sensors

MDPI

*Article*

# WildGait: Learning Gait Representations from Raw Surveillance Streams

**Adrian Cosma and Ion Emilian Radoi ***

Faculty of Automatic Control and Computer Science, University Politehnica of Bucharest, 060042 Bucharest, Romania; ioan_adrian.cosma@stud.acs.upb.ro

* Correspondence: emilian.radoi@cs.pub.ro

**Simple Summary:** In this work, we explore self-supervised pretraining for gait recognition. We gather the largest dataset to date of real-world gait sequences automatically annotated through pose tracking (UWG), which offers realistic confounding factors as opposed to current datasets. Results highlight the great performance in scenarios with low amounts of training data, and state-of-the-art accuracy on skeleton-based gait recognition when utilizing all available training data.

**Abstract:** The use of gait for person identification has important advantages such as being non-invasive, unobtrusive, not requiring cooperation and being less likely to be obscured compared to other biometrics. Existing methods for gait recognition require cooperative gait scenarios, in which a single person is walking multiple times in a straight line in front of a camera. We address the challenges of real-world scenarios in which camera feeds capture multiple people, who in most cases pass in front of the camera only once. We address privacy concerns by using only motion information of walking individuals, with no identifiable appearance-based information. As such, we propose a self-supervised learning framework, WildGait, which consists of pre-training a Spatio-Temporal Graph Convolutional Network on a large number of automatically annotated skeleton sequences obtained from raw, real-world surveillance streams to learn useful gait signatures. We collected and compiled the largest pretraining dataset to date of anonymized walking skeletons called Uncooperative Wild Gait, containing over 38k tracklets of anonymized walking 2D skeletons. We make the dataset available to the research community. Our results surpass the current state-of-the-art pose-based gait recognition solutions. Our proposed method is reliable in training gait recognition methods in unconstrained environments, especially in settings with scarce amounts of annotated data.

**Keywords:** gait recognition; pose estimation; graph neural networks; self-supervised learning



**Citation:** Cosma, A.; Radoi, I.E. WildGait: Learning Gait Representations from Raw Surveillance Streams. *Sensors* **2021**, *21*, 8387. https://doi.org/10.3390/s21248387

Academic Editor: Michele Calì

Received: 15 October 2021
Accepted: 9 December 2021
Published: 15 December 2021



## 1. Introduction

Human behaviour is complex, it defines us and is driven, in part, by the environment, social influences, life experiences, and internal emotional factors, attitudes and values. While some of the effects of individual actions and decisions are long reaching and abstract, low-level behaviour in the form of movement in physical space is highly indicative of a person's immediate intentions, state of mind and identity.

Walking is one of the most fundamental actions we perform and the study of gait (manner of walking) has gained increased attention in recent years, as it encodes important behavioral biometric information, and the recent advancements in machine and deep learning provide the necessary toolset to model this information. Walking patterns can be used to estimate the age and gender of a person [1], estimate emotions [2], and provide insight into various physiological conditions [3]. Moreover, aside from these soft-biometrics, gait information is used as a unique fingerprinting method for identifying individuals. Although face recognition has become the norm for person identification in a range of

suitable applications with good results, gait recognition from video is still a challenging task in real-world scenarios. The intrinsic dynamic nature of walking makes it susceptible to a multitude of confounding factors such as view angle, shoes and clothing, carrying variations, age, interactions with other people and various actions that the person is performing while walking.

Most of the existing studies that use gait to estimate various internal aspects of a person are performed in highly constrained conditions, requiring subjects' cooperation, often involving walking on a treadmill [4] in a laboratory setting. Since so far the study of gait recognition is mostly performed in controlled environments, only a subset of confounding factors were explored [5]. As of date, the conditions in which such a system powered by machine learning algorithms would operate in a real-world setting is poorly studied. Current datasets [4–7] focus on the change in view angle, clothing and carrying conditions, while ignoring other behavioral variations. As opposed to face recognition datasets, gait recognition datasets are harder to build, and require the cooperation of thousands of subjects in order to be relevant. Furthermore, privacy laws restrict the usage of these datasets in real world applications.

We propose WildGait, a framework for automatically learning useful, discriminative representations for gait recognition from raw, real-world data, without explicit human labels. We leverage surveillance streams of people walking and employ state of the art pose estimation methods (e.g., AlphaPose [8], OpenPose [9], LCRNet [10]) to extract skeleton sequences and pose tracking to construct a loosely annotated dataset. Making use of a diverse set of augmentation procedures, we pretrain our network with minimal direct supervision—the only indirect label employed is the pose tracking information uncovered automatically. The deep learning community has recently an increased interest in self-supervised learning: methods that leverage large datasets without labels to learn meaningful and informative representations of input samples. Traditionally, these methods involved using a pretext task [11–13] as a proxy to learn representations, but they have shifted towards, non-pretext, contrastive learning [14–17].

The reason for using pose estimations is that they do not contain identifiable appearance-based information of walking individuals. Current methods that use silhouettes [18] are unsuitable in real-world, dynamic scenarios in which changes in illumination and multiple overlapping individuals severely affect the quality of extracted silhouettes. Traditional silhouette-based methods also encode information about the subjects, as general clothing, hairstyles and accessories are distinguishable. While some approaches explicitly disentangle appearance features and pose information [7], we argue that appearance-based methods are too invasive in terms of the privacy of individuals. Skeletons extracted from human pose estimation methods encode only motion information, which is sufficient to determine if two skeleton representations belong to the same person, without holding any information about the person's appearance.

Pose information also enables leveraging information of performed actions and activities, and filtering out individuals that are not walking or have abnormal walking patterns [19]. For pretraining, we propose the Unconstrained Wild Gait dataset (UWG), which unlike current available datasets, contains anonymized skeleton sequences of a large number of people walking in a natural environment (over 38k tracklets), with many walking variations and confounding factors—the data is gathered from raw, real-world video streams (Figure 1). People walking in UWG are present only once, from a single viewing angle and with a constant array of intrinsic confounding factors (clothing, shoes etc.), making it a highly challenging dataset. We leverage this noisy information to pretrain a neural network to better handle controlled gait sequences in scenarios with few data samples.

**Figure 1.** Samples from the UWG dataset. We gather skeleton sequences by applying state-of-the-art pose estimation/pose tracking models on publicly available surveillance streams. This allows for learning discriminative gait representations without explicit human labels and without the cooperation of subjects.

This paper makes the following contributions:

1. We are among the first to explore self-supervised learning on gait recognition, and we propose WildGait, a data collection and pretraining pipeline, which enables learning meaningful gait representations in a self-supervised manner, from automatically extracted skeleton sequences in unconstrained environments.
2. The Unconstrained Wild Gait dataset (UWG), the largest dataset to date of noisily tracked skeleton sequences to enable the gait recognition research community to further explore ways to pretrain gait recognition systems in an self-supervised manner.
3. A study on transfer learning capabilities of our pretrained network on popular gait recognition databases, highlighting great performance in scenarios with low amounts of training data, and state-of-the-art accuracy on skeleton-based gait recognition when utilizing all available training data.

## 2. Related Work

The research attention received by gait recognition over the past decade has been increasing. A significant portion of this research was dedicated to gait recognition using wearable inertial sensors [20,21], however, our focus is on recognition of gait using camera sensors. Gait recognition approaches can be classified into two main categories: model-based and appearance-based.

