1 KR	R1 LBR	R1 GCR	R1 CG	R5 KR	R5 LBR	R5 GCR	R5 CG	mAP KR	mAP LBR	mAP GCR	mAP CG
GaitPart [10]	CVPR 20	26.5	23.3	26.0	**29.5**	42.7	47.1	45.7	**48.5**	21.50	18.24	21.68	**22.73**
GaitGL [24]	ICCV 21	26.0	27.3	22.4	**31.8**	42.4	48.0	41.6	**51.0**	20.83	18.69	18.10	**23.55**
GaitSet [4]	AAAI 19	34.8	33.0	35.7	**41.5**	53.1	60.4	56.4	**62.1**	30.26	26.92	30.53	**32.97**
GaitBase [14]	CVPR 23	60.0	63.8	63.1	**66.1**	77.6	82.7	79.2	**82.8**	**57.78**	51.43	53.12	57.66
DGV2-P3D [11]	ArXiv 23	65.8	58.7	74.2	**75.1**	83.5	86.6	87.0	**87.5**	65.71	54.48	65.28	**66.89**
SwinGait3D [11]	ArXiv 23	66.7	64.0	74.1	**76.3**	83.6	88.1	86.3	**88.6**	66.79	57.79	64.03	**67.59**
SG++ [15]	AAAI 24	69.7	61.7	76.1	**78.1**	85.6	90.2	89.6	**90.4**	70.30	58.99	68.72	**70.86**

Table 2. Rank-K accuracy [%] and mAP on *Gait3D* dataset [53] for different re-ranking methods: k-reciprocal (KR) [54], LBR [27], and GCR [52], compared to CarGait (CG). Best results are in bold.

Method	Initial	KR	LBR	CG
GaitPart [10]	88.5	68.4	80.6	**89.1**
GaitGL [24]	89.7	71.9	89.2	**89.8**
GaitSet [4]	87.1	65.5	70.3	**87.5**
GaitBase [14]	90.8	72.9	74.4	**91.1**
DGV2-P3D [11]	91.9	76.4	77.3	**92.0**

Table 3. Rank-1 accuracy [%] on *OU-MVLP* dataset [38] for different re-rankers: k-reciprocal (KR) [54] and LBR [27], compared to CarGait (CG). Initial - indicates the global result w/o re-ranking. This dataset is distinguished by having only a single positive sample in the gallery, making re-ranking particularly challenging. Best results are in bold. GCR [52] is excluded due to excessive runtime.

Method	Publication	Inference time [msec]
KR [54]	CVPR 17	214
LBR [27]	ICCV 19	19.81
GCR [52]	TMM 23	1866
CarGait	-	**6.52**

Table 4. Inference time per probe for various re-ranking methods, evaluated on the top-10 list. Best result is in bold.

4.5. Ablation Study and Analysis

To demonstrate the efficacy of the proposed method and assess the impact of various key components, we conducted an extensive ablation study using SwinGait-3D model trained on *Gait3D* dataset. The results are summarized in Tab. 5, with further analysis provided in the appendix.

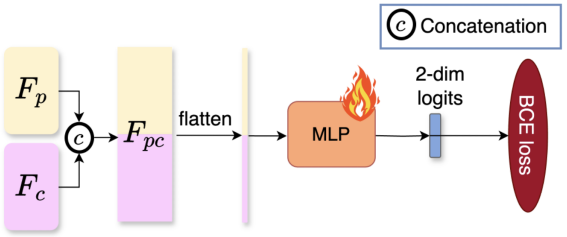

Figure 6. An overview of the baseline binary classifier architecture presented in row-3 in Tab. 5. In this setup, there is no cross-attention module. Instead, the features F_p, F_c are concatenated into F_{pc} which is flattened and fed into an MLP to predict if it represents a positive or a negative pair (same or different identities).

While our primary training objective is to optimize Rank-1 (R1), we also examine the impact on Rank-5 (R5).

