WildGait: Learning Gait Representations from Raw Surveillance Streams

Adrian Cosma[*1] , **Emilian Radoi**[1] ,
[1]University Politehnica of Bucharest
cosma.i.adrian@gmail.com, emilian.radoi@cs.pub.ro

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 aim to address the hard 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 novel weakly supervised learning framework, WildGait, which consists of 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. Our results show that, with fine-tuning, we surpass in terms of recognition accuracy 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. We obtain an accuracy of 84.43% on CASIA-B and 71.3% on FVG, while using only 10% of the available training data. This consists of 29% and 38% accuracy improvement on the respective datasets when using the same network without pretraining.

Figure 1: Samples from the UWG dataset. We gather skeleton sequences automatically annotated from pose estimation and pose tracking models from publicly available surveillance streams. This allows for learning discriminative gait representations without explicit human labels and without the cooperation of subjects.

1 Introduction

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 [Islam *et al.*, 2021], estimate emotions [Randhavane *et al.*, 2020], and provide insight into various physiological conditions [Ancillao, 2018]. 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 wide 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.

This led the study of gait recognition to be mostly performed in controlled environments, in which only a subset of confounding factors are explored [Shiqi Yu *et al.*, 2006]. Current datasets [Shiqi Yu *et al.*, 2006; An *et al.*, 2020; Zhang *et al.*, 2019] 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

*Corresponding Author

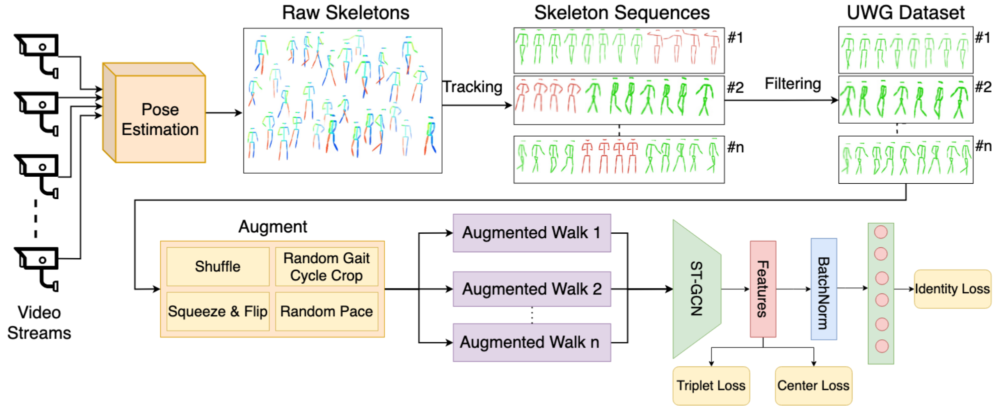

Figure 2: The WildGait framework. We process raw video streams from real-world settings of people walking and construct Uncooperative Wild Gait (UWG), a dataset of skeleton sequences loosely annotated through pose estimation and pose tracking. We use this raw and noisy data to pretrain a gait recognition model that generalizes well to different gait recognition scenarios.

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 [Li *et al.*, 2018]) to extract skeleton sequences and pose tracking to construct a loosely annotated dataset. Making use of augmentation procedures inspired from self-supervised learning, we pretrain our network with minimal direct supervision - the only indirect label employed is the pose tracking information uncovered automatically.

We chose to use pose estimations as they do not contain any identifiable appearance-based information of walking individuals. Current methods that use silhouettes [Han and Bhanu, 2006] are unsuitable in real-world, dynamic scenarios in which changes in illumination and multiple overlapping individuals severely affect the quality of extracted silhouettes. While some approaches explicitly disentangle appearance features and pose information [Zhang *et al.*, 2019], we argue that appearance-based methods do not respect 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 identity. Pose information also enables leveraging information of performed actions and activities, and filtering out individuals that are not walking or have abnormal walking patterns [Choi *et al.*, 2019].

We propose the Uncooperative Wild Gait dataset (UWG), which unlike current available datasets, contains anonymised skeleton sequences of a large number of people walking in a natural environment, 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.