### *2.1. Appearance-Based Methods*

One of the most prevalent approaches for appearance-based gait recognition is the use of a Gait Energy Image (GEI) [18]. GEIs are computed by averaging the silhouettes of walking individuals across a gait cycle. Such images are then processed using modern standard image processing approaches. Since GEIs have limitations, such as not taking into account temporal information, several variants are proposed to address these shortcomings, notably Gait Entropy Image (GEnI) [22], Gait Flow Image [23] and Chrono-Gait Image [24], all showing good performance on benchmark datasets.

More recent approaches tend to use appearance features to explicitly learn a disentangled representation. The authors of [7] propose a way to explicitly disentangle motion and appearance features via an autoencoder with carefully designed loss functions.

### *2.2. Model-Based Approaches*

As opposed to appearance-based methods, model-based methods process walking patterns as a set of human joint trajectories across time. The performance increase of

pose-estimation models [25], enabled the use of skeletons sequences in gait recognition. The authors of [26] directly use the joint probability heatmaps as an input for an Long-Short Term Memory (LSTM) [27] network. LSTM architectures are widely used for modelling temporal data, and model-based gait is naturally processed as a sequence. The authors of [28] used an LSTM and a CNN to process 2D skeleton sequences to account for the temporal and spatial variations of walking. The authors of [29] used an LSTM autoencoder with contrastive learning to further stabilize the joint trajectories of skeletons. Recently, the authors of [19] enhanced the robustness of estimated skeletons by constructing a quality-adjusted cost matrix between input frames and registered frames for frame-level matching. The authors of [30] propose a a fully-connected network to model a single skeleton, and temporal aggregation of skeletal features for the final classification. Both references [31,32] leverage 3D skeletons to model gait patterns, with incremental improvements over 2D skeletons. Similar to us, reference [33] applies a graph convolutional network to process skeleton sequences, but use a final pyramid pooling layer for recognition. An important aspect of gait recognition for deployment in practice is multi-gait, in which multiple people walk together, and their individual walking patterns change. The authors of [34] proposed an attribute discovery model in a max-margin framework to recognize a person based on gait while walking with multiple people.

### *2.3. Gait Recognition Datasets*

Several benchmark datasets have been proposed to test the performance of gait recognition systems in the presence of a standard array of confounding factors. The most popular dataset is CASIA-B [5], which is comprised of 126 different identities, that walk several times in front of 11 cameras. Each identity also has different walking variations (clothing/carrying conditions) that affect the walking patterns. More recently, Front-View Gait Database (FVG) [7] was proposed to tackle the most difficult camera view-angle for gait processing systems (i.e., walking towards the camera). The authors propose several confounding factors, including clothing and carrying conditions, background clutter, and the passage of time. Another recent dataset, OU-MVLP [4], is introduced as a benchmark for evaluating the scalability of gait representations, being one of the largest datasets, comprised of 10.000 identities captured from 14 cameras. However, OU-MVLP is still captured in a highly controlled environment, requiring the subjects to walk on a treadmill. This makes it less suitable to pretrain models capable of generalizing across walking scenarios, compared to our proposed UWG dataset. UWG is designed to be representative of the general walking population, with a multitude of camera angles and variations. Moreover, Nixon et al. [35] provided a comprehensive review of current datasets and applications for gait-based identification, and identified a need for developing datasets and methods for scaling methods outside of controlled environments.

### *2.4. Unsupervised Skeleton-Based Methods*

Little research was performed on gait recognition in more constrained scenarios, with little to no annotated data. Closer to the work proposed in this paper are the recent advancements in skeleton-based action recognition in scenarios of little or no supervision. The authors of [36] proposed to use an LSTM autoencoder to learn discriminative representations for activity recognition, without any supervision except the skeleton sequences. Similarly, reference [37] used an iterative approach in an active learning setting. The authors of [38] used a self-supervised approach to activity recognition, in which they propose several pretext tasks to pretrain the network, such as pose shuffling and motion prediction. However, different from the previous action recognition methods which aim to ambiguate the identity from various actions, we target the opposite problem: given a single action (i.e., walking), we aim to uncover the identity. As such, we posit that directly using unsupervised action recognition methods for pretraining is unsuitable for our case.

Different from self-supervised methods, we take a data-driven approach. We process large amounts of raw data from real-world video streams, automatically extracting

skeletons from each frame, performing intra-frame pose tracking and filtering unwanted skeletons (i.e., poorly extracted/non-walking). We train a Spatio-Temporal Graph Convolutional Network (ST-GCN) [39] using contrastive learning for gait recognition. ST-GCN is widely used for processing skeleton sequences for a diverse set of human actions and has also been successfully used in gait recognition settings [33]. Without any major architectural modifications we obtain exceptional results in scenarios with scarce amounts of data.

## 3. Method

This section offers an overview of the methodology for collecting the Unconstrained Wild Gait dataset from raw surveillance streams, and of the self-supervised pretraining procedure.

### *3.1. Dataset Construction*

Our aim is to learn good gait representations from human skeleton sequences in unconstrained environments, without explicit labels. For this purpose we collect a sizeable dataset of human walking skeleton sequences from surveillance camera feeds, with a high variance of walking styles and from various geographic locations, environments, weather conditions and camera angles. The dataset captures a multitude of confounding factors in the manner of walking of individuals. We named this dataset Unconstrained Wild Gait (UWG), and releasing it upon request to the research community to further advance the field of gait recognition. UWG is intended for pretraining gait analysis models, which will be further fine-tuned on downstream tasks. Figure 2 showcases our data collection procedure, from raw surveillance streams to anonymized skeleton sequences. All video streams captured for UWG are collected from static surveillance cameras publicly available on the internet. These streams are often "street cams" that continuously stream a busy street in a city. Real-world surveillance streams offer significantly greater variation of viewpoints and confounding factors, with people changing walking direction, having heavy clothing and carrying accessories in unconventional ways. Figure 3 shows samples from from UWG. Moreover, scenes can be very crowded (right-hand side of Figure 3), which impacts both the gathering of gait cycles, as some joints are not always visible, and the walking dynamics themselves (i.e., walking to avoiding other people, keeping the same walking speed).

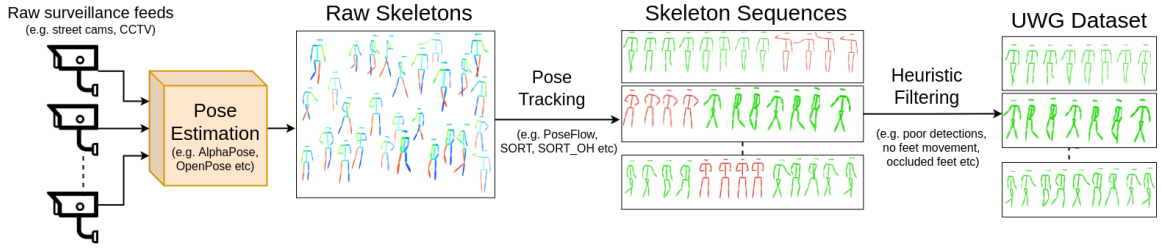


**Figure 2.** Data collection procedure. We process raw video streams from real-world settings of people walking and construct Unconstrained Wild Gait (UWG), a dataset of skeleton sequences loosely annotated through pose estimation and pose tracking. We use this raw and noisy data to pretrain a gait recognition model that generalizes well to different gait recognition scenarios. Skeletons are extracted from raw surveillance feeds with a well known multi-person pose estimation model, AlphaPose [8], which are then tracked intra-camera with a pose tracking model (SORT [40]). Finally, sub-par skeleton sequences are removed with simple heuristics. Best viewed in color.