#	Method	R1	R5	mAP
	Architectural and loss components			
1	Pre-trained model w/o re-ranker	75.0	86.7	66.69
2	Pre-trained model w/ extra training	75.0	86.9	66.87
3	Binary classification (Baseline)	71.6	87.8	65.47
4	w/o classification loss ($\alpha = 0$)	76.0	89.1	67.73
5	w/o loss damping ($\beta = 1$)	75.5	88.2	67.27
	Hyperparameters			
6	Inference re-ranking factor $K = 5$	76.5	86.7	67.30
7	Inference re-ranking factor $K = 20$	76.3	88.6	67.68
8	Dataset creation with $v = 20$	76.3	88.3	67.45
9	Dataset creation with $v = 40$	76.0	89.0	67.66
10	**CarGait**	76.3	88.6	67.59

Table 5. Ablation study with `SwinGait-3D` model trained on *Gait3D* dataset. Rank-1 (R1), Rank-5 (R5), and mAP are reported.

Baseline. To show that additional training alone is insufficient to enhance performance, we continue training the global model for the same number of iterations used for CarGait. The result in Tab. 5 (2nd row) suggests that the initial model (1st row) has already reached its full potential.

To further highlight the advantages of our strip-wise cross-attention module, we also evaluate a naive baseline. In this baseline, the re-ranker is treated as a classifier trained on the top-v global feature predictions, as illustrated in Fig. 2. To this end, we trained a binary classifier to identify whether a pair of probe and candidate sequences share the same identity (positives) or not (negatives). Practically, we trained a re-ranker without the cross-attention module, as illustrated in Fig. 6. In this setup, we concatenated the two gait global features F_p, F_c to F_{pc}, feeding it into an MLP that is further trained with BCE loss. At inference, the top K pre-trained model predictions of a probe sequence are sorted in descending order based on their positive class scores. The result in Tab. 5 (3rd row) shows a drop in Rank-1 (R1) performance compared to the initial reference state (1st row) for this baseline experiment. This decline can be attributed to naive interaction modeling and late fusion between the probe and candidates, whereas CarGait effectively learns internal part (strip) relationships, leading to improved recognition.

Loss Components. We conduct an ablation study on the classification loss to evaluate its impact during training. For this purpose, we train our model also without $\mathcal{L}_{CE}$ ($\alpha = 0$ in Eq. (2)). The result in row-4 confirms that incorporating the classification loss leads to additional improvements. Note that our optimization is focused on R1 and some components might negatively impact R5 or mAP. Then, in row-5, we present an experiment to evaluate the impact of our damping parameter β in the loss function (see Eq. (1)).

Hyperparameters. In rows 6-7, we evaluate the re-ranking performance in relation to K, the length of the top-ranked list on which re-ranking is performed during inference. The result in row 6 shows a slightly better R1, but CarGait with $K = 10$ significantly enhances Rank-5 accuracy. While the mAP for $K = 20$ (row 7) is slightly better, we chose $K = 10$ as a fixed value across all experiments as a compromise between performance and runtime (see runtime analysis in the appendix). Finally, in rows 8-9, we examine the impact of the candidate set size during training (top-v). To this end, we create the training set with different values of v (20 and 40, instead of 30). Although different values of v could be optimized for each model and dataset, we fix $v = 30$ across all experiments to ensure better generalization of our method. In the supplementary materials, we present a further analysis on K and v values.

5. Conclusion

In this paper, we introduce CarGait, Cross-Attention based Re-ranker for gait recognition. CarGait operates on pairs of probe and top-K candidate feature maps, obtained from a pre-trained single-stage model. It introduces a new approach to generate *conditioned* representations for each probe-candidate pair. To achieve this, we suggest a cross-attention module that captures fine-grained correlations between pairwise probe-candidate body strips, through their spatio-temporal feature maps. These new representations are then used to compute updated distances, allowing CarGait to re-rank the candidate list and enhance performance at both Rank-1 and Rank-5. We evaluate CarGait with multiple models and across various datasets, demonstrating consistent performance improvements after re-ranking, and superior results over existing re-rankers. We hope this study will encourage the exploration of re-ranking methods for gait recognition, a field that has received very limited attention in the past.

