

Gait Recognition in the Wild: A Benchmark

Zheng Zhu^{1*} Xianda Guo² Tian Yang² Junjie Huang²
Jiankang Deng³ Guan Huang² Dalong Du² Jiwen Lu^{1y} Jie Zhou¹
¹Tsinghua University ²XForwardAI ³Imperial College London

fzhengzhu, lujijweng@tsinghua.edu.cn fxianda.guo, guan.huang, dalong.du@xforwardai.com

Abstract

Gait benchmarks empower the research community to train and evaluate high-performance gait recognition systems. Even though growing efforts have been devoted to cross-view recognition, academia is restricted by current existing databases captured in the controlled environment. In this paper, we contribute a new benchmark for Gait REcognition in the Wild (GREW). The GREW dataset is constructed from natural videos, which contains hundreds of cameras and thousands of hours streams in open systems. With tremendous manual annotations, the GREW consists of 26K identities and 128K sequences with rich attributes for unconstrained gait recognition. Moreover, we add a distractor set of over 233K sequences, making it more suitable for real-world applications. Compared with prevailing predefined cross-view datasets, the GREW has diverse and practical view variations, as well as more natural challenging factors. To the best of our knowledge, this is the first large-scale dataset for gait recognition in the wild. Equipped with this benchmark, we dissect the unconstrained gait recognition problem. Representative appearance-based and model-based methods are explored, and comprehensive baselines are established. Experimental results show (1) The proposed GREW benchmark is necessary for training and evaluating gait recognizer in the wild. (2) For state-of-the-art gait recognition approaches, there is a lot of room for improvement. (3) The GREW benchmark can be used as effective pre-training for controlled gait recognition. Benchmark website is <https://www.grew-benchmark.org/>.

1. Introduction

Gait recognition aims to identify a person according to his/her walking style in a video. Compared with face, fingerprint, iris and palmprint, gait is hard to disguise and can work at a long distance, giving it unique potential for crime prevention, forensic identification, and social security.

Figure 1: Examples comparison for CASIA-B [74], OU-MVLP [51] and the proposed GREW. The first two are captured under constrained environments, while the GREW is constructed in the wild. Since OU-MVLP [51] does not release RGB data, visualization results from its original paper are adopted. Faces are masked in the GREW for privacy concern.

Recognizing gait under a controlled environment has achieved significant progress due to the boom of deep learning. The essential engines of recent gait recognition consist of network architecture evolution [20, 9, 62, 65, 16, 72, 71, 31, 44, 50, 4, 63, 67, 39], loss function design [78, 17, 75, 79], and growing gait benchmarks [42, 7, 74, 37, 51, 23]. Even though gait recognition has achieved impressive advance in past years and it possesses the unique advantage of long-distance recognition, this technique has not yet been widely deployed in real-world applications. A notable obstacle is that there is almost no public benchmark to train and evaluate gait recognizer in the wild.

To our knowledge, most gait datasets are captured in relatively fixed and constrained environments such as laboratory or static outdoors. CASIA-B [74] and OU-MVLP [51] are most popularly used datasets in recent gait recognition research as shown in Figure 1. CASIA-B contains 124 subjects and 13,640 sequences, which is constructed in 2006. OU-MVLP consists of 10,307 identities and 288,596 walking videos, making it a big gait dataset with respect to #subjects. The statistics of more datasets are shown

^{*}These authors contributed equally to this work.

^yCorresponding author.

Table 1: Comparison of the GREW with existing gait recognition datasets regarding statistics, data type, captured environment, view variations and challenging factors. Datasets are sorted in publication time. *#Id.*, *#Seq.* and *#Cam.* refer to numbers of identities, sequences and cameras. *Sil.*, *Inf.*, *D.* and *A.* mean silhouette, infrared, depth and audio. *VI*, *DIS*, *BA*, *CA*, *DR*, *OCC*, *ILL*, *SU*, *SP*, *SH*, and *WD* are abbreviations of view, distractor, background, carrying, dressing, occlusion, illumination, surface, speed, shoes, and walking directions.

