GaitW: Enhancing Gait Recognition in the Wild using Dynamic Information

Daksh Thapar*[1], Jayesh Chaudhari*[2], Sunny Manchanda[3], Aditya Nigam[1], and Chetan Arora[2]

[1] Indian Institute of Technology Mandi, India
[2] Indian Institute of Technology Delhi, India
[3] Defense young Scientist Laboratory, India
d18033@students.iitmandi.ac.in, jayeshc.cstaff@iitd.ac.in
faculty.iitmandi.ac.in/~aditya/, www.cse.iitd.ac.in/~chetan/
sunny.dysl-ai@gov.in

Abstract. Success of modern deep neural networks (DNNs) for gait recognition on in-the-lab datasets such as CASIA-B and OU-MVLP have encouraged the community to aim for more challenging, and in-the-wild datasets such as GREW and Gait3D. The new datasets contain large variations in silhouettes due to change in camera pose, clothing, accessories, as well as occlusion, thus posing huge challenges to existing techniques and training strategies for gait recognition. We posit that to achieve high accuracy in in-the-wild datasets, explicitly leveraging dynamic information in gait samples during training is imperative. We propose a novel transformer based architecture for gait recognition specifically leveraging such dynamic information. The novel contributions include: (1) We propose interleaved spatial and temporal encoders to attend to positioning of various body parts in a frame, and movement of a body part across the sample, respectively. (2) We propose a novel dynamic information inspired curriculum, where we first determine the hardness of a sample based on the disparity between representations of its frame-wise silhouettes (FWSs) and GEI. The model is trained using easier samples first, followed by progressively difficult samples. (3) We propose mask-annealing for silhouettes using Gait Energy Images (GEIs), which attends to silhouette contours and allows a model to learn robust silhouette shape representation. We report a significant improvement in accuracy (in %) of 96.9, 92.9, 81.2, and 67.7 on benchmark CASIA-B, OU-MVLP, GREW, and Gait3D datasets respectively using our technique, against the current state-of-the-art (SOTA) accuracy of 96.9, 92.4, 77.4, and 67.0 by MSGR [43], HSTGait [41], SkeletonGait [13], and QAGait [45] respectively. In a significant departure from the current trend, and as evident from the above numbers, the proposed technique sets up simultaneous SOTA on most prominent in-the-lab as well as in-the-wild datasets. Complete source code and trained models of our method will be publicly available.

Keywords: Gait Recognition · Silhouettes · Curriculum Learning

* These authors contributed equally to this work

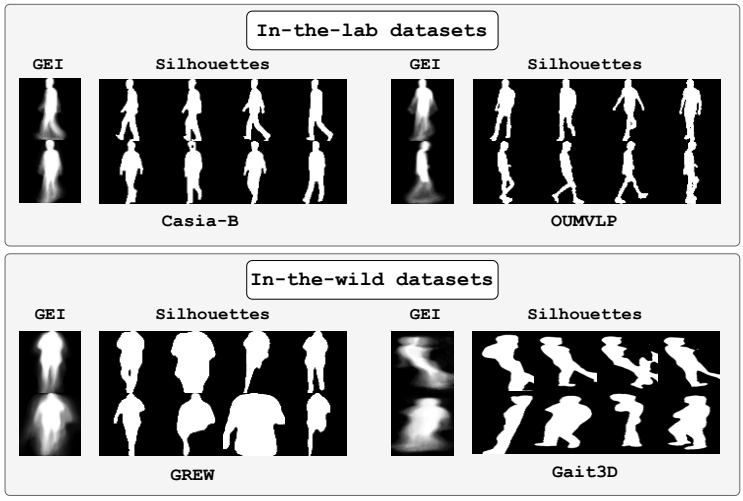

Fig. 1: [Challenges in the new, in-the-wild gait datasets] Comparative analysis between traditional gait datasets, `CASIA-B` and `OU-MVLP`, and the new and more challenging `GREW` and `Gait3D` datasets. For each dataset, the first row presents Gait Energy Images (`GEI`) and frame-wise silhouettes (`FWSs`) of a training sample, while the second row displays a testing sample of the same subject. The key observation of this work is that for simple samples with minimal movement variation, the `GEI` and silhouette representations bear close resemblance, encapsulating similar informational content. In contrast, for complex samples characterized by significant dynamic motion, there is a notable disparity in the information conveyed by the silhouettes and `GEI`. This divergence underscores the need for a newer gait recognition model, like our `GaitW`, which explicitly focuses on the regions with high dynamic information, as well as account for the sample hardness due to high dynamic information content in the learning strategy.

1 Introduction

Background Gait recognition is a well explored problem in computer vision [5–9, 14, 15, 18, 20, 21, 28, 30, 31, 35, 39–42] due to its ability to recognize a subject from a distant camera, even without the subject's active cooperation. This makes it a technique of choice for forensic identification, as well as security and surveillance [4, 25]. Due to the distances involved, input images to a gait recognition are typically low resolution with large illumination changes in all weather conditions. Hence, most gait recognition techniques have avoided use of a subject's pose or skeleton which can not be robustly extracted from such views. Instead, the input of choice has been binary silhouettes, that ignores background, respects privacy of the subjects, and can be robustly computed using automated techniques. However, wide variations in silhouettes due to walking, loose clothing, accessories, and unseen viewpoints makes analyzing such inputs highly challenging. Sample silhouettes and `GEIs` are shown in Fig. 1.

Current Status and Challenges Earlier in-the-lab datasets (e.g. `CASIA-B` and `OU-MVLP`) for gait recognition were typically captured indoors with a fixed number of known camera positions and controlled changes in clothing of a sub-

ject between train and test sequences. Number of accessories (such as handbag etc.) on the subject were also kept minimal to prevent extensive silhouette change. This makes the problem relatively easier than in-the-wild datasets. Hence, whereas several gait recognition systems [7, 30, 41] achieve good performance on such benchmarks, their generalization to newer in-the-wild datasets (`GREW`, and `Gait3D`) is not as impressive (c.f. Tab. 1). This has also led to the fragmentation in the community with several techniques exclusively focusing on dataset specific cues and reporting impressive results only on one kind of datasets. Fig. 1 visually motivates some of the critical differences between the two kind of datasets.

Our proposal We observe that, whereas classical techniques have mostly used `GEIs` as the inputs, with the improvements in `DNN` architectures and larger datasets, recent techniques have increasingly used raw frame-wise silhouettes (`FWSs`). However, as the datasets became more complex (e.g. `Gait3D`, and `GREW`), researchers have increasingly focused their attention towards capturing more accurate dynamic information by attending to larger moving parts [42], or adaptive region based motion extraction [41], or precise silhouette alignment using skeleton coordinates [13]. We believe that leveraging dynamic information in a gait sample is critical for robust recognition and higher accuracy for in-the-wild datasets. With this objective, we propose a novel transformer based `DNN` model which leverages dynamic information in a gait sample using interleaved spatial and temporal encoders, `GEI` based silhouette mask annealing, and a novel easy-to-hard curriculum learning strategy which determines a sample hardness based on disparity between the representations of `FWSs` and its `GEI`.