We postulate that constructing UWG through mining real-world surveillance streams would be a necessary step in making gait analysis models more robust to "out-of-distribution" walks, as opposed to combining multiple existing datasets [4,5,7]. An ensemble of controlled datasets will still provide only controlled walks, which would make the model susceptible to outliers. Works in out-of-distribution detection have suggested that "exposing" models to a large and diverse set of samples increases robustness to outliers [41].

**Figure 3.** Samples of people walking from UWG. As opposed to strictly controlled mainstream datasets, the skeleton sequences provided by UWG were captured in real-world conditions (including sequences from crowded scenes) and with real-world confounding factors (in the form of heavy clothing, a variety of footwear, and multiple carrying conditions). Faces are blurred for privacy reasons. UWG contains only anonymized, out-of-context skeleton sequences—samples shown here are only for illustrating the real-world confounding factors.

Since people from different geographic locations have different manners of walking influenced by cultural and societal norms [42], we gathered data from 3 major continents (Europe, Asia and North America—Table 1) to capture subject diversity. To obtain skeleton sequences for people in a video, we first preprocess the streams to a common 24 fps and 720p resolution. Then, a crowd pose estimation method (in our case, AlphaPose [8]) was used to extract human skeletons for each frame, and skeletons were tracked intra-camera, across time, using SORT [40]. We chose SORT [40] as it has high performance and reasonable accuracy. However, as in the case of the pose estimation algorithm, the exact method is not crucial, so long as skeleton sequences of reasonable accuracy are uncovered. Moreover, we chose SORT as it only operates on the coordinates of the bounding box, and not on the appearance of individuals, as in the case of Deep SORT [43]. We aimed to use as little appearance information as possible in our processing pipeline. The output of the AlphaPose model consists of 17 joints with X and Y coordinates of the joints and the confidence, in the COCO [44] pose format.

Unstructured environments invariably introduce noise in the final skeleton sequences due to unreliable extracted skeletons, lost tracking information, or people standing who perform activities other than walking. To address this, we only keep sequences with a mean confidence on extracted joints of over 60% and with no more than 3 consecutive frames with feet confidence of less than 50%. This way we eliminate poorly extracted poses and ensure that the feet are visible throughout the sequence. Moreover, we enforce a minimum tracking sequence of 54 frames, which corresponds to approximately two full gait cycles [45].

**Table 1.** Geographic locations from where we gathered skeleton sequences. We aimed for a multi-cultural representation of walking people. * Approximate number given by pose tracker.

| Continent | # IDs * | Walk Length (hr) | Avg. Run Len. (Frames) |
|---|---|---|---|
| Asia | 3635 | 4.79 | 104.5 |
| Europe | 12,993 | 19.76 | 110.1 |
| North America | 21,874 | 34.47 | 108.2 |

Further, based on known body proportions we normalize each skeleton to be invariant to the height of the person by first zero-centering each skeleton by subtracting the pelvis coordinates, and then normalizing the Y coordinate by the length between the neck and the pelvis, and the X coordinate by length from the right to the left shoulder (Equations (1) and (2)). The normalization procedure ensures the skeleton sequences are aligned and similar poses have close coordinates. Moreover, by removing information

related to the height of the person, information related to the stature and particular body characteristics of a person are eliminated.

$$x_{joint} = \frac{x_{joint} - x_{pelvis}}{|x_{R.shoulder} - x_{L.shoulder}|} \quad (1)$$

$$y_{joint} = \frac{y_{joint} - y_{pelvis}}{|y_{neck} - y_{pelvis}|} \quad (2)$$

We address the issue of non-walking people with a heuristic on the movement of the joints corresponding to the legs (feet and knees). We compute the average movement velocity of the feet, which is indicative of the activity of a person. Thus, we filter out individuals with an average velocity magnitude of less than 0.01. Extremely long tracklets are also filtered out, as it was noticed that individuals tracked for a longer time are usually standing (not walking). Figure 4 showcases the distribution of the final tracklet durations.

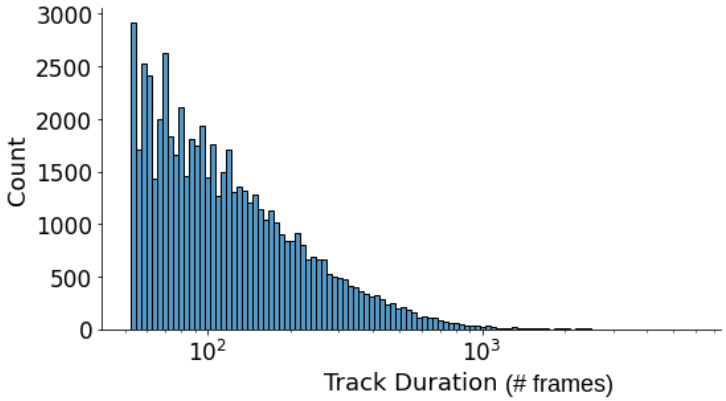


**Figure 4.** Distribution of tracklets durations in the UWG dataset. Longer tracklets capture more diversity within the same subject's walk.

The proposed framework does not rely on appearance information at any step in the processing pipeline, except when extracting the pose information. Table 2 shows a comparison between UWG and CASIA-B and FVG in terms of diversity and size. A total of 38,502 tracklets were obtained, with an average walk duration of 108 frames. While this is an approximation of the number of different identities in the dataset, it is possible that some tracklets belong to the same person, due to lost tracking information by the pose tracking model. This is not an issue since identity-related information is not used in the pretraining stage. The total walking sequences duration in the dataset is of approximately 60 h. The scenario for this dataset is more restrictive compared to other benchmark datasets, as it does not include multiple runs of the same person from multiple camera angles and with multiple walking styles (such as different carrying and clothing conditions). Still, even from this restrictive scenario, the large amount of data is leveraged to learn good gait representations that transfer well to other benchmark datasets. In the proposed configuration, the UWG dataset is used only for pretraining, and the learned embeddings are evaluated on popular gait recognition datasets.

**Table 2.** Comparison of gait datasets. UWG differs in purpuse, as it is intended for pretraining, and not for evaluation. It is a large-scale dataset, noisily annotated, agnostic to confounding factors and camera viewpoints. * Approximate number given by pose tracker.

| Dataset | # IDs | Views | Total Walk Length (hr) | Avg. Run Length (frames) | Runs/ID |
|---|---|---|---|---|---|
| CASIA-B | 124 | 11 | 15.8 | 100 | 110 |
| FVG | 226 | 3 | 3.2 | 97 | 12 |
| UWG (ours) | **38,502 *** | **1** | **59.0** | **108.5** | **1** |

### *3.2. Learning Procedure*

Figure 5 highlights the proposed methodology for learning informative gait representations from unconstrained scenarios. Walking sequences for each tracked person are obtained after processing the data from the video streams. A Spatio-Temporal Graph Convolutional Network (ST-GCN) [46] is employed to process the walking sequences, which was chosen due to its good results in the area of action recognition. Moreover, applying graph computation on skeletons allows modelling both the local interactions between joints and the global time variation between individual skeletons. A graph model was used for the implementation, as it is more appropriate to model the spatio-temporal relationships between joints compared to a simple LSTM network [26]. Moreover, the authors of [33] show that a ST-GCN can successfully be used for skeleton-based gait recognition. In this ST-GCN implementation, a standard $1x\Gamma$ 2D convolution is performed on a (C, V, T) feature map tensor, where $\Gamma$ represents the temporal window size, C is the number of channels, V is the number of vertices and T the number of frames. The resulting tensor is then multiplied with the normalized adjacency matrix $\mathbf{\Lambda}^{-\frac{1}{2}}(\boldsymbol{A}+\boldsymbol{I})\mathbf{\Lambda}^{-\frac{1}{2}}$. In our case, the adjacency matrix $\boldsymbol{A}$ is given by the COCO skeleton structure and the identity matrix $\boldsymbol{I}$ represents the self-connections. Moreover, learnable edge-importance weights between joints are implemented through a mask matrix $\boldsymbol{M}$ which is multiplied element-wise with the adjacency matrix:

$$f_{out} = \mathbf{\Lambda}^{-\frac{1}{2}}((\boldsymbol{A}+\boldsymbol{I})\otimes \boldsymbol{M})\mathbf{\Lambda}^{-\frac{1}{2}} f_{in}\boldsymbol{W} \tag{3}$$

where $\mathbf{\Lambda}^{-\frac{1}{2}} = \sum_j (A^{ij} + I^{ij})$.