Dataset	Publication	#Id.	#Seq.	#Cam.	Data types	# Distractor	Environment	View var.	Challenges
CMU MoBo [12]	TR2001	25	600	6	RGB, Sil.	None	Controlled	Predefined	VI, CA, SP, SU
CASIA-A [57]	TPAMI2003	20	240	3	RGB	None	Controlled	Predefined	VI
SOTON [45]	ASSC2004	115	2,128	2	RGB, Sil.	None	Controlled	Predefined	VI
USF [42]	TPAMI2005	122	1,870	2	RGB	None	Controlled	Predefined	VI, CA, SU, SH
CASIA-B [74]	ICPR2006	124	13,640	11	RGB, Sil.	None	Controlled	Predefined	VI, CA, DR
CASIA-C [52]	ICPR2006	153	1,530	1	Inf., Sil.	None	Controlled	None	CA, SP
OU-ISIR Speed [54]	CVPR2010	34	612	1	Sil.	None	Controlled	None	SP
OU-ISIR Cloth [19]	PR2010	68	2,764	1	Sil.	None	Controlled	None	DR
OU-ISIR MV [38]	ACCV2010	168	4,200	25	Sil.	None	Controlled	Predefined	VI
OU-LP [23]	TIFS2012	4,007	7,842	2	Sil.	None	Controlled	Predefined	VI
ADSC-AWD [35]	TIFS2014	20	80	1	Sil.	None	Controlled	None	WD
TUM GAID [18]	JVCIR2014	305	3,370	1	RGB, D., A.	None	Controlled	None	CA, SH
OU-LP Age [68]	CVA2017	63,846	63,846	1	Sil.	None	Controlled	None	Age
OU-MVLP [51]	CVA2018	10,307	288,596	14	Sil.	None	Controlled	Predefined	VI
OU-LP Bag [55]	CVA2018	62,528	187,584	1	Sil.	None	Controlled	None	CA
OU-MVLP Pose [2]	TBIOM2020	10,307	288,596	14	2D Pose	None	Controlled	Predefined	VI
GREW	-	26,345	128,671	882	Sil. Flow 2/3D Pose	233,857	Wild	Diverse	VI, DIS, BA, CA, DR, OCC, ILL, SU

in Table 1, which are mainly constructed under controlled settings and designed for predefined cross-view gait recognition. However, in real scenarios, gait recognition would encounter fully-unconstrained challenges, such as diverse view, occlusion, various carrying and dressing, complex and dynamic background clutters, illumination, walking style, surface influence *et al.* Existing benchmarks are far behind the requirements of practical gait recognition. Considering the remarkable success of face recognition [49, 43, 56, 8, 21, 70, 3, 13, 24, 84] and person re-identification (ReID) [77, 48, 36, 66, 17, 10, 82, 80, 81, 83, 27, 61], it is time to move to benchmark gait recognition in the wild.

In this paper, we present the Gait REcognition in the Wild (GREW) benchmark, which is the first work delving into this open problem to the best of our knowledge. The GREW dataset is constructed from natural streams with multiple cameras as shown in Figure 1. Identity information from raw videos is manually annotated, resulting in 26K subjects, 128K sequences and 14M boxes for unconstrained gait recognition. Besides, rich human attributes including gender, age group, carrying and dressing styles are labelled for fine-grained performance analysis. In practice, the gallery scale is a vital problem for recognition accuracy. To this end, we add a distractor set of over 233K sequences, making it more suitable for real-world applications. Since there are a series of gait recognition frameworks using different input data types, the GREW provides silhouettes, Gait Energy Images (GEIs) [14], optical flow, 2D and 3D poses by automatical processing. Compared with controlled gait dataset such as CASIA-B and OU-MVLP, our GREW is fully-unconstrained and has more diverse and practical view variations instead of predefined ones. Meanwhile, there are various challenging factors in the GREW

such as distractor set, complex background, occlusion, carrying, dressing *et al.* as shown in Table 1 and Figure 2.

Equipped with the proposed GREW, the unconstrained gait recognition problem is deeply investigated. Firstly, representative appearance-based and model-based baselines are performed on the GREW, which indicates a lot of room for improvement. For example, top-performed GaitSet [4] obtains 46.28% Rank-1 accuracy on the GREW test set, while it scores more than 80% on the CASIA-B and OU-MVLP. With the distractor set, gait recognition in the wild would become more challenging, while the best model scores only 41.97% Rank-1. Secondly, the influence of the data scale is explored, including the number of training identities and gallery size. Increasing training subjects consistently boosts the performance, while large-scale test set with distractor is still very difficult for CNN-based recognizer. Thirdly, performance on different attributes (gender, age group, carrying, and dressing) is reported, which gives in-depth analysis results. Lastly, we validate the effectiveness of the GREW for pre-training. Fine-tuning models pre-trained on the GREW shows superior performance for cross-dataset gait recognition.