Contributions The key contributions of this paper are:

1. **Gait specific spatial and temporal encoders:** Unlike for natural images/videos [22, 33], gait datasets are much smaller. Hence, in this paper we propose extracting accurate dynamic information by focusing on most informative and dynamic regions in a gait sample. In a binary silhouette most of the information is contained in the contour shape and relative motion of a particular part across the frames. To help proposed model focus on these specific features, we divide each frame into tokens and then leverage spatial (for contour shape), and temporal attention (for movement of a part across frames). Our model contains 3 pairs of such interleaved spatial and temporal encoders trained using triplet loss.
2. **Curriculum learning with gait specific hardness scoring:** We posit that samples where frame-wise silhouette masks are much different than corresponding `GEI` has high degree of dynamism and thus more difficult for a network to learn. Inspired from the success of Curriculum Learning (`CL`) in other similar vision tasks, we propose a gait specific curriculum, where samples with low dynamism are fed for the training first, followed by gradually more difficult samples.
3. **GEI based silhouette mask annealing:** To help our model focus on the silhouette contours, we ignore those regions which are always foreground across frames. We create multiple relaxed versions of such regions to ignore and fine-

tune using them at different training iterations starting from relaxed to strict version (more details in Sec. 4.3) . As we show through experiments such silhouette mask annealing helps our model focus on the most dynamic and informative regions of a silhouette and learn more robust gait representations.

4. **Simultaneous state-of-the-art on all major datasets:** A standout feature of our framework is its ability to achieve SOTA results across all prominent gait datasets. We report an accuracy (%) of 96.9, 92.9, 81.2, and 67.7 on benchmark CASIA-B, OU-MVLP, GREW, and Gait3D datasets respectively using our technique, against the current SOTA accuracy of 96.9, 92.4, 77.4, and 67.0 by MSGR [43](TMM'23), HSTGait [41] (ICCV'23), SkeletonGait [13] (AAAI'24), and QAGait [45] (AAAI'24) respectively.

2 Related Work

Methods focusing on in-the-lab datasets CASIA-B and OU-MVLP datasets stand out in this category as meticulously crafted under controlled setting, capturing diverse angles, clothing variations, and accessories. CASIA-B provides RGB and silhouette data, while OU-MVLP provides the silhouette and pose information. GaitSet [6] uses silhouettes and devised statistical functions to construct gait templates. Recently proposed, DANet [30] (CVPR'23) works on silhouettes and models global motion patterns through discrete local motion, giving competitive performance on both CASIA-B and OU-MVLP datasets. 2D poses/skeleton information is utilized by PoseGait [36] that leverages human body priors to extract robust pose features via a CNN model. Similarly, GaitGraph [38] also employs 2D pose data to extract robust gait information using Graph Convolutional Network (GCN), achieving competitive results on OU-MVLP. In another recent work, MSGR [43] (TMM'23) fuses both skeleton and silhouette features and demonstrates competitive results on CASIA-B and OU-MVLP.

Methods focusing on in-the-wild datasets Recently introduced GREW (2021) and Gait3D (2022) datasets challenged SOTA gait recognition techniques by capturing data under real-world, outdoor settings, encompassing diverse environmental and occlusion conditions. DyGait [42] (ICCV'23) utilizes silhouettes and introduces a module to extract dynamic gait features and delivers impressive results on both GREW, and Gait3D datasets. A hierarchical body parts approach, HSTGait [41] (ICCV'23), has been proposed using silhouettes, and leveraging local motion patterns by learning region-independent spatio-temporal representation. In another recent work, GaitGCI [9] (CVPR'23), silhouette data is used and a generative counterfactual intervention approach has been deployed showcasing its effectiveness on both in-the-wild datasets. QA-Gait [45] proposes quality aware and quality assessment framework for silhouettes which tries to mitigate the noise present in silhouettes. In order to explore cross-dataset generalizability, GPGait [15] (ICCV'23) utilizes 2D poses, with a Human-Oriented Transformation Module. In ICCV'23, Physics Augumented Autoencoder (PAA) [17] has been proposed using 2D poses, and training an autoencoder architecture to model skeleton joint and forces between them. It has shown good results on the Gait3D

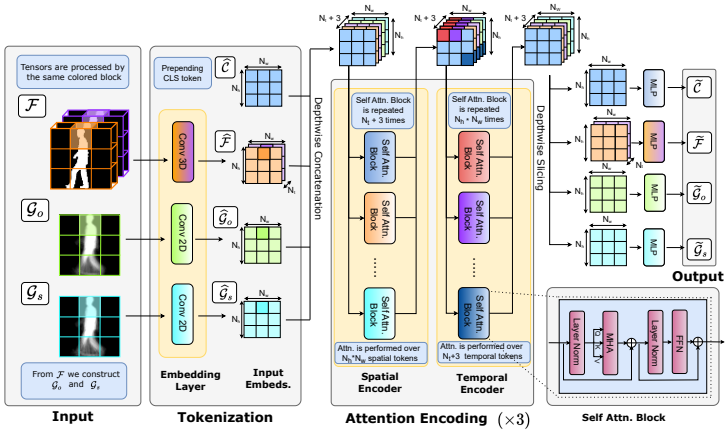

Fig. 2: [**Proposed `GaitW` architecture**] The model takes `FWSs` ($\mathcal{F}$), `GEI` ($\mathcal{G}_o$) and subsampled `GEI` ($\mathcal{G}_s$) as an input. Initially, input tokenization is performed, with the addition of a classification token. Subsequently, spatial attention is applied, followed by temporal attention, resulting in robust output feature vectors: $\widetilde{\mathcal{F}}$, $\widetilde{\mathcal{G}}_o$, $\widetilde{\mathcal{G}}_s$ and $\widetilde{\mathcal{C}}$. These features are then utilized for metric learning in the first stage followed by gait sample hardness score based curriculum learning in the second stage.

as compared with other pose based models. However, it is noteworthy that both skeleton based `GPGait` and `PAA` significantly lags behind silhouette-based methods, due to the poor performance of pose extraction under these settings. This is further showcased by `SkeletonGait` [13], which uses skeleton data to generate skeletal maps composing Gaussian maps of joints and limbs, and gives `SOTA` performance on `GREW`, but fails on `Gait3D` and `OU-MVLP` due to view angle variations. Tab. 1 in Sec. 6, shows accuracy (%) of recent techniques over four prominent gait datasets and confirms the observation.

Curriculum Learning In recent years, curriculum learning [3, 11] has received growing research interests in natural language processing [23, 24, 29, 32, 49, 52, 53] and computer vision [26, 44, 47]. For example, [32] proposed sentence length, or the rarity of the words appearing in it [53], as the measure of difficulty. Similarly, [46] utilized manual annotation to describe the difficulty, and [16, 52] utilized another meta learner network to predict the difficulty of a sample. To the best of our knowledge ours is the first work proposing gait sample hardness to design a suitable curriculum. We speculate that our work may inspire similar curriculum for other video analysis tasks as well.

3 Proposed Architecture

Overall architecture of proposed `GaitW` is visually depicted in Fig. 2. We describe the architecture below.

Input Originally, input to our model is frame-wise silhouettes (`FWSs`), denoted as $\mathcal{F} \in \mathbb{R}^{T \times H \times W}$, where T is the length of gait sequence and $(H \times W)$ indicates the

image size of each frame. While FWSs offers locally rich and dynamic information they lack a global gait representation. To overcome this limitation, we generate the global representations by averaging FWSs over time, resulting in Gait Energy Images (GEIs) denoted as $\mathcal{G}_o$. Furthermore, we perform global averaging at a different temporal resolution, by temporally subsampling FWSs, and produce subsampled GEIs, denoted as $\mathcal{G}_s$. We use $\mathcal{G}_s$ to achieve robustness against walking speed variations. Hence, the input for our model can be represented as a tuple, $(\mathcal{F}, \mathcal{G}_o, \mathcal{G}_s)$, where $\mathcal{F} = \{f_i \mid i = 1, 2, \ldots, T\}$, and $f_i \in \mathbb{R}^{H \times W}$ is the i^{th} frame of the input gait sequence. Further, $\mathcal{G}_o = \frac{1}{T} \sum_{i+=1}^{T}{}_{i=1} f_i$, and $\mathcal{G}_s = \frac{1}{T} \sum_{i+=2}^{T}{}_{i=1} f_i$. The input, $\mathcal{I}$, then undergoes tokenization at the patch level through raster scanning, moving from left to right and top to bottom.