References

[1] Dasom Ahn, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. Star-transformer: a spatio-temporal cross attention transformer for human action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3330–3339, 2023. 2

[2] Song Bai, Peng Tang, Philip HS Torr, and Longin Jan Latecki. Re-ranking via metric fusion for object retrieval and person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 740–749, 2019. 2, 3

[3] Imed Bouchrika, Michaela Goffredo, John Carter, and Mark Nixon. On using gait in forensic biometrics. *Journal of forensic sciences*, 56(4):882–889, 2011. 1

[4] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8126–8133, 2019. 1, 2, 3, 4, 5, 7

[5] Xin Chen and Jiaming Xu. Uncooperative gait recognition: Re-ranking based on sparse coding and multi-view hypergraph learning. *Pattern Recognition*, 53:116–129, 2016. 2, 3

[6] Yufeng Cui and Yimei Kang. Multi-modal gait recognition via effective spatial-temporal feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17957, 2023. 2

[7] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67, 2005. 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[9] Huanzhang Dou, Pengyi Zhang, Wei Su, Yunlong Yu, Yining Lin, and Xi Li. Gaitgci: Generative counterfactual intervention for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5578–5588, 2023. 7

[10] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. 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. 1, 2, 3, 4, 5, 7

[11] Chao Fan, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Exploring deep models for practical gait recognition, 2023. 1, 2, 3, 4, 5, 7

[12] Chao Fan, Saihui Hou, Jilong Wang, Yongzhen Huang, and Shiqi Yu. Learning gait representation from massive unlabelled walking videos: A benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[13] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9707–9716, 2023. 6

[14] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9707–9716, 2023. 1, 2, 3, 5, 7

[15] Chao Fan, Jingzhe Ma, Dongyang Jin, Chuanfu Shen, and Shiqi Yu. Skeletongait: Gait recognition using skeleton maps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1662–1669, 2024. 1, 2, 3, 5, 6, 7

[16] Yang Fu, Shibei Meng, Saihui Hou, Xuecai Hu, and Yongzhen Huang. Gpgait: Generalized pose-based gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19595–19604, 2023. 2

[17] Albert Gordo, Filip Radenovic, and Tamara Berg. Attention-based query expansion learning. In *European Conference on Computer Vision*, pages 172–188. Springer, 2020. 3

[18] Hongji Guo and Qiang Ji. Physics-augmented autoencoder for 3d skeleton-based gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19627–19638, 2023. 2

[19] Gavriel Habib, Noa Barzilay, Or Shimshi, Rami Ben-Ari, and Nir Darshan. Watch where you head: A view-biased domain gap in gait recognition and unsupervised adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6109–6119, 2024. 1

[20] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European conference on computer vision*, pages 382–398. Springer, 2020. 7

[21] Xiaohu Huang, Duowang Zhu, Hao Wang, Xinggang Wang, Bo Yang, Botao He, Wenyu Liu, and Bin Feng. Context-sensitive temporal feature learning for gait recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12909–12918, 2021. 7

[22] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2991–2999, 2024. 2

[23] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020. 2

[24] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14648–14656, 2021. 2, 3, 5, 7

[25] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8599–8608, 2024. 2

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[27] Chuanchen Luo, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Spectral feature transformation for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4976–4985, 2019. 3, 6, 7

[28] Kang Ma, Ying Fu, Dezhi Zheng, Chunshui Cao, Xuecai Hu, and Yongzhen Huang. Dynamic aggregated network for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22076–22085, 2023. 7

[29] Kang Ma, Ying Fu, Chunshui Cao, Saihui Hou, Yongzhen Huang, and Dezhi Zheng. Learning visual prompt for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 593–603, 2024. 2, 7

[30] Sumit Majumder, Tapas Mondal, and M Jamal Deen. A simple, low-cost and efficient gait analyzer for wearable healthcare applications. *IEEE Sensors Journal*, 19(6):2320–2329, 2018. 1

[31] Jashila Nair Mogan, Chin Poo Lee, Kian Ming Lim, and Kalaiarasi Sonai Muthu. Gait-vit: Gait recognition with vision transformer. *Sensors*, 22(19):7362, 2022. 2

[32] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019. 3

[33] Anubha Parashar, Apoorva Parashar, Andrea F Abate, Rajveer Singh Shekhawat, and Imad Rida. Real-time gait biometrics for surveillance applications: A review. *Image and Vision Computing*, page 104784, 2023. 1