This paper makes the following contributions:

1. We are among the first to explore unsupervised learning on gait recognition, and propose a novel framework, WildGait, which describes a data collection and pretraining pipeline, that enables learning meaningful gait representations in a weakly supervised manner, from automatically extracted skeleton sequences in unconstrained environments.
2. A study on transfer learning capabilities of our pretrained network on popular gait recognition databases, highlighting great data-efficiency when fine-tuning: we obtain an improvement in terms of recognition accuracy of 29% on CASIA-B and 38% on FVG, when using only 10% of the available training data.
3. The Uncooperative Wild Gait dataset (UWG), dataset of noisily tracked skeleton sequences to allow the gait recognition research community to further explore ways to pretrain gait recognition systems in an unsupervised manner.

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 [Sprager and Juric, 2015], 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. One of the most prevalent approaches for appearance-based gait recognition is the use of a Gait Energy Image (GEI) [Han and Bhanu, 2006]. 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) [Bashir *et al.*, 2009], Gait Flow Image [Lam *et al.*, 2011] and Chrono-Gait Image [Wang *et al.*, 2010], all showing good performance on benchmark datasets.

More recent approaches tend to use appearance features to explicitly learn a disentangled representation. [Zhang *et al.*, 2019] propose a way to explicitly disentangle motion and appearance features via an autoencoder with carefully designed loss functions.

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 [Chen *et al.*, 2020], enabled the use of skeletons sequences in gait recognition. [Yang Feng *et al.*, 2016] directly use the joint probability heatmaps as an input for an LSTM network. [Liao *et al.*, 2017] used an LSTM and a CNN to process 2D skeleton sequences to account for the temporal and spatial variations of walking. [Sheng and Li, 2020] used an LSTM autoencoder with contrastive learning to further stabilize the joint trajectories of skeletons. Recently, [Choi *et al.*, 2019] enhance the robustness of estimated skeletons by constructing a quality-adjusted cost matrix between input frames and registered frames for frame-level matching. [Lima *et al.*, 2020] propose a a fully-connected network to model a single skeleton, and temporal aggregation of skeletal features for the final classification. Both [Liao *et al.*, 2020] and [An *et al.*, 2018] leverage 3D skeletons to model gait patterns, with incremental improvements over 2D skeletons. Similar to us, [Li *et al.*, 2020] 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. [Chen *et al.*, 2018] proposed an attribute discovery model in a max-margin framework to recognize a person based on gait while walking with multiple people.

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. [Su *et al.*, 2019] proposed to use an LSTM autoencoder to learn discriminative representations for activity recognition, without any supervision except the skeleton sequences. Similarly, [Li and Shlizerman, 2020] used an iterative approach in an active learning setting. [Lin *et al.*, 2020] 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 action recognition models 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). To construct a supervisory signal, instead of relying only on skeleton sequences as in self-supervised approaches, we leverage the pose tracking ID and treat the problem as a gait recognition problem. We train an ST-GCN for gait recognition without any major architectural modifications, and obtain exceptional results in scenarios with scarce amounts of data.

3 Method

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 Uncooperative Wild Gait (UWG), and publicly released it for the research community to further advance the field of gait recognition.

Since people from different geographic locations have different manners of walking influenced by cultural and societal norms [Al-Obaidi *et al.*, 2003], we gathered data uniformly from 3 major continents (Europe, Asia and America). To obtain skeleton sequences for people in a video, a crowd pose estimation method (e.g. AlphaPose [Li *et al.*, 2018]) was used to extract human skeletons for each frame, and skeletons were tracked intra-camera, across time, with a simple Kalman filter [Bewley *et al.*, 2016].

Unstructured environments invariably introduce noise in the final skeleton sequences due to unreliable extracted skeletons, lost tracking information, or people standing who perform other activities 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 50 frames, which corresponds to approximately two full gait cycles [Murray *et al.*, 1964]. 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 half this length. 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. We address the issue of non-walking people with an heuristic on the movement of the joints corresponding to the legs (feet and knees). We compute the average gradient magnitude across time, which is indicative of the activity of a person. Thus, we filter out individuals with an average gradient magnitude of less than 0.01. Tracks longer than 10 seconds are also filtered out, as it was noticed that individuals tracked for a longer time are usually standing (not walking).

The proposed framework does not rely on appearance information at any step in the processing pipeline, except when extracting the pose information. A total of 18.212 identities were obtained, with an average walking duration of 110 frames. The total walking sequences duration in the dataset is of approximately 26 hours. 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 training, and the learned embeddings are evaluated on popular gait recognition datasets.

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**)	**18212***	**1**	26.1	110	**1**

Table 1: 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.

3.2 Learning Procedure

Figure 2 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) [Yan *et al.*, 2018] 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 [Yang Feng *et al.*, 2016]. Moreover, [Li *et al.*, 2020] show that a ST-GCN can successfully be used for skeleton-based gait recognition.