Tokenizing frame-wise silhouette Given frame-wise silhouettes $\mathcal{F} \in \mathbb{R}^{H \times W \times T}$ for a particular gait sample, we extract non-overlapping, 3D spatio-temporal patches (tokens) of size (h, w, t) from the input volume. Each of this patch is projected onto d-dimensional vector embedding using d, 3D convolution filters of size (h, w, t) and stride of (h, w, t). This results in a 4D tensor, $\widehat{\mathcal{F}}$, of size $N_h \times N_w \times N_t \times d$, where $N_h = H/h$, $N_w = W/w$, and $N_t = T/t$.

GEI Tokenization We describe here tokenization process for $\mathcal{G}_o$. Tokenization of $\mathcal{G}_s$ is performed similarly. Given a GEI image $\mathcal{G}_o \in \mathbb{R}^{H \times W}$, we extract non-overlapping, 2D patches of size (h, w), and project them to a d-dimensional token embedding, using d convolution filters of size (h, w), and a stride of (h, w). This results in a tensor, $\widehat{\mathcal{G}}_o$, of size $N_h \times N_w \times d$ (similarly $\widehat{\mathcal{G}}_s$ corresponding to $\mathcal{G}_s$).

Tensor concatenation and classification token We concatenate tensors $\widehat{\mathcal{G}}_o$, and $\widehat{\mathcal{G}}_s$ with $\widehat{\mathcal{F}}$ along the temporal dimension. Additionally, another tensor, $\widehat{\mathcal{C}}$ of size $N_h \times N_w \times d$ is appended. The role of $\widehat{\mathcal{C}}$ is similar to classification token in many transformer models [2], as the output token ($\widetilde{\mathcal{C}}$) corresponding to $\widehat{\mathcal{C}}$ is finally used as the embedding vector for the gait sample. Overall, a tensor of size $N_h \times N_w \times (N_t + 3) \times d$ is passed onto the next stage comprising of proposed spatial and temporal encoders as shown in Fig. 2.

Spatial and temporal encoders In the proposed spatial encoder block, we deploy $(N_t + 3)$ attention heads, each of which compute attention over $(N_h \times N_w)$ tokens independently. The purpose of spatial encoder is to focus on the spatial arrangement of various body parts in a particular frame, as well as GEI images. Similarly in the proposed temporal encoder, we slice along the temporal axis, and deploy $(N_h \times N_w)$ attention heads, each of which computes attention independently over $(N_t + 3)$ tokens. Complementing spatial encoders, role of temporal encoder is to focus on movement of various body parts over time, as each token represents a specific spatial region corresponding to original silhouette and GEI. The proposed architecture contains 3 blocks of spatial-temporal encoder pairs. The output tensor after the spatial-temporal encoder blocks is of size $N_h \times N_w \times (N_t + 3) \times d$ which is further transformed using separate MLP's $(\mathcal{O})$ and then sliced into 4 different tensors, $\widetilde{\mathcal{C}}$, $\widetilde{\mathcal{G}}_o$, $\widetilde{\mathcal{G}}_s$ and $\widetilde{\mathcal{F}}$ (each of size $N_h \times N_w \times d$). Here $\widetilde{\mathcal{F}}$ is averaged in N_t dimension.

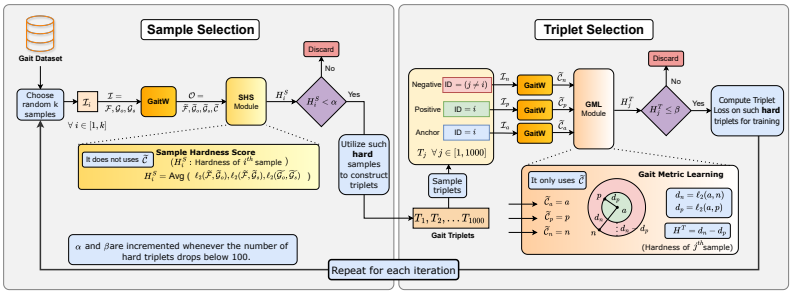

Fig. 3: [**Proposed Curriculum (Stage-II)**] Input comprises `FWSs` and `GEI`, which undergo processing to produce $\widetilde{\mathcal{C}}$ using `GaitW`. The system utilizes a decision mechanism to detect hard samples based on sample hardness ($H^S < \alpha$). These samples are subsequently utilized to form triplets for Gait Metric Learning, employing only $\widetilde{\mathcal{C}}$. Later, hard triplets violating the margin criterion (i.e., $(H^T = d_n - d_p) \leq \beta$) are selected, and triplet loss is computed solely from them.

4 Proposed Training Methodology

Our proposed training methodology contains three distinct steps/stages, each with a different strategy. In the first step we train our model using metric learning with triplet loss. The key novelty introduced in this stage is dynamic margin adaptation. In the second stage of the training we introduce sample hardness score and train with an easy-to-hard curriculum learning strategy. In the third stage of the training, we introduce mask-annealing. Each of these stages is described in detail in this section below.

4.1 Stage-I: Gait Metric Learning

Triplet loss The first stage of training involves gait metric learning with triplet loss [34]. Here a gait sample is randomly chosen first as anchor, then a positive and negative sample is chosen at random, such that a positive sample is from the same subject as anchor, and negative implies from any subject but the one in anchor. We then generate embedding vectors $\widetilde{\mathcal{C}}_a, \widetilde{\mathcal{C}}_p, \widetilde{\mathcal{C}}_n$ corresponding to anchor, positive, and negative samples respectively as described in Sec. 3. The triplet loss for the three chosen samples is then computed as:

$$\mathcal{L}_t = \max\left(0, d_p - d_n + \beta\right), \quad \text{where} \quad d_p = \ell_2(\widetilde{\mathcal{C}}_a, \widetilde{\mathcal{C}}_p), \quad d_n = \ell_2(\widetilde{\mathcal{C}}_a, \widetilde{\mathcal{C}}_n). \quad (1)$$

Here β is a chosen constant called *margin*.

Hard negative mining Effectively sampling triplets for training is critical in gait recognition systems using triplet loss. Random triplet selection often results in the inclusion of easy triplets, where positive pairs are highly similar and negative pairs are markedly dissimilar. This can cause the network to converge quickly to a local minimum, primarily catering to easy triplets to meet the margin requirement. The key lies in judiciously choosing only the *hard* triplets in a training batch violating the margin constraint (i.e., $(d_n - d_p) \leq \beta$).

Margin adaptation In our proposed approach, we introduce margin adaptation to dynamically adjust the margin parameter (β). As the training progresses, the representations of anchor, positive, and negative samples evolve, resulting in a decrease in d_p and an increase in d_n. Consequently, number of hard triplets decreases over iterations, indicating the network's improved ability to differentiate within the specified margin. Hence, we propose to initialize β by ($\beta_{\min} = 1$), but increase the margin by raising the β value by 0.1, whenever the proportion of hard triplets in the selected triplets falls below 10%. We keep increasing β until it reaches $\beta_{\max} = 1.5$. The proposed margin adaptation enables efficient learning from simpler samples during phases of low margin ($\beta \approx \beta_{\min}$) and from more complex samples during phases of high margin ($\beta \approx \beta_{max}$). By dynamically regulating the margin, we achieve a balance between under-fitting and over-fitting, thereby maximizing learning across a wide range of triplet complexities.