The main contributions can be summarized as follows:

A large-scale benchmark is constructed for the research community towards gait recognition in the wild. The proposed GREW consists of 26K subjects and 128K sequences with rich attributes from flexible data streams, which makes it the first dataset for unconstrained gait recognition to the best of our knowledge. To constitute the GREW benchmark, we collect thousands of hours of streams from multiple cameras in open systems. With automatical pre-processing and tremendous manual identity annotations, there are

Figure 2: Identities examples of the GREW dataset. The first two rows show 2 subjects with various challenges. The last row shows a subject in distractor set. Faces are masked to protect privacy.

more than 14M boxes that simultaneously provide silhouettes and human poses. Besides, we enrich the GREW by a distractor set with 233K sequences, making it more suitable for real-world applications. Enabled by the new benchmark, we perform extensive gait recognition experiments and establish comprehensive baselines, including representative methods, scale influence, attributes analysis and pre-training. Results indicate that the GREW is necessary and effective for gait recognition in the wild. Besides, recognizing unconstrained gait is a very challenging task for current SOTA approaches. Lastly, the proposed dataset can be employed as effective pre-training data for controlled gait recognition to achieve higher performance.

2. The GREW Dataset

2.1. Overview of GREW

Qualitative and quantitative comparisons between the GREW and representative gait recognition datasets are illustrated in Figure 1 and Table 1, respectively. The GREW consists of 26,345 subjects and 128,671 sequences, which come from 882 cameras in open environments. Furthermore, we propose the first distractor set in the gait research community, which contains 233,857 sequences. Silhouettes, GEIs and 2D/3D human poses data types are provided for both appearance-based and model-based algorithms as shown in Figure 3. Since the raw data is captured in natural environments, recognizing identities by gait in the GREW is more challenging compared with popular CASIA-B and OU-MVLP. For example, detecting and segmenting the human body from the complex and dynamic background is a difficult task, considering occlusion, truncation, illumination *et al.* As shown in Figure 2, unconstrained setting also

brings new challenging factors for gait patterns, such as diverse view, dressing, carrying, crowd and distractor.

2.2. Data Collection and Annotation

The raw videos are collected from 882 cameras in large public areas, during one day of July, 2020. About 70% cameras have non-overlapping views, and all cameras cover more than 600 positions. *We are authorized by administrations, and all of involved subjects are told to collect data for research purposes.* 7,533 video clips are used, containing near 3,500 hours 1080 1920 streams.

Before annotation, HTC detector [6] is performed to provide initial human boxes. Then annotators select the boxes from the same subject as a trajectory (sequence). Since there are multiple cameras and a certain person may enter/leave the same camera view, one identity always has multiple sequences. We ensure that each subject in GREW *train, val* and *test* set appears at more than 1 camera, which guarantees view diversity. Other sequences are utilized as distractor set as shown in Section 2.5.

In Table 1, we compare GREW with previous gait datasets regarding #identities, #sequences, #cameras, provided data types, #distractor set, environment, view variations and challenging factors. Finally, a total of 128,671 sequences are manually annotated to obtain 26,345 identities, which contains 14,185,478 human boxes. Current #identities in the GREW is lower than OU-LP Bag/Age [55, 68]. Besides, the distractor set consists of 233,857 sequences and 9,676,016 human boxes. It takes 20 annotators working for 3 months for this tremendous labelling, and we hope the proposed GREW benchmark would facilitate future research of unconstrained gait recognition. *It is worth noting that only silhouettes, optical flow and poses (shown in Figure 3 and 4) will be utilized and released, which do not contain any personal visual information.*

Comparison with Video-based and Long-term Person ReID. Most related computer vision tasks are person ReID in the videos and long-term (cloth changing) ReID. Gait recognition approaches aim to identify a certain subject by silhouettes (GEIs) or poses information, instead of RGB input in ReID. This feature makes gait recognizer more friendly for preserving privacy, which may be more easily accepted by the public. Meanwhile, gait pattern is harder to disguise. Moreover, compared with popular video ReID [60, 80, 64, 26, 46, 25] and long-term ReID [76, 73, 69] datasets, our GREW has more #identities and #cameras as shown in Table 2 and 3.

2.3. Automatical Pre-processing

Representative gait recognition approaches can be roughly divided into appearance-based [44, 65, 4, 9, 29, 20] and model-based [53, 32, 30, 2, 28, 34] categories, which take silhouettes (GEIs) and human poses as input, respec-

Table 2: Comparison with video-based person ReID datasets.

Dataset	#Identities	#Cameras	#Boxes
iLIDS-VID [60]	300	2	44K
MARS [80]	1,261	6	1M
Duke-Video [64]	1,812	8	-
Duke-Tracklet [26]	1,788	8	-
LPW [46]	2,731	4	590K
LS-VID [25]	3,772	15	3M
GREW	26,345	882	14M

Table 3: Comparison with long-term person ReID datasets.