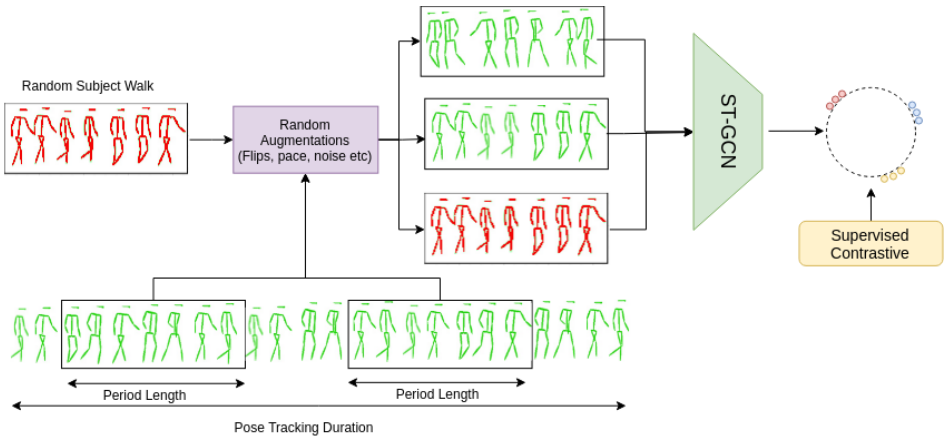


**Figure 5.** WildGait training methodology. Given a tracked sequence of skeletons, we employ Supervised Contrastive (SupCons) loss on multiple augmented views (randomly cropped, flipped etc.) of the same person and a walk from a random subject in the dataset as a negative sample.

Since the setting for the UWG dataset does not include multiple runs of the same walking person, with different confounding factors, a diverse set of data augmentations are employed to create augmented walking sequences for the same person. The network receives a randomly sampled gait sequence of 54 frames, out of the full tracked sequence. If

the tracked sequence is less than 54 frames, the start of the sequence is repeated. Moreover, the sequence is dilated or contracted, in accordance to pace prediction [47]. Modifying the pace of a video sequence has been shown to allow for learning meaningful semantic information in a self-supervised manner. The time modification factor is uniformly sampled from $\{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$. This procedure allows for the model to be robust to changes in video framerate and subject walking speed.

Further, squeezing, flipping, mirroring and randomly dropping out joints and frames are employed to further introduce diversity in the dataset.

Existing literature training procedures for recognition problems, employ either triplet loss [48], center loss [49], direct classification, or a combination of them [50]. However, manipulating the loss weights of multiple loss functions is cumbersome, are requires significant tuning. Moreover, having a direct classification head of the person identity does not scale well with the number of identities, and many models employ such a classification head for regularisation and preventing the triplet loss from collapsing embeddings. The use of triplet loss also involves careful hard negative mining [51], which can have a significant negative impact on performance if performed improperly.

To alleviate some of the optimization problems present in current approaches, we employed the Supervised Contrastive (SupCons) [17] loss, which is a generalization of the triplet loss, allowing for multiple positive examples per identity. Supervised Contrastive loss assumes a multiviewed batch, of multiple augmenations for the same sample. Let $i \in I \equiv \{1 \dots 2N\}$ be the index of an arbitrary augmentated sample and $j(i)$ the index of the other augmented sample from the same source. In supervised contrastive, a generalization of the self-supervised loss [52–54] is achieved by incorporating supervision, which allows for the presence of an arbitrary number of positives:

$$\mathcal{L}^{sup} = \sum_{i \in I} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)} \tag{4}$$

In this equation, $z_l = Proj(Enc(\widetilde{x}_l)$, the embedding of a skeleton sequence, $\cdot$ denotes the dot product operation, $\tau \in \mathcal{R}^+$ is a temperature parameter and $A(i) \equiv I \setminus \{i\}$. Moreover, $P(i) \equiv \{p \in A(i) : \widetilde{y}_p = \widetilde{y}_i\}$ is the set of indices of all positives in the multiviewed batch distinct from $i$. By using supervised contrastive instead of a traditional self-supervised loss enables an intrinsic ability to perform hard positive/negative mining, which is computationally expensive. Moreover, the authors highlight the increased contrastive power with the presence of more negative examples in the batch and the ability to generalize to an arbitrary number of positives. In our case, the positives are given by different crops across time of the same skeleton sequence. In our implementation, we chose two crops corresponding to two positive examples. The variability of the two crops is higher if the skeleton sequence is longer, as the walker might change direction or perform different actions while walking.

Figure 5 showcases our simplified learning procedure. After the feature extraction, embeddings are normalized to the unit sphere. Following the author's recommendations, we employ a loss temperature of 0.01, as smaller temperatures benefits training [17].

By not using a direct classification head of the identities, as in previous gait recognition works, we are able to scale to tens of thousands of identities with the same model size. This was not a requirement when working with smaller datasets such as CASIA-B and FVG, but it is in the case of UWG, since it has 38k tracklets.

## 4. Experiments & Results

### *4.1. Benchmark Datasets*

Popular gait recognition datasets, CASIA-B and FVG (Front-View Gait), were chosen to evaluate our unsupervised pretraining scheme.

**CASIA-B.** We chose CASIA-B as it is a popular benchmark for evaluating the effect of viewpoint variation, which allows us to compare to other pose-based gait recognition

methods. CASIA-B has 124 identities captured from 11 different viewpoints. All subjects walk indoors, and their walk is captured with synchronized cameras. Each identity has three walking variations corresponding to normal walking (NM), clothing variation (CL) and carry bag (BG).

**Front-View Gait (FVG).** is used to further evaluate the front-view angle, across multiple confounding factors that are not present in CASIA-B. The 226 subjects from FVG are walking outdoors and are captured with a single camera. Besides normal walking and changing clothes, subjects also change their walking speed to be slower or faster than their normal cadence.

### *4.2. Evaluation Procedure*

The performance evaluations presented in this paper abide by the evaluation guidelines of each dataset. For CASIA-B, we use the first 62 identities for training and the final 62 for evaluation, and show the average recognition accuracy across all viewing angles, except when gallery and probe angles are the same. For FVG, we used 136 identities for training and the rest for testing, and report results for each evaluation protocol: Walk Speed (WS), Carrying Bag (CB), Changing Clothes (CL), Cluttered Background (CBG) and ALL. The gallery set is comprised of only run #2 for each identity from sessions 1 and 2, as the authors suggested [7].

### *4.3. Quantitative Evaluation*

#### 4.3.1. Direct Transfer Performance

We initially tested our network's capability to generalize to CASIA-B and FVG by pretraining on UWG and directly evaluating on CASIA-B/FVG, without actually training or fine-tuning on these datasets. We evaluated the direct transfer recognition accuracy using increasingly larger random samples from UWG to highlight the impact of the size of the pretraining dataset. Each experiment was run 5 times and the results were averaged to avoid a favorable configuration for our setting. The results in Figure 6 show that transfer accuracy on downstream tasks benefits from a larger size of the pretraining dataset. It is more evident in the case of FVG, since it has fewer viewpoint variations pertaining to each subject.