4.2 Stage-II: Sample hardness based curriculum

For in-the-lab datasets such as `CASIA-B` and `OU-MVLP`, only Stage-I training is typically enough. However, for in-the-wild `GREW` and `Gait3D` datasets, many samples have large pose variations, making it extremely hard for a network to learn their representation in the initial training phases. Hence, we propose to use `SHS` score to design a curriculum utilising easy to hard gait samples in Stage-II as shown in Fig. 3.

Sample hardness score and loss We propose a novel sample hardness score (`SHS`) based on the disparity in representations of `FWSs` and `GEIs` for a gait sample. The `SHS` score for a gait sample is computed as:

$$\mathcal{H} = \frac{1}{3} \sum \left(\ell_2(\widetilde{\mathcal{F}}, \widetilde{\mathcal{G}}_o), \ell_2(\widetilde{\mathcal{F}}, \widetilde{\mathcal{G}}_s), \ell_2(\widetilde{\mathcal{G}}_o, \widetilde{\mathcal{G}}_s) \right), \tag{2}$$

where $\ell_2(\cdot, \cdot)$ represents Euclidean distance between the two vectors. We introduce an additional loss $\mathcal{L}_h$ which minimizes the hardness score defined as the average hardness of the anchor, positive, and negative samples. Hence, for Stage-II, the overall loss is computed as, $\mathcal{L} = \mathcal{L}_t + \lambda \mathcal{L}_h$, where λ is an appropriately chosen hyperparameter.

Training with easy-to-hard samples We train our model in Stage-I for 400 iterations, or until $\beta < 1.5$. In Stage-II as shown in Fig. 3, we only select samples with a `SHS` score lower than α, and generate triplets using only these samples. We then utilize hard negative mining, same as in Stage-I, to select triplets that violate the margin criterion and use them for the training (recall that $\beta \geq 1.5$ in Stage-II). We initialize α with $\alpha_{\min} = 0.5$ in Stage-II. As training progresses, number of hard triplets decreases, and when their proportion fall below 10% of the total samples, we increment α by 0.15, and β by 0.1 (same as in Stage-I). We cap the β value at $\beta_{\max} = 2.0$, and α at $\alpha_{\max} = 6.0$. We run Stage-II for fixed 150 iterations.

4.3 Stage-III: GEI based silhouette mask annealing

To allow our model to focus on dynamic, contour regions of a silhouette, we fine-tune it further using silhouette mask annealing. Within a `GEI` image, pixels corresponding to the inside of the body are characterized by intensity values of 255, whereas dynamic, silhouette contour regions display varying intensities. We denote the former as *foreground*, and later as *boundary* pixels. We derive a binary mask, m_0, from $\mathcal{G}_o$ by designating foreground pixels as 0 and boundary pixels as 1. All the pixels marked 0 in $\mathcal{G}_o$ are also kept at 0. This mask is progressively refined to m_1 and m_2 using image dilation. In Stage-III, we train our model in the same fashion as Stage-I, except for the fact that each silhouette image is masked using m_2 for first 50 iterations, using m_1 for next 50 iterations, and then using m_0 for next 50 iterations. This implies that for all pixels in a silhouette image where the corresponding mask pixel is 1, we retain their original value, and set the remaining pixels in the `GEI` image to 0. For more detailed description please refer the supplementary material.

5 Datasets and Evaluation Methodology

Datasets (1) `CASIA-B` **[51]:** It is the most popular gait dataset containing 124 subjects, 13,640 sequences with 3 walking conditions and 11 views (0°, 18°, ..., 180°), acquired under controlled lab environment. The walking condition contains normal (NM) (6 sequences per subject), walking with a bag (BG) (2 sequences per subject) and wearing a coat or jacket (CL) (2 sequences per subject). Each subject has $11 \times (6+2+2) = 110$ sequences. **(2)** `OU-MVLP` **[37]:** It is one of the largest public gait datasets, captured under the protocol described in [27,48]. It contains 10,307 subjects each of which have two groups of gait sequences (with a total of 288,596 sequences). Each group contains gait sequences at 14 angles (0°; 15°; ...; 270°). **(3)** `GREW` **[57]:** It is a recently introduced, one of the most challenging gait datasets, captured in unconstrained outdoor conditions. It is a massive dataset consisting of 26,345 subjects with 128,671 sequences acquired using 882 cameras under realistic and varying environmental conditions. Similar to other datasets, `GREW` provides silhouette masks along with human pose data, but no RGB images (to maintain anonymity). We use this dataset to analyze the effectiveness of our gait recognition technique in in-the-wild settings. **(4)** `Gait3D` **[55]:** The `Gait3D` dataset represents a large-scale repository featuring 3D perspective. Comprising a total of 4,000 subjects, the dataset captures individuals in an unconstrained setting through 39 cameras.

Experimental Setup (1) `CASIA-B`**:** As there is no officially suggested training-test split, hence, in order to have a fair comparison, we conduct experiments on three settings as suggested in most `SOTA` techniques [27,48]. We name these three settings as Small-sample Training (ST), Medium-sample Training (MT) and Large-sample Training (LT). In ST (referred to as ST24), the first 24 subjects (labeled from 001-024) are used for training, and the remaining 100 subjects are left for testing. In MT (referred to as MT62), the first 62 subjects are used

for training, and the rest 62 subjects are left for testing. In LT (referred to as LT74), the first 74 subjects are used for training, and the remaining 50 subjects are left for testing. In test sets of all three settings, first, 4 sequences of NM condition (NM 1-4) are kept in the gallery, and the remaining 6 sequences are divided into 66 probe subsets, i.e. NM subsets containing NM 5-6, BG subsets containing BG 1-2 and CL subsets containing CL 1-2. **(2)** `OU-MVLP`**:** For this dataset, sequences of first 5153 subjects have been used for training and the remaining 5154 subjects are utilized for testing as suggested in [27, 48], in order to have a fair comparison. In the test set, sequences with NM#01 are kept in the gallery, and NM#02 is used as probes. **(3)** `GREW`**:** The dataset contains training data with 20,000 identities and 102,887 sequences [57]. The testing set of `GREW` contains 6,000 identities and 24,000 sequences. Each subject in the test set has 4 sequences, 2 for probe and 2 for the gallery. The dataset does not provide labels for probe sequences. For calculating results, the output has to be uploaded on their website [1]. Hence, for ablation study we have created a subset test data `GREW-1K`. The first 1000 subjects in `GREW` are chosen. In this, the gallery data containing 2 sequences is labelled. Hence, we chose 1 sequence as a gallery and other sequence as a probe. For comparison with `SOTA` we have utilized full `GREW` test data, and for ablation and development study, we have utilized `GREW-1K`. **(4)** `Gait3D`**:** This dataset is partitioned into a training set of 3,000 subjects, and a test set of another 1,000 subjects. To facilitate evaluation, we follow the strategy utilized by `SOTA` techniques [41, 42, 55], one sequence from subject is considered to be a probe set, while the remaining sequences are designated as the gallery set for matching purposes.

Evaluation Metric Gait recognition is $(1 : N)$ search process that retrieves the same person from the gallery, given a probe subject. When evaluated on test set, probe sample is matched with all gallery samples. For `CASIA-B` and `OU-MVLP` datasets, we adopt Rank-1 matching accuracy as the evaluation metric. For `GREW`, we adopt Rank-k matching accuracy as the evaluation metric which denotes the possibility to locate at least one true positive in the top-k ranks (with k ranging from 1 to 20). For `Gait3D` we report Rank-1 accuracy, Rank-5 accuracy, mAP (mean average precision) and mINP (mean inverse negative penalty) [50].