Since the setting for the UWG dataset does not include multiple runs of the same walking person, with different confounding factors, data augmentations are employed to create augmented walking sequences for the same person. The network receives a randomly sampled gait sequence of 72 frames, out of the full tracked sequence. If the tracked sequence is less than 72 frames, the start of the sequence is repeated. Moreover, the sequence is dilated or contracted, in accordance to pace prediction [Wang *et al.*, 2020]. 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 randomly sampled from $\{0.3, 0.5, 0.9, 1, 2, 2.5\}$ with weights of $\{0.1, 0.1, 0.1, 0.5, 0.1, 0.1\}$, setting higher weight to normal walking. This procedure allows for the model to be robust to changes in video framerate and subject walking speed.

Further, portions of the skeleton sequence are randomly shuffled to model temporal patterns, similarly to [Lin *et al.*, 2020]. Shuffling the order of the skeleton sequence introduces spatio-temporal ambiguity, that forces to model to learn invariant representations for use in downstream tasks.

Finally, squeezing and flipping is introduced by multiplying the X-axis with a number between (-1, $-1+\epsilon$), or just squeezing with a number between ($1-\epsilon$, 1), ϵ being randomly sampled from $[0, 0.15]$. Squeezing approximates a smaller rotation, but since we use 2D skeletons, perfect camera-invariant rotations are not possible, due to the depth ambiguity of the joints.

Note that we do not have a separate prediction head for any of our transformations. These procedures are only used to disentangle confounding factors related to gait (i.e. make the network invariant to walking speed). In [Dosovitskiy *et al.*, 2014], the authors heavily augment a set of images and task the network to identify the source image. Similarly, each intra-frame tracking pose ID in our scenario is its own class.

Advances in person re-identification [Luo *et al.*, 2019] inspired a combination of losses for our network. Triplet loss [Hermans *et al.*, 2017] is initially used with batch hard-triplet mining, with 16 different track IDs per batch and a margin of $\alpha = 0.3$:

$$\mathcal{L}_T = \max(||f(A) - f(P)||^2 - ||f(A) - f(N)||^2 + \alpha, 0) \quad (1)$$

A, P and N represent the anchor, positive input and negative input, respectively. Center loss [Wen *et al.*, 2016] is used to account for the downsides of triplet loss, such as ignoring the absolute difference between embeddings:

$$\mathcal{L}_C(X) = \frac{1}{2}\sum_{j=1}^{B} ||f(X)_j - c_{y_j}||_2^2 \quad (2)$$

c_y represents the center embedding for the input with track ID y. Further, regularization of the training regime is performed by directly predicting the track ID (i.e. the identity). This is achieved by a classification head with softmax activation, and applying cross-entropy loss. The final loss is a weighted combination of the three:

$$\mathcal{L} = \mathcal{L}_T + \lambda_1 \cdot \mathcal{L}_C + \lambda_2 \cdot \mathcal{L}_{ID} \quad (3)$$

A significant influence on the final accuracy of appearance-based person re-identification pipelines found by [Luo *et al.*, 2019] is due to the batch-normalization head. As such, instead of operating directly on the unnormalized final embedding vector (or unit-sphere projection), the output embeddings of the network are normalized through a batch-normalization layer before the identity prediction layer, as proposed by [Luo *et al.*, 2019]. This ensures that features are normally distributed and the triplet loss converges faster. We used an embedding size of 512 in all experiments and a batch size of 256. Since the UWG dataset is subject to noise, and track IDs are not a reliable source of information, a widely used technique is to employ label smoothing to further regularize the model, and penalize overconfident predictions. However, we did not observe any improvement in results by using this technique, and we chose not to use it in our final experiments.

4 Experiments and Results

Popular gait recognition datasets, CASIA-B and FVG, were chosen to evaluate our unsupervised pretraining scheme. 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.

We initially tested our network's capability to generalize to CASIA-B and FVG without actually training or fine-tuning on these datasets. We evaluated recognition accuracy through transfer learning using increasingly larger random samples from UWG (25%, 50%, 75% and 100%) 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 3 show that transfer learning 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.