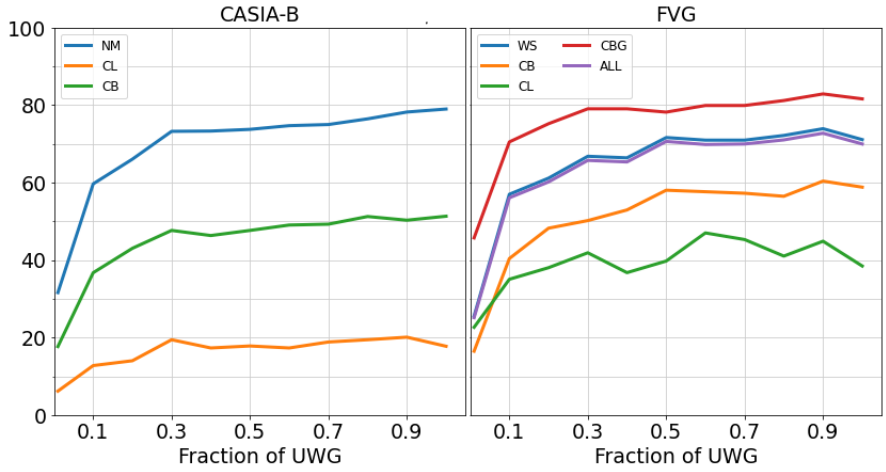


**Figure 6.** UWG dataset size influence over the transfer learning performance of WildGait on downstream evaluation benchmarks. The network was pretrained on UWG, but not trained on CASIA-B or FVG. For CASIA-B, we show mean accuracy where the gallery set contains all viewpoints except the probe angle.

#### 4.3.2. Supervised Fine-Tuning

Further, we evaluated the performance of our pretrained network when fine-tuned using limited amounts of training data, in a few-shot learning regime. For both CASIA-B

and FVG we used random samples of 10%, 20%, 30%, 50%, 70% and 100% of the runs of each person. Each experiment was run 5 times and the results were averaged. The entire network was trained with a high learning rate at deeper levels in the architecture and a decreasingly lower rate for the lower-level representations, as proposed by Kirkpatrick et al. [55].

The pretrained network was compared to a randomly initialized one, with results presented in Figure 7. On both datasets, the benefits of pretraining are showcased in the constant superior performance over random weight initialization, especially with lower amounts of training data. When fine-tuning with only 10% of the available data, the gap is more dramatic in the case of FVG—in FVG, each subjects walk is captured 12 times, whereas in CASIA-B, each subject is captured 110 times from all cameras. As such, 10% of the FVG data represents one walk per person, compared to 11 walks per person offered by 10% CASIA-B dataset. This is significantly less data per person used for fine-tuning, and highlights the benefits of pretraining the network on UWG. Moreover, the training regime for the pretrained network is more stable regardless of the amount of data, shown by the small standard deviations in Figure 7.

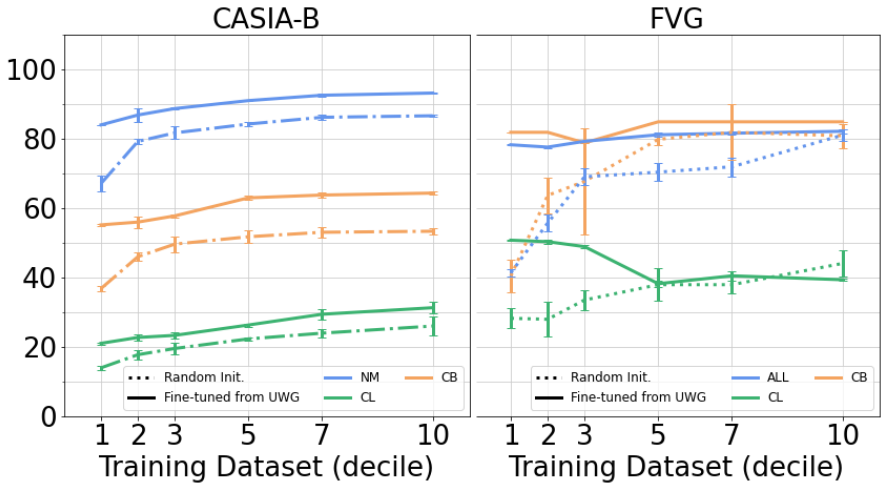


**Figure 7.** Performance of fine-tuning the proposed network on downstream evaluation benchmarks, with fractions of the training data. For FVG, runs are randomly sampled per subject. For CASIA-B, runs are randomly sampled uniformly from all angles per subject, and the average accuracy is across all viewpoints, where the gallery contains all angles except the probe angle. Pretraining on UWG results in a more stable training regime and significantly increased performance in scenarios with little labelled training data available.

#### 4.3.3. Comparison with Unsupervised Skeleton-Based Methods

We compared WildGait to other relevant methods that leverage skeleton sequences to learn meaningful representations. A ST-GCN pretrained on Kinetics [56] was selected to evaluate the transfer learning capabilities from supervised action recognition to gait recognition. Kinetics is a popular dataset of training and evaluating action recognition models, including skeleton-based models. Since it offers a large variety of actions and movements from skeletons, we used it to measure its performance in the particular case of gait analysis. We used the publicly available pretrained model provided by the authors [39]. Self-supervised approaches such as MS$^2$L [38] and Pace Prediction [47] were also chosen for comparison, along with a popular method for unsupervised pretraining in the field of skeleton-based action recognition, Predict and Cluster [36]. This latter method uses a sequence-to-sequence LSTM network with fixed decoder to learn discriminative representations. We followed the authors' implementation and pretrain on UWG. For Pace Prediction [47] and MS$^2$L [38], we implemented the methods according to their respective papers and pretrain the models on UWG. The results for direct transfer learning (without fine-tuning) on CASIA-B and FVG are presented in Tables 3 and 4, and show that WildGait outperforms existing approaches by a large margin. Our results show that, in the case

of gait recognition, the information captured in tracked skeleton sequences of walking people is sufficient for a strong supervisory signal, while plain unsupervised approaches are unsuitable. Moreover, our results show that transferring knowledge from the action recognition domain is inappropriate for gait analysis: pretraining on skeleton-based action recognition datasets does not lead to meaningful representations for particular instances of walks, nor are the algorithms developed in this domain suitable for gait recognition.