6 Experiments and Results

Our experimental analysis consists of following main components. **(1)** We compare our technique with other `SOTA` methods across four public gait datasets: `CASIA-B`, `OU-MVLP`, `GREW`, and `Gait3D`. **(2)** We present the robustness analysis of `GaitW` against occlusion on `GREW-1K` dataset. **(3)** We conduct a thorough ablation study for open-set recognition specifically on the `GREW-1K` dataset. **(4)** We have also reported qualitative analysis along with extended results in the supplementary material to validate our proposed technique.

Implementation Details We train our `GaitW` using Adam optimizer, with an initial learning rate and weight decay set to 1e-5. In our `GaitW` architecture, we employ 3 spatio-temporal attention blocks with the number of attention heads

		CASIA-B	OU-MVLP	GREW				Gait3D			
Method	**Venue**	**Mean**	**Mean**	**R-1**	**R-5**	**R-10**	**R-20**	**R-1**	**R-5**	**mAP**	**mINP**
GaitPart [14]	CVPR'20	88.8	88.7	44.0	60.6	67.25	73.4	29.9	50.6	23.3	13.15
GLN [18]	ECCV'20	89.46	89.18	-	-	-	-	42.2	64.5	33.1	19.5
3D-Local [21]	ICCV'21	92.7	90.9	-	-	-	-	-	-	-	-
CSTL [20]	ICCV'21	93.4	90.2	50.6	65.9	71.9	76.9	12.2	21.7	6.4	3.3
SRN+CB [19]	TBBIS'21	93.4	90.1	-	-	-	-	-	-	-	-
GaitGL [28]	ArxiV'22	93.5	92.0	68.0	80.7	85.0	88.2	63.8	80.5	55.8	36.7
LangGait [5]	CVPR'22	92.3	90.0	-	-	-	-	-	-	-	-
SMPLGait [55]	CVPR'22	-	-	-	-	-	-	53.2	71.0	42.4	25.9
MetaGait [8]	ECCV'22	93.4	91.9	-	-	-	-	-	-	-	-
MTSGait [54]	ACM'22	-	-	55.3	71.3	76.9	81.6	48.7	67.1	37.6	21.9
GaitGCI [9]	CVPR'23	94.5	92.1	68.5	80.8	84.9	87.7	57.2	74.5	45.0	27.6
MMGaitf. [7]	CVPR'23	96.4	90.1	-	-	-	-	-	-	-	-
DANet [30]	CVPR'23	94.6	90.7	-	-	-	-	48.0	69.7	-	-
GaitBase [12]	CVPR'23	89.6	90.8	60.1	-	-	-	65.6	-	-	-
GaitRef [56]	IJCB'23	94.0	90.2	53.0	67.9	73.0	77.5	49.0	69.3	40.6	25.2
STANet [31]	ICCV'23	94.6	90.7	-	-	-	-	-	-	-	-
DyGait [42]	ICCV'23	94.1	-	71.4	83.2	86.8	89.5	66.3	80.8	56.4	**37.3**
HSTgait [41]	ICCV'23	94.3	**92.4**	62.7	76.5	81.3	85.2	61.3	76.3	55.5	34.7
MSGR [43]	TMM'23	**96.9**	91.2	-	-	-	-	-	-	-	-
MSAFF [58]	IJCB'23	**96.5**	-	57.4	72.9	78.2	82.8	48.1	66.6	38.4	23.4
QA-Gait [45]	AAAI'24	90.2	-	59.1	74.0	79.2	83.1	**67.0**	**81.5**	**56.5**	-
Skel.Gait [13]	AAAI'24	-	67.4	**77.4**	**87.9**	**91.0**	**93.2**	38.1	56.7	28.9	16.1
CLASH [10]	TIP'24	93.9	91.9	67.0	78.9	83.0	85.8	58.9	75.7	47.3	29.4
GaitW	Proposed	**96.9**	**92.9**	**81.2**	**89.9**	**92.6**	**94.3**	**67.7**	**83.3**	**56.7**	**37.9**

Table 1: Aggregated performance across 4 major datasets: In this table we report average accuracy achieved under the NM, BG, and CL conditions for CASIA-B dataset. We report average Rank-1 accuracy from all angles (excluding identical viewpoints) for the OU-MVLP dataset, and Rank-1 to Rank-20 accuracy for the GREW. For Gait3D dataset we have reported Rank-1, Rank-5, mAP and mINP. The accuracy for compared techniques are taken as reported by the respective authors in their papers. Hence, wherever a particular technique has not reported result on a dataset, we indicate it with a "-". Further detailed results for each dataset are provided in the supplementary material for thorough examination. For each column best performance is highlighted in Green color while second best is shown in Red color.

in each encoder set to 8. The 2D/3D convolution kernels for generating input embedding are trained from scratch, alongside other parameters of the model. For all four datasets (CASIA-B, OU-MVLP, Gait3D, and GREW), we utilize gait sequences (FWSs) of size (25, 64, 64), where 25 represents the number of frames and (64, 64) denotes the frame resolution. We then append the classification token ($\widehat{\mathcal{C}}$), and GEIs $\mathcal{G}_o$, and $\mathcal{G}_s$, each of shape (1, 64, 64), to obtain a final tensor of size (28, 64, 64) as shown in Fig. 2. The hidden dimension size (d) is adjusted according to the dataset: 128 for CASIA-B, 256 for OU-MVLP, and 512 for both Gait3D and GREW. We use data augmentation as suggested in various SOTA [45], including Horizontal Flip (HF), Rotation (R), Perspective Transformation (PT), Affine Transformation (AT), and Random Erasing (RE).

6.1 Comparative Analysis

Simultaneous SOTA over CASIA-B, OU-MVLP, GREW and Gait3D datasets Tab. 1 offers an in-depth comparison showcasing how our proposed model, GaitW, stands against recent advancements in SOTA gait recognition techniques across all 4 datasets. It is evident from the comparison that GaitW consistently achieves