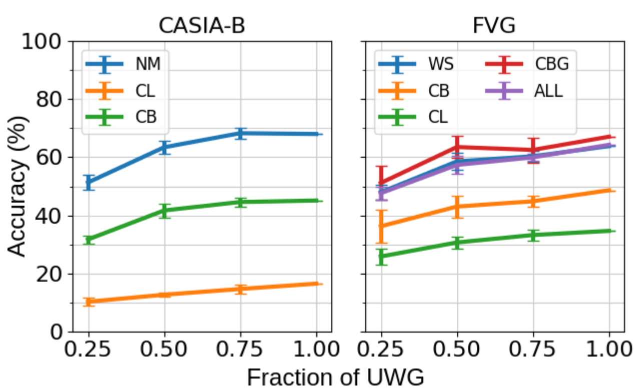

Figure 3: UWG dataset size influence over the transfer learning performance of WildGait on downstream evaluation benchmarks. The network was pretrained on UWG, but not trained on CASIA-B or FVG. For CASIA-B, we show mean accuracy where the gallery set contains all viewpoints except the probe angle.

Further, we evaluated the performance of our pretrained network when fine-tuned using limited amounts of training data. 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.*, 2016]. The pretrained network was compared to a randomly initialized one, with results presented in Figure 4. 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. Moreover, the training regime is more stable regardless of the amount of data, as evidenced by very small standard deviations.

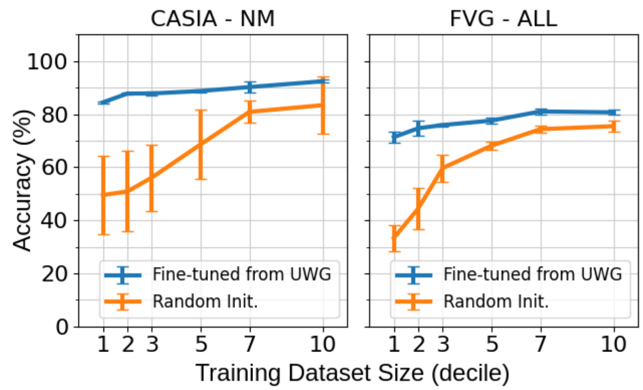

Figure 4: Performance of fine-tuning the proposed network on downstream evaluation benchmarks, with fractions of the training data. For FVG, runs are randomly sampled per subject. For CASIA-B, runs are randomly sampled uniformly from all angles per subject, and the average accuracy is across all viewpoints, 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.

We compared WildGait to other relevant methods that leverage skeleton sequences to learn meaningful representations. A ST-GCN pretrained on Kinetics [Kay *et al.*, 2017] was selected to evaluate the transfer learning capabilities from supervised action recognition to gait recognition. Self-supervised approaches such as MS^{2}L [Lin *et al.*, 2020] and Pace Prediction [Wang *et al.*, 2020] were also chosen for comparison, along with a popular method for unsupervised pretraining in the field of skeleton-based action recognition, Predict & Cluster [Su *et al.*, 2019]. 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. The results for direct transfer learning (without fine-tuning) on both CASIA-B and FVG are presented in Table 2, 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 or self-supervised approaches are unsuitable.

Finally, we compared with state-of-the-art skeleton-based gait recognition methods on CASIA-B, with the results presented in Table 3. 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 (CB) by a significant margin. While we have reasonable performance on clothing variation (CL) compared to other methods, [Li *et al.*, 2020] has exceptionally good results in this scenario. However, the results in Figure 3 show potential for improvement, especially in the case of CL, by further increasing the size of the UWG dataset. Our state-of-the-art results are attributed to the large pretraining dataset and the augmentation procedures we employ to make the model in-

	CASIA-B - Normal Walking												FVG				
	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean	WS	CB	CL	CBG	ALL
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	24.00	**54.55**	28.63	43.16	22.33
Predict & Cluster	32.26	42.74	44.35	31.45	29.03	48.39	44.35	43.55	29.03	16.94	16.94	34.46	33.47	35.69	20.51	36.32	34.13
MS^2L	32.26	36.29	47.58	41.13	21.77	44.35	47.58	49.19	37.9	23.39	19.35	36.44	42.88	38.43	27.78	49.57	40.87
Pace Prediction	42.34	63.31	70.56	56.05	52.02	62.5	62.1	**74.19**	**67.74**	37.5	45.97	57.66	44.15	37.84	25.00	51.28	43.72
WildGait **(ours)**	**58.06**	**74.19**	**83.06**	**74.19**	**60.48**	**75.81**	**74.19**	**74.19**	65.32	**54.03**	**54.84**	**68.04**	**63.84**	48.63	**34.62**	**67.09**	**64.29**

Table 2: Transfer learning comparison with other unsupervised skeleton-based training methods. For CASIA-B we report accuracy where the gallery set contains all viewpoints except the probe angle (top row).