[34] Leigang Qu, Meng Liu, Wenjie Wang, Zhedong Zheng, Liqiang Nie, and Tat-Seng Chua. Learnable pillar-based re-ranking for image-text retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1252–1261, 2023. 3

[35] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 420–429, 2018. 3

[36] Shihao Shao, Kaifeng Chen, Arjun Karpur, Qinghua Cui, André Araujo, and Bingyi Cao. Global features are all you need for image retrieval and reranking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11036–11046, 2023. 2, 3

[37] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is chatgpt good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542*, 2023. 3

[38] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ transactions on Computer Vision and Applications*, 10:1–14, 2018. 1, 2, 5, 6, 7

[39] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 12105–12115, 2021. 3

[40] 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)*, pages 2314–2318. IEEE, 2021. 2

[41] Torben Teepe, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Towards a deeper understanding of skeleton-based gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1569–1577, 2022. 2

[42] Issar Tzachor, Boaz Lerner, Matan Levy, Michael Green, Tal Berkovitz Shalev, Gavriel Habib, Dvir Samuel, Noam Korngut Zailer, Or Shimshi, Nir Darshan, et al. Effovpr: Effective foundation model utilization for visual place recognition. *arXiv preprint arXiv:2405.18065*, 2024. 2

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[44] Lei Wang, Bo Liu, Fangfang Liang, and Bincheng Wang. Hierarchical spatio-temporal representation learning for gait recognition. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19582–19592. IEEE, 2023. 7

[45] Ming Wang, Beibei Lin, Xianda Guo, Lincheng Li, Zheng Zhu, Jiande Sun, Shunli Zhang, Yu Liu, and Xin Yu. Gaitstrip: Gait recognition via effective strip-based feature representations and multi-level framework. In *Proceedings of the Asian conference on computer vision*, pages 536–551, 2022. 2

[46] Ming Wang, Xianda Guo, Beibei Lin, Tian Yang, Zheng Zhu, Lincheng Li, Shunli Zhang, and Xin Yu. Dygait: Exploiting dynamic representations for high-performance gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13424–13433, 2023. 7

[47] Datao Xu, Huiyu Zhou, Wenjing Quan, Xinyan Jiang, Minjun Liang, Shudong Li, Ukadike Chris Ugbolue, Julien S Baker, Fekete Gusztav, Xin Ma, et al. A new method proposed for realizing human gait pattern recognition: Inspirations for the application of sports and clinical gait analysis. *Gait & Posture*, 107:293–305, 2024. 1

[48] Dingqiang Ye, Chao Fan, Jingzhe Ma, Xiaoming Liu, and Shiqi Yu. Biggait: Learning gait representation you want by large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 200–210, 2024. 1

[49] Cun Zhang, Xing-Peng Chen, Guo-Qiang Han, and Xiang-Jie Liu. Spatial transformer network on skeleton-based gait recognition. *Expert Systems*, 40(6):e13244, 2023. 2

[50] Hao Zhang, Xin Chen, Heming Jing, Yingbin Zheng, Yuan Wu, and Cheng Jin. Etr: An efficient transformer for re-ranking in visual place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5665–5674, 2023. 2

[51] Renjie Zhang, Yu Fang, Huaxin Song, Fangbin Wan, Yanwei Fu, Hirokazu Kato, and Yang Wu. Specialized re-ranking: A novel retrieval-verification framework for cloth changing

person re-identification. *Pattern Recognition*, 134:109070, 2023. 3

[52] Yuqi Zhang, Qi Qian, Hongsong Wang, Chong Liu, Weihua Chen, and Fan Wang. Graph convolution based efficient re-ranking for visual retrieval. *IEEE Transactions on Multimedia*, 26:1089–1101, 2023. 3, 6, 7

[53] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20228–20237, 2022. 1, 2, 3, 5, 7

[54] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1318–1327, 2017. 2, 3, 6, 7

[55] Yunhao Zhou, Yi Wang, and Lap-Pui Chau. Moving towards centers: Re-ranking with attention and memory for re-identification. *IEEE Transactions on Multimedia*, 25:3456–3468, 2022. 2

[56] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19370–19380, 2023. 2, 3

[57] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14789–14799, 2021. 2, 3, 5, 6