Gallery NM#1-4		Angles from 0°-108°											
Method	**Venue**	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
GaitPart [14]	CVPR'20	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
GLN [18]	ECCV'20	70.6	82.4	85.2	82.7	79.2	76.4	76.2	78.9	77.9	78.70	64.3	77.5
3DLocal [21]	ICCV'21	78.2	90.2	92.0	87.1	83.0	76.8	83.1	86.6	86.8	84.1	70.9	83.7
CSTL [20]	ICCV'21	78.1	89.4	91.6	86.6	82.1	79.9	81.8	86.3	88.7	86.6	75.3	84.2
SRN+CB [19]	TBBIS'21	75.1	88.2	89.9	86.3	81.2	78.8	80.0	84.0	86.3	80.7	68.8	81.8
GaitGL [28]	ArXiv'22	82.6	92.6	94.2	91.8	86.1	81.3	87.2	90.2	90.9	88.5	75.4	87.3
LangGait [5]	CVPR'22	77.4	90.6	93.2	90.2	84.7	80.3	85.2	87.7	89.3	86.6	71.0	85.1
MetaGait [8]	ECCV'22	80.0	91.8	93.0	87.8	86.5	82.9	85.2	90.0	90.8	89.3	78.4	86.9
GaitGCI [9]	CVPR'23	-	-	-	-	-	-	-	-	-	-	-	88.5
MMGaitF. [7]	CVPR'23	**93.9**	**98.0**	96.9	**96.0**	**93.7**	**91.6**	**93.5**	**96.4**	**96.5**	**95.7**	**90.2**	**94.8**
DANet [30]	CVPR'23	82.8	94.8	96.9	94.3	89.0	83.9	87.9	92.3	95.1	92.0	80.3	89.9
GAitBase [12]	CVPR'23	-	-	-	-	-	-	-	-	-	-	-	77.4
GaitRef [56]	IJCB'23	81.4	93.3	94.3	91.6	87.8	83.9	88.5	91.7	91.6	89.1	75.0	88.0
STANet [31]	ICCV'23	83.4	94.6	96.8	93.9	89.6	86.0	88.2	92.1	93.9	90.2	78.1	89.7
DyGait [42]	ICCV'23	82.2	93.0	95.2	91.6	87.1	83.4	87.2	90.1	92.4	88.2	75.8	87.8
HSTGait [41]	ICCV'23	82.4	94.2	95.0	91.7	88.2	83.3	88.0	92.3	93.1	91.0	78.5	88.9
MSGR [43]	TMM'23	91.8	95.8	**97.3**	95.2	**93.0**	91.1	93.4	96.0	95.5	**95.5**	89.7	94.0
MSAFF [58]	IJCB'23	92.1	94.6	95.6	93.8	91	90.6	92.5	94	95.3	94.8	**91.7**	93.3
QA-Gait [45]	AAAI'24	-	-	-	-	-	-	-	-	-	-	-	78.2
CLASH [10]	TIP'24	-	-	-	-	-	-	-	-	-	-	-	88.0
GaitW		**94.2**	**97.7**	**97.0**	**96.3**	**93.7**	**92.1**	**93.8**	**96.7**	**96.8**	95.4	89.7	**94.9**

Table 2: Rank-1 accuracy (%) on CASIA-B on all angles, CL#1-2 conditions, under LT-74 setting. Symbol "-" indicates results for this configuration were not reported by the said method. For each column best performance is highlighted in Green color while second best is shown in Red color.

unparalleled SOTA performance across all datasets, underscoring its effectiveness and robustness in gait recognition. Interestingly, it also reveals that the runner-up techniques, in terms of performance, keep on varying for each dataset. This is indicative of competitive landscape in gait recognition technologies.

CASIA-B **Dataset** In Tab. 1, we presented average accuracy achieved under the NM, BG, and CL conditions. Interestingly, despite MSGR's approach of integrating both silhouettes and pose data to improve gait recognition, our model, GaitW, achieved comparable SOTA accuracy utilizing only silhouette information. In Tab. 2 we show detailed results for all angles in CL condition, as the condition is known for its challenge due to clothing variations. Notably, GaitW excels here also, surpassing MSGR by a margin of 0.9%, even though MSGR showed comparable mean overall accuracy in Tab. 1. Under the CL condition, MMGaitF. (another method which fuses silhouettes and pose data) emerges as the runner-up with a mean accuracy of 94.8%, closely trailing behind GaitW's 94.9%. For a thorough analysis, including results under the NM and BG conditions, please refer to the supplementary material.

OU-MVLP **Dataset** Detailed comparative performance on OU-MVLP dataset is shown in Tab. 3. Similar to other SOTA works, for a probe sample from a particular view, we exclude samples with the same angle from the gallery. We report averaged results over all views in the table. The result shows that GaitW has a good generalization under wide view variations. GaitW outperforms SOTA by a margin of 0.5%, improving the performance from 92.4% to 92.9%.

GREW **Dataset** As shown in Tab. 1, one can observe that GaitW sets a new state-of-the-art in gait recognition in the wild with its performance on GREW. Our technique gives a score of 81.2% in Rank-1 metric, which exceeds SOTA SkeletonGait by about 3.8%. It should be noted that GREW contains 24,000 sequences test

Method	Venue	Probe View														
		0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean
GaitPart [14]	CVPR'20	82.6	88.9	90.8	91.0	89.7	89.9	89.5	85.2	88.1	90.0	90.1	89.0	89.1	88.2	88.7
GLN [18]	ECCV'20	83.8	90.0	91.0	91.2	90.3	90.0	89.4	85.3	89.1	90.5	90.6	89.6	89.3	88.5	89.2
3D Local [21]	ICCV'21	86.1	91.2	92.6	92.9	92.2	91.3	91.1	86.9	90.8	92.2	92.3	91.3	91.1	90.2	90.9
CSTL [20]	ICCV'21	87.1	91.0	91.5	91.8	90.6	90.8	90.6	89.4	90.2	90.5	90.7	89.8	90.0	89.4	90.2
SRN+CB [19]	TBBIS'21	85.6	90.7	91.5	91.7	90.6	90.6	90.1	86.8	90.0	90.9	91.1	89.9	90.0	89.3	89.9
GaitGL [28]	Arxiv'22	91.1	92.6	92.3	92.5	92.8	92.2	92.1	92.4	91.9	91.7	91.9	92.1	91.5	91.4	92.0
LangGait [5]	CVPR'22	85.9	90.6	91.3	91.5	91.2	91.0	90.6	88.9	89.2	90.5	90.6	89.9	89.8	89.2	90.0
MetaGait [8]	ECCV'22	88.2	92.3	**93.0**	**93.5**	**93.1**	**92.7**	**92.6**	89.3	91.2	**92.0**	**92.6**	**92.3**	**91.9**	91.1	91.9
GaitGCI [9]	CVPR'23	91.2	92.3	92.6	92.7	93.0	92.3	92.1	92.0	91.8	91.9	**92.6**	**92.3**	91.4	91.6	92.1
MMGaitF. [7]	CVPR'23	-	-	-	-	-	-	-	-	-	-	-	-	-	-	90.1
DANet [30]	CVPR'23	87.7	91.3	91.6	91.8	91.7	91.4	91.1	90.4	90.3	90.7	90.9	90.5	90.3	89.9	90.7
GaitBase [12]	CVPR'23	-	-	-	-	-	-	-	-	-	-	-	-	-	-	90.8
GaitRef [56]	IJCB'23	85.7	90.5	91.6	91.9	91.3	91.3	90.9	89.3	89.0	90.8	90.8	90.1	90.1	89.5	90.2
STANet [31]	ICCV'23	87.7	91.4	91.6	91.9	91.6	91.4	91.2	90.4	90.3	90.8	91.0	90.5	90.3	90.1	90.7
HSTGait [41]	ICCV'23	**91.4**	**92.9**	92.7	93.0	92.9	92.5	92.5	**92.7**	**92.3**	**92.1**	92.3	92.2	**91.8**	**91.8**	**92.4**
MSGR [43]	TMM'23	89.3	91.9	91.9	92.1	92.0	91.6	91.3	90.9	90.9	91.2	91.3	91.1	90.8	90.5	91.2
Skel.Gait [13]	AAAI'24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	67.4
CLASH [10]	TIP'24	91.0	92.2	92.3	92.6	92.7	92.0	92.0	91.8	91.6	91.6	92.5	92.1	91.2	91.3	91.9
GaitW		**92.5**	**92.7**	**94.5**	**93.1**	**93.3**	**92.8**	**92.7**	**93.1**	**92.8**	91.7	**96.3**	**92.5**	**91.8**	**92.0**	**92.9**

Table 3: Per view Rank-1 accuracy for the OU-MVLP. We also report averaged results over all views. Symbol "-" denotes that the particular technique did not report this result. For each column best performance is highlighted in Green color while second best is shown in Red color.