**Table 3.** Transfer learning comparison with other unsupervised skeleton-based training methods on CASIA-B. WildGait consistently outperforms standard methods for unsupervised learning from skeleton sequences. We report accuracy where the gallery set contains all viewpoints except the probe angle (top row). We highlight with bold our method (WildGait) and best results in each walking variation.

| | | **CASIA-B—Direct Transfer** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0°** | **18°** | **36°** | **54°** | **72°** | **90°** | **108°** | **126°** | **144°** | **162°** | **180°** | **Mean** |
| NM | Pretrained Kinetics | 19.35 | 19.35 | 27.42 | 29.03 | 25.81 | 38.71 | 27.42 | 27.42 | 20.97 | 12.9 | 6.45 | 23.17 |
| | Predict & Cluster | 41.93 | 45.16 | 45.96 | 34.67 | 20.16 | 11.29 | 30.64 | 32.25 | 21.77 | 19.35 | 12.09 | 28.66 |
| | $MS^2L$ | 37.90 | 39.51 | 40.32 | 51.61 | 24.19 | 17.74 | 25.80 | 46.74 | 40.32 | 33.06 | 34.67 | 35.62 |
| | Pace Prediction | 43.34 | 39.51 | 50.00 | 54.03 | 38.70 | 62.90 | **70.96** | 70.16 | 54.03 | 66.93 | 64.51 | 55.91 |
| | WildGait (ours) | **72.58** | **84.67** | **90.32** | **83.87** | **63.70** | **62.90** | 66.12 | **83.06** | **86.29** | **84.67** | **83.06** | **78.29** |
| CL | Pretrained Kinetics | 9.68 | 8.87 | 16.94 | **19.35** | 6.45 | 12.1 | 16.13 | 15.32 | 7.26 | 4.03 | 4.84 | 11.0 |
| | Predict & Cluster | 15.32 | 19.35 | 25.81 | 16.13 | 6.45 | 4.03 | 17.74 | 17.74 | 8.06 | 8.06 | 7.26 | 13.27 |
| | $MS^2L$ | 10.48 | 17.74 | 13.71 | 12.1 | 6.45 | 12.9 | 13.71 | 18.55 | 13.71 | 8.06 | 9.68 | 12.46 |
| | Pace Prediction | **13.71** | 10.48 | 8.87 | 8.06 | 11.29 | **13.71** | 16.94 | 17.74 | 12.9 | 13.71 | 15.32 | 12.98 |
| | WildGait (ours) | 12.1 | **33.06** | **25.81** | 18.55 | **12.9** | 11.29 | **21.77** | **24.19** | **20.16** | **26.61** | **16.13** | **20.23** |
| CB | Pretrained Kinetics | 17.74 | 15.32 | 12.9 | 14.52 | 18.55 | 12.9 | 15.32 | 17.74 | 14.52 | 8.87 | 8.06 | 14.22 |
| | Predict & Cluster | 24.19 | 34.68 | 27.42 | 26.61 | 10.48 | 9.68 | 17.74 | 20.97 | 11.29 | 15.32 | 12.1 | 19.13 |
| | $MS^2L$ | 25.0 | 33.87 | 31.45 | 26.61 | 11.29 | 16.13 | 20.97 | 27.42 | 25.81 | 20.97 | 20.16 | 23.61 |
| | Pace Prediction | 37.1 | 29.03 | 37.1 | 31.45 | 31.45 | **43.55** | **42.74** | 34.68 | 31.45 | 33.06 | 34.68 | 35.12 |
| | WildGait (ours) | **67.74** | **60.48** | **63.71** | **51.61** | **47.58** | 39.52 | 41.13 | **50.0** | **52.42** | **51.61** | **42.74** | **51.69** |

**Table 4.** Transfer learning comparison with other unsupervised skeleton-based training methods on FVG dataset. We highlight with bold our method (WildGait) and best results in each walking variation.

| | **FVG** | | | | |
|---|---|---|---|---|---|
| | **WS** | **CB** | **CL** | **CBG** | **ALL** |
| Pretrained Kinetics | 24.00 | 54.55 | 28.63 | 43.16 | 22.33 |
| Predict & Cluster | 32.79 | 33.72 | 20.08 | 44.01 | 32.40 |
| $MS^2L$ | 42.33 | 40.78 | 31.62 | 53.84 | 41.93 |
| Pace Prediction | 45.65 | 40.78 | 28.63 | 55.55 | 44.84 |
| WildGait (ours) | **75.66** | **81.81** | **48.71** | **84.61** | **75.66** |

#### 4.3.4. Comparison with State-of-the-Art

Finally, we compared with state-of-the-art skeleton-based gait recognition methods on CASIA-B, with the results presented in Table 5. We fine-tuned our network using all the available training data: 62 subjects, all viewpoints and runs. We achieve state-of-the-art results in skeleton-based gait recognition on normal walking (NM) and carry bag variations (BG) by a significant margin. When handling clothing (CL) variation, our method achieves good results relative to other methods, but clothing variation remains a challenging problem for gait-based person identification using skeletons, as heavy clothing significantly affects the manner of walking and also makes certain joints less visible to the pose estimation model. Moreover, UWG, by design, does not contain walking sequences of the same subject with different confounding factors (such as clothing variation). As such,

complete disentanglement is cumbersome in our setting. This is noticeable in Figure 7, in the case of FVG, where the pretrained model is negatively affected on the clothing variation through the addition of more data, while other variations are more stable.

Our state-of-the-art results are attributed to the large pretraining dataset and the diverse augmentation procedures we employ to make the model invariant to different walking variations and camera viewpoints.

**Table 5.** Comparison with other skeleton-based gait recognition methods on CASIA-B dataset. In all methods the gallery set contains all viewpoints except the proble angle. WildGait achieves state-of-the art results in normal walking (NM) and carry-bag variation (CB) by a large margin, being able to generalize well across camera viewpoints. We highlight with bold our network (WildGait) and best results in each walking variation.

| Probe | Method | 0° | 18° | 36° | 54° | 72° | 90° | 108° | 126° | 144° | 162° | 180° | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NM | PTSN [28] | 34.5 | 45.6 | 49.6 | 51.3 | 52.7 | 52.3 | 53 | 50.8 | 52.2 | 48.3 | 31.4 | 47.4 |
| | PTSN-3D [32] | 38.7 | 50.2 | 55.9 | 56 | 56.7 | 54.6 | 54.8 | 56 | 54.1 | 52.4 | 40.2 | 51.9 |
| | PoseGait [31] | 48.5 | 62.7 | 66.6 | 66.2 | 61.9 | 59.8 | 63.6 | 65.7 | 66 | 58 | 46.5 | 60.5 |
| | PoseFrame [30] | 66.9 | 90.3 | 91.1 | 55.6 | 89.5 | **97.6** | **98.4** | 97.6 | 89.5 | 69.4 | 68.5 | 83.1 |
| | WildGait network (ours) | **86.3** | **96.0** | **97.6** | **94.3** | **92.7** | 94.3 | 94.3 | **98.4** | **97.6** | **91.1** | **83.8** | **93.4** |
| CL | PTSN [28] | 14.2 | 17.1 | 17.6 | 19.3 | 19.5 | 20 | 20.1 | 17.3 | 16.5 | 18.1 | 14 | 17.6 |
| | PTSN-3D [32] | 15.8 | 17.2 | 19.9 | 20 | 22.3 | 24.3 | 28.1 | 23.8 | 20.9 | 23 | 17 | 21.1 |
| | PoseGait [31] | 21.3 | 28.2 | 34.7 | 33.8 | 33.8 | 34.9 | 31 | 31 | **32.7** | 26.3 | 19.7 | 29.8 |
| | PoseFrame [30] | 13.7 | 29.0 | 20.2 | 19.4 | **28.2** | **53.2** | **57.3** | **52.4** | 25.8 | 26.6 | 21.0 | 31.5 |
| | WildGait network (ours) | **29.0** | **32.2** | **35.5** | **40.3** | 26.6 | 25.0 | 38.7 | 38.7 | 31.4 | **34.6** | **31.4** | **33.0** |
| BG | PTSN [28] | 22.4 | 29.8 | 29.6 | 29.2 | 32.5 | 31.5 | 32.1 | 31 | 27.3 | 28.1 | 18.2 | 28.3 |
| | PTSN-3D [32] | 27.7 | 32.7 | 37.4 | 35 | 37.1 | 37.5 | 37.7 | 36.9 | 33.8 | 31.8 | 27 | 34.1 |
| | PoseGait [31] | 29.1 | 39.8 | 46.5 | 46.8 | 42.7 | 42.2 | 42.7 | 42.2 | 42.3 | 35.2 | 26.7 | 39.6 |
| | PoseFrame [30] | 45.2 | 66.1 | 60.5 | 42.7 | **58.1** | **84.7** | **79.8** | **82.3** | **65.3** | 54.0 | 50.0 | 62.6 |
| | WildGait network (ours) | **66.1** | **70.1** | **72.6** | **65.3** | 56.4 | 64.5 | 65.3 | 67.7 | 57.2 | **66.1** | **52.4** | **64.0** |