Probe	Method	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
NM	PTSN [Liao *et al.*, 2017]	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 [An *et al.*, 2018]	38.7	50.2	55.9	56	56.7	54.6	54.8	56	54.1	52.4	40.2	51.9
	PoseGait [Liao *et al.*, 2020]	48.5	62.7	66.6	66.2	61.9	59.8	63.6	65.7	66	58	46.5	60.5
	JointsGait [Li *et al.*, 2020]	68.1	73.6	77.9	76.4	77.5	79.1	78.4	76	69.5	71.9	70.1	74.4
	PoseFrame [Lima *et al.*, 2020]	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)**	**84.6**	**96.7**	**95.1**	**98.3**	**95.1**	**98.3**	**99.1**	**100.0**	**95.1**	**90.3**	**70.9**	**93.1**
CL	PTSN [Liao *et al.*, 2017]	14.2	17.1	17.6	19.3	19.5	20	20.1	17.3	16.5	18.1	14	17.6
	PTSN-3D [An *et al.*, 2018]	15.8	17.2	19.9	20	22.3	24.3	28.1	23.8	20.9	23	17	21.1
	PoseGait [Liao *et al.*, 2020]	21.3	28.2	34.7	33.8	33.8	34.9	31	31	32.7	26.3	19.7	29.8
	JointsGait [Li *et al.*, 2020]	**48.1**	**46.9**	**49.6**	**50.5**	**51**	52.3	49	46	**48.7**	**53.6**	**52**	**49.8**
	PoseFrame [Lima *et al.*, 2020]	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)**	31.4	37.9	33.8	42.7	31.4	37.1	35.4	37.9	30.6	29.4	24.1	33.8
BG	PTSN [Liao *et al.*, 2017]	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 [An *et al.*, 2018]	27.7	32.7	37.4	35	37.1	37.5	37.7	36.9	33.8	31.8	27	34.1
	PoseGait [Liao *et al.*, 2020]	29.1	39.8	46.5	46.8	42.7	42.2	42.7	42.2	42.3	35.2	26.7	39.6
	JointsGait [Li *et al.*, 2020]	54.3	59.1	60.6	59.7	63	65.7	62.4	59	58.1	58.6	**50.1**	59.1
	PoseFrame [Lima *et al.*, 2020]	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)**	**68.5**	**79.8**	**80.6**	**76.6**	**76.6**	74.1	73.3	76.6	**70.1**	**68.5**	45.1	**71.8**

Table 3: Comparison with other skeleton-based gait recognition methods on CASIA-B dataset. The evaluation was performed in the "leave-one-out" setting, in which the gallery set contains all viewpoints except the one in the probe set. 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.

variant to different walking variations.

5 Conclusions

This work presents a novel weakly-supervised framework, WildGait, for learning informative gait representations from unconstrained environments, without direct human supervision. We show that we can leverage large amounts of video data, surveillance streams, 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 automatically construct 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.

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 suggest 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 18K extracted walking skeleton sequences.

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.

References

[Al-Obaidi *et al.*, 2003] Saud Al-Obaidi, James C. Wall, Alia Al-Yaqoub, and Muneera Al-Ghanim. Basic gait parameters: A comparison of reference data for normal subjects 20 to 29 years of age from kuwait and scandinavia. *The Journal of Rehabilitation Research and Development*, 40(4):361, 2003.

[An *et al.*, 2018] Weizhi An, Rijun Liao, Shiqi Yu, Yongzhen Huang, and Pong C. Yuen. Improving gait recognition with 3d pose estimation. In Jie Zhou, Yunhong Wang, Zhenan Sun, Zhenhong Jia, Jianjiang Feng, Shiguang Shan, Kurban Ubul, and Zhenhua Guo, editors, *Biometric Recognition*, pages 137–147, Cham, 2018. Springer International Publishing.

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

[Ancillao, 2018] Andrea Ancillao. *Modern Functional Evaluation Methods for Muscle Strength and Gait Analysis*. 01 2018.

[Bashir *et al.*, 2009] Khalid Bashir, Tao Xiang, and Shaogang Gong. Gait recognition using gait entropy image. IET, 2009.

[Bewley *et al.*, 2016] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. *2016 IEEE International Conference on Image Processing (ICIP)*, Sep 2016.