Table 4: Impact of Occlusion on Accuracy: Rank-1 (R-1) and Rank-5 (R-5) accuracy on subsets of GREW-1K, categorized into three occlusion levels: (1) none or minor, (2) moderate, and (3) severe occlusion. For each column, best performance is in Green color and second best is in Red.

Method	No/minor		Moderate		Severe	
	R1	R5	R1	R5	R1	R5
Gaitset [6]	69.7	85.9	59.8	79.9	45.2	62.7
GaitGL [28]	69.9	85.1	60.8	77.3	45.2	57.6
GaitPart [14]	69.9	87.0	60.8	77.1	48.1	62.7
GaitBase [12]	**79.2**	**90.6**	**67.8**	**84.9**	**52.5**	**69.3**
GaitW	**82.2**	**97.1**	**74.3**	**89.5**	**60.9**	**75.8**

samples. Hence, the gain of 3.8% is statistically significant in the case of GREW dataset.

Gait3D Dataset Similar to GREW, Gait3D also contains challenging, realistic scenarios with 3D views and large-scale data. GaitW shows superior performance in all metrics (c.f. Tab. 1), indicating its ability to model gait characteristics in realistic scenarios. The proposed GaitW sets up a new SOTA of 67.7% against the current 67.0% by QAGait.

6.2 Robustness against the occlusion

To evaluate our model's effectiveness in handling occlusion, we meticulously reviewed 1,000 test samples from the GREW-1K dataset, classifying them based on occlusion levels into three categories: **(1)** No or minor occlusion, **(2)** Moderate occlusion, and **(3)** Severe occlusion (occlusion labels will be made publicly available). The distribution of samples showed that 47% had no or minor, 40% had moderate, and 13% had severe occlusion. Tab. 4 shows performance comparison between GaitW and recent models (whose code was publically available), separately across these occlusion categories. Our findings, based on using publicly available pre-trained weights of SOTA models (for the GREW dataset), reveal that GaitW performs 3% better than SOTA in "no or minor occlusion". But notably, GaitW shows significant improvements in moderate and severe occlusion, with

Table 5: Ablation study. Rank-1 and Rank-5 accuracy metrics for our model on the GREW-1K dataset across different configurations.

Config.	Rank-1	Rank-5	Configuration	Rank-1	Rank-5
1. $\mathcal{G}_o$	38.22	39.67	7. $\mathcal{F}$+$\mathcal{G}_o$+$\mathcal{G}_s$ (Stage-I only)	49.23	65.67
2. $\mathcal{G}_s$	35.86	36.72	8. Stage-I and Stage-III	52.11	67.68
3. $\mathcal{G}_o$+ $\mathcal{G}_s$	39.14	45.62	9. Stage-I and Stage-II	58.42	74.86
4. $\mathcal{F}$	41.82	46.47	10. All modules, without Alignment	63.61	78.34
5. $\mathcal{F}$+$\mathcal{G}_o$	49.54	60.68	11. All modules, with Alignment	71.83	85.17
6. $\mathcal{F}$+$\mathcal{G}_s$	47.45	59.21			

accuracy gain of 6.5% (from 67.8% to 74.3%) and 8.4% (from 52.5% to 60.9%), respectively.

6.3 Ablation Study

Tab. 5, outlines the results of our ablation study, which assesses the impact of individual components in our model's performance. We have performed this study on GREW-1K. Each configuration, denoted by combinations of $\mathcal{F}$, $\mathcal{G}_o$ and $\mathcal{G}_s$, tests different inputs to our model, illustrating their unique contributions. Note that while ablating for input combination, we use only Stage-I training. Hence, the row corresponding to "$\mathcal{F}$+$\mathcal{G}_o$+$\mathcal{G}_s$", essentially represents all inputs, but with only Stage-I training. Further, in the adjacent rows, we ablate under inclusion of various training stages in our model.

Alignment pre-processing It is important to describe "Without alignment" row as it has not been discussed so far in this manuscript. Alignment is a fundamental pre-processing step involving silhouette centering, cropping, and aligning for uniform representation across videos, is a common practice in prior studies, and now standardized in the OpenGait [12] library. Although alignment is usually a prerequisite for consistent GEI generation, we evaluate our model's performance with and without this step. Highlighting our model's effectiveness without alignment is crucial, considering alignment might not always be precise, particularly in high occlusion, and extreme viewpoint variation scenarios. The good performance of our model, even without alignment, underscores the model's adaptability, which has not been reported by any previous work.

7 Conclusion

Datasets are important trigger for the development of novel models in many computer vision problems. For gait recognition as well, datasets like CASIA-B, and OU-MVLP served an important for the progress of the field. However, availability of newer datasets like Gait3D and GREW, provides an opportunity to understand implicit assumptions, and limitations of current techniques, and help develop newer and more robust models. This paper is an attempt towards that challenge. With more unconstrained motion, and camera views, the level of variations in silhouettes has increased significantly. Developing a curriculum which can separate easier samples from difficult looks promising to simplify learning. Coupled with carefully designed architecture, and training schema to focus on dynamic and salient regions of a sample allows our proposed model to significantly improve the state of the art on in-the-lab as well as in-the-wild datasets. We hope that our effort will encourage more such explorations to develop robust techniques which improve the performance on all kinds of datasets simultaneously.

References

1. Grew submission link, https://codalab.lisn.upsaclay.fr/competitions/3409
2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6836–6846 (2021)
3. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009)
4. Bouchrika, I., Goffredo, M., Carter, J., Nixon, M.: On using gait in forensic biometrics. Journal of forensic sciences **56**(4), 882–889 (2011)
5. Chai, T., Li, A., Zhang, S., Li, Z., Wang, Y.: Lagrange motion analysis and view embeddings for improved gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20249–20258 (June 2022)
6. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8126–8133 (2019)
7. Cui, Y., Kang, Y.: Multi-modal gait recognition via effective spatial-temporal feature fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17949–17957 (June 2023)
8. Dou, H., Zhang, P., Su, W., Yu, Y., Li, X.: Metagait: Learning to learn an omni sample adaptive representation for gait recognition (2023)
9. Dou, H., Zhang, P., Su, W., Yu, Y., Lin, Y., Li, X.: Gaitgci: Generative counterfactual intervention for gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5578–5588 (June 2023)
10. Dou, H., Zhang, P., Zhao, Y., Jin, L., Li, X.: Clash: Complementary learning with neural architecture search for gait recognition. IEEE Transactions on Image Processing pp. 1–1 (2024). https://doi.org/10.1109/TIP.2024.3360870
11. Elman, J.L.: Learning and development in neural networks: The importance of starting small. Cognition **48**(1), 71–99 (1993)
12. Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., Yu, S.: Opengait: Revisiting gait recognition toward better practicality. arXiv preprint arXiv:2211.06597 (2022)
13. Fan, C., Ma, J., Jin, D., Shen, C., Yu, S.: Skeletongait: Gait recognition using skeleton maps (2023)
14. Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., He, Z.: Gaitpart: Temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14225–14233 (2020)
15. Fu, Y., Meng, S., Hou, S., Hu, X., Huang, Y.: Gpgait: Generalized pose-based gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19595–19604 (October 2023)
16. Gong, C., Tao, D., Maybank, S.J., Liu, W., Kang, G., Yang, J.: Multi-modal curriculum learning for semi-supervised image classification. IEEE Transactions on Image Processing **25**(7), 3249–3260 (2016)
17. Guo, H., Ji, Q.: Physics-augmented autoencoder for 3d skeleton-based gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19627–19638 (October 2023)
18. Hou, S., Cao, C., Liu, X., Huang, Y.: Gait lateral network: Learning discriminative and compact representations for gait recognition. In: Vedaldi, A., Bischof, H., Brox,