### *4.4. Qualitative Evaluation*

Further, to better understand the behaviour of our model, we make a visualization of the embeddings provided by the model pretrained on UWG of walks from CASIA-B. Figure 8 shows the t-SNE [57] visualization of the test set from CASIA-B. t-SNE has stood the test of time with regards to the visualization power of high dimensional datasets, and can give us important information towards the model performance. As is the case in the numerical evaluation, of interest is the network ability to generalize across viewpoints. We color the same plot two different ways: color by tracking ID—each color is a different person—and by camera viewpoint—each color is a different viewpoint. It is evident that the model clearly clusters walking sequences pertaining to the same subject, regardless of camera viewpoint.

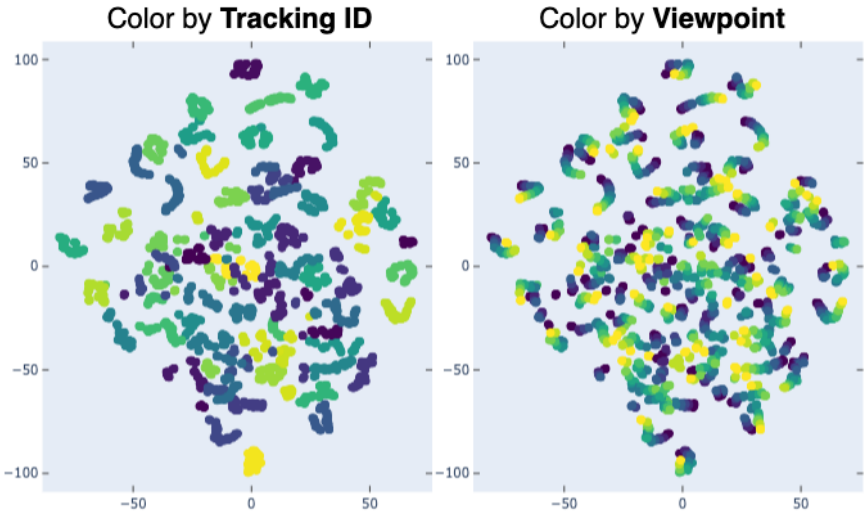


**Figure 8.** t-SNE [57] visualization of the embeddings from the evaluation set from CASIA-B, as outputted by WildGait, colored by tracking id (**left**) and camera viewpoint (**right**). It is clear that sequences pertaining to the same IDs are close to each other, irrespective of viewpoints.

## 5. Conclusions

This work presents a self-supervised framework, WildGait, for learning informative gait representations from unconstrained environments, without direct human supervision. We show that large amounts of video data such as surveillance streams can be leveraged by automatically annotating, filtering and processing walking people to learn discriminative embeddings that generalize well to new individuals, with good disentanglement of confounding factors. We leverage state-of-the-art pose estimation and pose tracking methods to construct Unconstrained Wild Gait (UWG), a large dataset of anonymized skeleton sequences. As far as we know, we are among the first to study pretraining in the context of gait recognition from raw video. Through pretraining on UWG and fine-tuning on downstream recognition tasks, we achieve state-of-the-art results in skeleton-based gait recognition on the CASIA-B benchmarking dataset.

The accuracy of pose-based gait recognition methods is highly dependent on the quality of extracted poses. Large and accurate pose-estimation models come with heavy computational burdens on the processing pipeline, especially in crowded scenes. This suggests a trade-off between accuracy and computational demand/inference time of model-based approaches. We are partly addressing this limitation by releasing the UWG dataset for the pretraining stage, containing over 38K extracted walking skeleton sequences. To raise the accuracy for clothing variation scenarios, we aim to further improve our data collection pipeline to include information regarding different clothing for the same person. This implies collecting the same scene (e.g., an office buildings entrance hallway) over a long period of time (ideally a year to include clothing changes due to seasonal variation), and combining information from both facial identification (e.g., FaceNet [48]) and person attribute identification models (i.e., HydraNet [58]) for a richer set of automatic annotations.

One of the main concerns in regards to the broader impact of biometrics-based human identification is privacy. Making use of skeletons for gait recognition softens this concern by relying solely on motion information of people walking, and no identifiable appearance-based information. Furthermore, pose estimation approaches are constantly advancing in terms of performance and efficiency, aiming for real-time inference with negligible to no accuracy loss and requiring less computational resources. This enables pushing skeleton-extraction computation to edge devices, removing the need to upload videos for cloud processing.

**Author Contributions:** Conceptualization, A.C. and I.E.R.; Methodology, Software, A.C.; Validation, I.E.R.; Formal Analysis, Investigation, A.C.; Resources, I.E.R.; Data Curation, A.C.; Writing—original draft preparation, A.C. and I.E.R.; Writing—review and editing, A.C. and I.E.R.; Visualization, A.C.; Supervision, I.E.R.; Project Administration, I.E.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partly supported by CRC Research Grant 2021, with funds from UEFISCDI in project CORNET (PN-III 1/2018).

**Data Availability Statement:** Data available on request.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Islam, T.U.; Awasthi, L.K.; Garg, U. Gender and Age Estimation from Gait: A Review. In Proceedings of the International Conference on Innovative Computing and Communications, New Delhi, India, 21–23 February 2020; Gupta, D., Khanna, A., Bhattacharyya, S., Hassanien, A.E., Anand, S., Jaiswal, A., Eds.; Springer: Singapore, 2021; pp. 947–962.
2. Randhavane, T.; Bhattacharya, U.; Kapsaskis, K.; Gray, K.; Bera, A.; Manocha, D. Identifying Emotions from Walking using Affective and Deep Features. *arXiv* **2020**, arXiv:1906.11884.
3. Ancillao, A. *Modern Functional Evaluation Methods for Muscle Strength and Gait Analysis*; Springer International Publishing: Cham, Switzerland, 2018. [CrossRef]
4. An, W.; Yu, S.; Makihara, Y.; Wu, X.; Xu, C.; Yu, Y.; Liao, R.; Yagi, Y. Performance Evaluation of Model-based Gait on Multi-view Very Large Population Database with Pose Sequences. *IEEE Trans. Biom. Behav. Identity Sci.* **2020**, *2*, 421–430. [CrossRef]
5. Shiqi, Y.; Tan, D.; Tan, T. A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; Volume 4, pp. 441–444. [CrossRef]
6. Hofmann, M.; Geiger, J.; Bachmann, S.; Schuller, B.; Rigoll, G. The TUM Gait from Audio, Image and Depth (GAID) database: Multimodal recognition of subjects and traits. *J. Vis. Commun. Image Represent.* **2014**, *25*, 195–206. [CrossRef]
7. Zhang, Z.; Tran, L.; Yin, X.; Atoum, Y.; Wan, J.; Wang, N.; Liu, X. Gait Recognition via Disentangled Representation Learning. In Proceeding of IEEE Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
8. Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.S.; Lu, C. CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark. *arXiv* **2018**, arXiv:1812.00324.
9. Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; Sheikh, Y.A. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
10. Rogez, G.; Weinzaepfel, P.; Schmid, C. LCR-Net++: Multi-Person 2D and 3D Pose Detection in Natural Images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 1146–1161. [CrossRef] [PubMed]
11. Doersch, C.; Zisserman, A. Multi-task Self-Supervised Visual Learning. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2070–2079. [CrossRef]
12. Gidaris, S.; Singh, P.; Komodakis, N. Unsupervised representation learning by predicting image rotations. *arXiv* **2018**, arXiv:1803.07728.
13. Noroozi, M.; Favaro, P. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 69–84.
14. Caron, M.; Bojanowski, P.; Joulin, A.; Douze, M. Deep clustering for unsupervised learning of visual features. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 132–149.
15. Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv* **2020**, arXiv:2006.09882.
16. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. *arXiv* **2021**, arXiv:2103.00020.
17. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised Contrastive Learning. *arXiv* **2020**, arXiv:2004.11362.
18. Han, J.; Bhanu, B. Individual recognition using gait energy image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 316–322. [CrossRef]
19. Choi, S.; Kim, J.; Kim, W.; Kim, C. Skeleton-Based Gait Recognition via Robust Frame-Level Matching. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 2577–2592. [CrossRef]
20. Sprager, S.; Juric, M. Inertial Sensor-Based Gait Recognition: A Review. *Sensors* **2015**, *15*, 22089–22127. [CrossRef]
21. Zeng, X.; Zhang, X.; Yang, S.; Shi, Z.; Chi, C. Gait-Based Implicit Authentication Using Edge Computing and Deep Learning for Mobile Devices. *Sensors* **2021**, *21*, 4592. [CrossRef] [PubMed]