[Chen *et al.*, 2018] X. Chen, J. Weng, W. Lu, and J. Xu. Multi-gait recognition based on attribute discovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(7):1697–1710, 2018.

[Chen *et al.*, 2020] Yucheng Chen, Yingli Tian, and Mingyi He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192:102897, Mar 2020.

[Choi *et al.*, 2019] S. Choi, J. Kim, W. Kim, and C. Kim. Skeleton-based gait recognition via robust frame-level matching. *IEEE Transactions on Information Forensics and Security*, 14(10):2577–2592, 2019.

[Dosovitskiy *et al.*, 2014] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin A. Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *CoRR*, abs/1406.6909, 2014.

[Han and Bhanu, 2006] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(2):316–322, 2006.

[Hermans *et al.*, 2017] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification, 2017.

[Islam *et al.*, 2021] Tawqeer Ul Islam, Lalit Kumar Awasthi, and Urvashi Garg. Gender and age estimation from gait: A review. In Deepak Gupta, Ashish Khanna, Siddhartha Bhattacharyya, Aboul Ella Hassanien, Sameer Anand, and Ajay Jaiswal, editors, *International Conference on Innovative Computing and Communications*, pages 947–962, Singapore, 2021. Springer Singapore.

[Kay *et al.*, 2017] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[Kirkpatrick *et al.*, 2016] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016.

[Lam *et al.*, 2011] Toby HW Lam, King Hong Cheung, and James NK Liu. Gait flow image: A silhouette-based gait representation for human identification. *Pattern recognition*, 44(4):973–987, 2011.

[Li and Shlizerman, 2020] Jingyuan Li and Eli Shlizerman. Iterate & cluster: Iterative semi-supervised action recognition, 2020.

[Li *et al.*, 2018] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018.

[Li *et al.*, 2020] Na Li, Xinbo Zhao, and Chong Ma. Jointsgait:a model-based gait recognition method based on gait graph convolutional networks and joints relationship pyramid mapping, 2020.

[Liao *et al.*, 2017] Rijun Liao, Chunshui Cao, Edel B Garcia, Shiqi Yu, and Yongzhen Huang. Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In *Chinese Conference on Biometric Recognition*, pages 474–483. Springer, 2017.

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

[Lima *et al.*, 2020] Vítor C. de Lima, Victor H. C. Melo, and William R. Schwartz. Simple and efficient pose-based gait recognition method for challenging environments. *Pattern Analysis and Applications*, Nov 2020.

[Lin *et al.*, 2020] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. Ms^{2}l: Multi-task self-supervised learning for skeleton based action recognition. In *ACM Multimedia*, 2020.

[Luo *et al.*, 2019] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification, 2019.

[Murray *et al.*, 1964] M. Pat Murray, A. Bernard Drought, and Ross C. Kory. Walking patterns of normal men. *JBJS*, 46(2), 1964.

[Randhavane *et al.*, 2020] Tanmay Randhavane, Uttaran Bhattacharya, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha. Identifying emotions from walking using affective and deep features, 2020.

[Sheng and Li, 2020] Weijie Sheng and Xinde Li. Siamese denoising autoencoders for joints trajectories reconstruction and robust gait recognition. *Neurocomputing*, 395:86 – 94, 2020.

[Shiqi Yu *et al.*, 2006] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 4, pages 441–444, 2006.

[Sprager and Juric, 2015] Sebastijan Sprager and Matjaz Juric. Inertial sensor-based gait recognition: A review. *Sensors*, 15(9):22089–22127, September 2015.

[Su *et al.*, 2019] Kun Su, Xiulong Liu, and Eli Shlizerman. Predict & cluster: Unsupervised skeleton based action recognition, 2019.

[Wang *et al.*, 2010] Chen Wang, Junping Zhang, Jian Pu, Xiaoru Yuan, and Liang Wang. Chrono-gait image: A novel temporal template for gait recognition. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 257–270, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[Wang *et al.*, 2020] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction, 2020.

[Wen *et al.*, 2016] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 499–515, Cham, 2016. Springer International Publishing.

[Yan *et al.*, 2018] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition, 2018.

[Yang Feng *et al.*, 2016] Yang Feng, Yuncheng Li, and Jiebo Luo. Learning effective gait features using lstm. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 325–330, 2016.

[Zhang *et al.*, 2019] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Jian Wan, Nanxin Wang, and Xiaoming Liu. Gait recognition via disentangled representation learning. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Long Beach, CA, June 2019.