T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 382–398. Springer International Publishing, Cham (2020)
19. Hou, S., Liu, X., Cao, C., Huang, Y.: Set residual network for silhouette-based gait recognition. IEEE Transactions on Biometrics, Behavior, and Identity Science **3**(3), 384–393 (2021)
20. Huang, X., Zhu, D., Wang, H., Wang, X., Yang, B., He, B., Liu, W., Feng, B.: Context-sensitive temporal feature learning for gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12909–12918 (October 2021)
21. Huang, Z., Xue, D., Shen, X., Tian, X., Li, H., Huang, J., Hua, X.S.: 3d local convolutional neural networks for gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14920–14929 (2021)
22. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset (2017)
23. Kocmi, T., Bojar, O.: Curriculum learning and minibatch bucketing in neural machine translation. arXiv preprint arXiv:1707.09533 (2017)
24. Kumar, G., Foster, G., Cherry, C., Krikun, M.: Reinforcement learning based curriculum optimization for neural machine translation. arXiv preprint arXiv:1903.00041 (2019)
25. Larsen, P.K., Simonsen, E.B., Lynnerup, N.: Gait analysis in forensic medicine. Journal of forensic sciences **53**(5), 1149–1153 (2008)
26. Li, Q., Huang, S., Hong, Y., Zhu, S.C.: A competence-aware curriculum for visual concepts learning via question answering. In: European Conference on Computer Vision. pp. 141–157. Springer (2020)
27. Lin, B., Zhang, S., Bao, F.: Gait recognition with multiple-temporal-scale 3d convolutional neural network. In: Proceedings of the 28th ACM international conference on multimedia. pp. 3054–3062 (2020)
28. Lin, B., Zhang, S., Wang, M., Li, L., Yu, X.: Gaitgl: Learning discriminative global-local feature representations for gait recognition. arXiv preprint arXiv:2208.01380 (2022)
29. Liu, X., Lai, H., Wong, D.F., Chao, L.S.: Norm-based curriculum learning for neural machine translation. arXiv preprint arXiv:2006.02014 (2020)
30. Ma, K., Fu, Y., Zheng, D., Cao, C., Hu, X., Huang, Y.: Dynamic aggregated network for gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22076–22085 (June 2023)
31. Ma, K., Fu, Y., Zheng, D., Peng, Y., Cao, C., Huang, Y.: Fine-grained unsupervised domain adaptation for gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11313–11322 (October 2023)
32. Platanios, E.A., Stretcu, O., Neubig, G., Poczos, B., Mitchell, T.M.: Competence-based curriculum learning for neural machine translation. arXiv preprint arXiv:1903.09848 (2019)
33. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge (2015)
34. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)

35. Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Geinet: View-invariant gait recognition using a convolutional neural network. In: 2016 international conference on biometrics (ICB). pp. 1–8. IEEE (2016)
36. Sokolova, A., Konushin, A.: Pose-based deep gait recognition (2018)
37. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Transactions on Computer Vision and Applications **10**(1), 1–14 (2018)
38. Teepe, T., Khan, A., Gilg, J., Herzog, F., Hormann, S., Rigoll, G.: Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In: 2021 IEEE International Conference on Image Processing (ICIP). IEEE (Sep 2021). https://doi.org/10.1109/icip42928.2021.9506717
39. Thapar, D., Jaswal, G., Nigam, A., Arora, C.: Gait metric learning siamese network exploiting dual of spatio-temporal 3d-cnn intra and lstm based inter gait-cycle-segment features. Pattern Recognition Letters **125**, 646–653 (2019)
40. Thapar, D., Nigam, A., Aggarwal, D., Agarwal, P.: Vgr-net: A view invariant gait recognition network. In: 2018 IEEE 4th international conference on identity, security, and behavior analysis (ISBA). pp. 1–8. IEEE (2018)
41. Wang, L., Liu, B., Liang, F., Wang, B.: Hierarchical spatio-temporal representation learning for gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19639–19649 (October 2023)
42. Wang, M., Guo, X., Lin, B., Yang, T., Zhu, Z., Li, L., Zhang, S., Yu, X.: Dygait: Exploiting dynamic representations for high-performance gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13424–13433 (October 2023)
43. Wang, R., Shi, Y., Ling, H., Li, Z., Zhao, C., Wei, B., Li, H., Li, P.: Gait recognition with multi-level skeleton-guided refinement. IEEE Transactions on Multimedia pp. 1–12 (2023). https://doi.org/10.1109/TMM.2023.3323887
44. Wang, Y., Gan, W., Yang, J., Wu, W., Yan, J.: Dynamic curriculum learning for imbalanced data classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5017–5026 (2019)
45. Wang, Z., Hou, S., Zhang, M., Liu, X., Cao, C., Huang, Y., Li, P., Xu, S.: Qagait: Revisit gait recognition from a quality perspective. arXiv preprint arXiv:2401.13531 (2024)
46. Wei, J., Suriawinata, A., Ren, B., Liu, X., Lisovsky, M., Vaickus, L., Brown, C., Baker, M., Nasir-Moin, M., Tomita, N., et al.: Learn like a pathologist: curriculum learning by annotator agreement for histopathology image classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2473–2483 (2021)
47. Weinshall, D., Cohen, G., Amir, D.: Curriculum learning by transfer learning: Theory and experiments with deep networks. In: International Conference on Machine Learning. pp. 5238–5246. PMLR (2018)
48. Wolf, T., Babaee, M., Rigoll, G.: Multi-view gait recognition using 3d convolutional neural networks. In: 2016 IEEE international conference on image processing (ICIP). pp. 4165–4169. IEEE (2016)
49. Xu, C., Hu, B., Jiang, Y., Feng, K., Wang, Z., Huang, S., Ju, Q., Xiao, T., Zhu, J.: Dynamic curriculum learning for low-resource neural machine translation. arXiv preprint arXiv:2011.14608 (2020)
50. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H.: Deep learning for person re-identification: A survey and outlook (2021)

51. Yu, S., Tan, D., Tan, T.: 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). vol. 4, pp. 441–444 (2006). https://doi.org/10.1109/ICPR.2006.67
52. Zhang, X., Kumar, G., Khayrallah, H., Murray, K., Gwinnup, J., Martindale, M.J., McNamee, P., Duh, K., Carpuat, M.: An empirical exploration of curriculum learning for neural machine translation. arXiv preprint arXiv:1811.00739 (2018)
53. Zhao, M., Wu, H., Niu, D., Wang, X.: Reinforced curriculum learning on pre-trained neural machine translation models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 9652–9659 (2020)
54. Zheng, J., Liu, X., Gu, X., Sun, Y., Gan, C., Zhang, J., Liu, W., Yan, C.: Gait recognition in the wild with multi-hop temporal switch (2022)
55. Zheng, J., Liu, X., Liu, W., He, L., Yan, C., Mei, T.: Gait recognition in the wild with dense 3d representations and a benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
56. Zhu, H., Zheng, W., Zheng, Z., Nevatia, R.: Gaitref: Gait recognition with refined sequential skeletons (2023)
57. Zhu, Z., Guo, X., Yang, T., Huang, J., Deng, J., Huang, G., Du, D., Lu, J., Zhou, J.: Gait recognition in the wild: A benchmark. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14789–14799 (2021)
58. Zou, S., Xiong, J., Fan, C., Yu, S., Tang, J.: A multi-stage adaptive feature fusion neural network for multimodal gait recognition (2023)