22. Bashir, K.; Xiang, T.; Gong, S. Gait recognition using gait entropy image. In Proceedings of the 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009), London, UK, 3 December 2009.
23. Lam, T.H.; Cheung, K.H.; Liu, J.N. Gait flow image: A silhouette-based gait representation for human identification. *Pattern Recognit.* **2011**, *44*, 973–987. [CrossRef]
24. Wang, C.; Zhang, J.; Pu, J.; Yuan, X.; Wang, L. Chrono-Gait Image: A Novel Temporal Template for Gait Recognition. In *Proceedings of the Computer Vision—ECCV 2010, Crete, Greece, 5–11 September 2010*; Daniilidis, K., Maragos, P., Paragios, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 257–270.
25. Chen, Y.; Tian, Y.; He, M. Monocular human pose estimation: A survey of deep learning-based methods. *Comput. Vis. Image Underst.* **2020**, *192*, 102897. [CrossRef]
26. Feng, Y.; Li, Y.; Luo, J. Learning effective Gait features using LSTM. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancún, Mexico, 4–8 December 2016; pp. 325–330. [CrossRef]
27. Hochreiter, S.; Schmidhuber, J. Long Short-term Memory. *Neural Comput.* **1997**, *9*, 1735–80. [CrossRef] [PubMed]
28. Liao, R.; Cao, C.; Garcia, E.B.; Yu, S.; Huang, Y. Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations. In *Proceedings of the Chinese Conference on Biometric Recognition, Shenzhen, China, 28–29 October 2017*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 474–483.
29. Sheng, W.; Li, X. Siamese denoising autoencoders for joints trajectories reconstruction and robust gait recognition. *Neurocomputing* **2020**, *395*, 86–94. doi: 10.1016/j.neucom.2020.01.098. [CrossRef]
30. Lima, V.C.d.; Melo, V.H.C.; Schwartz, W.R. Simple and efficient pose-based gait recognition method for challenging environments. *Pattern Anal. Appl.* **2020**, *24*, 497–507. [CrossRef]
31. Liao, R.; Yu, S.; An, W.; Huang, Y. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognit.* **2020**, *98*, 107069. doi: 10.1016/j.patcog.2019.107069. [CrossRef]
32. An, W.; Liao, R.; Yu, S.; Huang, Y.; Yuen, P.C. Improving Gait Recognition with 3D Pose Estimation. In *Biometric Recognition*; Zhou, J., Wang, Y., Sun, Z., Jia, Z., Feng, J., Shan, S., Ubul, K., Guo, Z., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 137–147.
33. Li, N.; Zhao, X.; Ma, C. JointsGait:A model-based Gait Recognition Method based on Gait Graph Convolutional Networks and Joints Relationship Pyramid Mapping. *arXiv* **2020**, arXiv:2005.08625.
34. Chen, X.; Weng, J.; Lu, W.; Xu, J. Multi-Gait Recognition Based on Attribute Discovery. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1697–1710. [CrossRef]
35. Makihara, Y.; Matovski, D.; Carter, J.; Yagi, Y. Gait Recognition: Databases, Representations, and Applications. In *Computer Vision*; Springer: Cham, Switzerland, 2015. [CrossRef]
36. Su, K.; Liu, X.; Shlizerman, E. Predict & cluster: Unsupervised skeleton based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9631–9640.
37. Li, J.; Shlizerman, E. Iterate & Cluster: Iterative Semi-Supervised Action Recognition. *arXiv* **2020**, arXiv:2006.06911.
38. Lin, L.; Song, S.; Yang, W.; Liu, J. MS$^2$L: Multi-Task Self-Supervised Learning for Skeleton Based Action Recognition. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020.
39. Yang, Z.; Li, Y.; Yang, J.; Luo, J. Action Recognition with Spatio-Temporal Visual Attention on Skeleton Image Sequences. *arXiv* **2018**, arXiv:cs.CV/1801.10304.
40. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016. [CrossRef]
41. Hendrycks, D.; Mazeika, M.; Dietterich, T. Deep Anomaly Detection with Outlier Exposure. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
42. Al-Obaidi, S.; Wall, J.C.; Al-Yaqoub, A.; Al-Ghanim, M. Basic gait parameters: A comparison of reference data for normal subjects 20 to 29 years of age from Kuwait and Scandinavia. *J. Rehabil. Res. Dev.* **2003**, *40*, 361. [CrossRef] [PubMed]
43. Wojke, N.; Bewley, A.; Paulus, D. Simple Online and Realtime Tracking with a Deep Association Metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649. [CrossRef]
44. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 740–755.
45. Murray, M.P.; Drought, A.B.; Kory, R.C. Walking Patterns of Normal Men. *JBJS* **1964**, *46*, 335–360. [CrossRef]
46. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
47. Wang, J.; Jiao, J.; Liu, Y.H. Self-supervised video representation learning by pace prediction. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 504–521.
48. 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, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
49. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A Discriminative Feature Learning Approach for Deep Face Recognition. In *Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 499–515.

50. Luo, H.; Gu, Y.; Liao, X.; Lai, S.; Jiang, W. Bag of tricks and a strong baseline for deep person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
51. Xuan, H.; Stylianou, A.; Liu, X.; Pless, R. Hard negative examples are hard, but useful. In *Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 126–142.
52. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Montréal, QC, Canada, 6–8 July 2020; pp. 1597–1607.
53. Tian, Y.; Krishnan, D.; Isola, P. Contrastive multiview coding. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XI 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 776–794.
54. Weinberger, K.Q.; Saul, L.K. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* **2009**, *10*, 207–244.
55. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A.A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; others. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526. [CrossRef] [PubMed]
56. Zisserman, A.; Carreira, J.; Simonyan, K.; Kay, W.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; et al. The kinetics human action video datasets. *arXiv* **2017**, arXiv:1705.06950.
57. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
58. Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; Wang, X. Hydraplus-net: Attentive deep features for pedestrian analysis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 350–359.