Dataset	#Identities	#Cameras	#Boxes
CVID-reID [76]	90	-	77K
COCAS [73]	5,266	30	62K
PRCC [69]	221	3	33K
GREW	26,345	882	14M

tively. In the GREW benchmark, we provide both two data types by automatical pre-processing. Specifically, silhouettes are produced by segmenting the foreground human body utilizing HTC [6] algorithm. We also try the Mask R-CNN [15], which results in inferior gait recognition accuracy. It is worth noting that human detection and segmentation may be less accurate as shown in Figure 3. Compared with near-perfect results of CASIA-B and OU-MVLP in the static background, the GREW enables assessing the influence of less heuristic pre-processing for gait recognition. This is a topic of great interest for practical applications but rarely considered in previous datasets. For GEIs, we do not adopt the gait cycle due to imperfect detection and segmentation in the wild. For human pose estimation, we provide 2D and 3D keypoints by [47] and [5] as illustrated in Figure 3. Furthermore, optical flow [22, 1] is extracted for potential usage as shown in Figure 4.

Figure 3: Examples of silhouette, GEI, 2D and 3D human pose from the GREW dataset.

Figure 4: Examples of optical flow from the GREW dataset

2.4. Human Attributes

For fine-grained recognition analysis, we annotate each sequence with rich attributes. Soft biometric features including gender and age are labelled for all subjects. Ages are categorized into 5 groups, which adopt 14-year intervals for adults (*i.e.* 16 to 30, 31 to 45, 46 to 60). Children (under 16) and elders (over 60) are treated as separate groups. The statistics of gender and age group are given in Figure 5. For each age group, there is an almost balanced male and female distribution. Since carrying and dressing are influential for gait pattern extraction, the GREW benchmark further provides 5 carrying conditions (*i.e.* none, backpack, shoulder bag, handbag, and lift-stuff) and 6 dressing styles (*i.e.* upper-long-sleeve, upper-short-sleeve, upper-sleeveless, lower-long-trousers, lower-shorts, and lower-skirt). Detailed statistics of these attributes is illustrated in Figure 5. Subjects in more than 70% sequences carry something, while upper-short-sleeve and lower-long-trousers form the majority of cloth styles.

Figure 5: Age group, gender, carrying and dressing attributes in the GREW. In (c), upper body dressing styles contain long-sleeve, short-sleeve, and sleeveless, while lower body includes long-trousers, shorts, and skirt.

2.5. Distractor Set

In real-world applications of gait recognition, the gallery scale is a vital factor. Therefore, we further augment the GREW benchmark with an additional distractor set. This dataset contains 233,857 sequences and 9,676,016 boxes, consisting of extra walking trajectories not belonging to the GREW *train*, *val* and *test*. Specifically, identities that are labelled but only appear at 1 camera would be categorized into distractor set. In Section 4.2, apart from the GREW *test* set, we also report baseline results on the GREW *test* + *distractor* set.

2.6. Evaluation Protocol

The GREW dataset is divided into 3 parts: a *train* set with 20,000 identities and 102,887 sequences, a *val* set with 345 identities and 1,784 sequences, a *test* set with 6,000 identities and 24,000 sequences. Identities in 3 sets are captured in different cameras. Each subject in *test* set has 4 sequences, 2 for probe while 2 for gallery. Besides, there is a distractor set with 233,857 sequences. Detailed statistics of the splits are presented in Table 4.

As shown in Figure 6, in the inference stage, recognizing gait in the wild firstly detects the subject from raw videos.

Then the segmentation or pose estimation module is performed to obtain gait input. Gait recognition is always a 1:N searching process, which aims to retrieve the same person from the gallery given a probe subject. When evaluated on test set, gait probe and gallery are all paired. When evaluated on a certain attribute, a subset of probe (sequences with the corresponding attribute) is chosen to perform gait recognition. We adopt prevailing Rank- k as the evaluation metric, which denotes the possibility to locate at least one true positive in the top- k ranks.

Figure 6: The pipeline of gait recognition in the wild, consisting of pre-processing and recognition steps. Pre-processing part detects human from raw sequences, and provides silhouettes (GEIs) or poses information. Given a certain probe, the recognition part performs 1:N searching from the gallery.

Table 4: Statistics of different splits.

Split	#Identities	Sequences	Frames
Train	20,000	102,887	10,166,842
Val	345	1,784	238,532
Test	6,000	24,000	3,780,104
Distractor	-	233,857	9,676,016

3. Baselines on GREW

To establish baselines, representative appearance-based methods [44, 65, 4, 9] and model-based methods [32, 53] are explored. Overview of input type, network and loss is shown in Table 5, and details are described as follows. All models are re-implemented in one codebase using PyTorch [40] and trained on cluster (each with 8 2080TI GPUs, Intel E5-2630-v4@2.20GHz CPU, 256G RAM). For GREW training, we train both models for 250K iterations with batch size of ($p = 32, k = 8$) and Adam. The learning rate starts at 10^{-4} and decreases to 10^{-5} after 150K iterations. For CASIA-B fine-tuning, the models are trained for extra 50K iterations with a constant learning rate of 10^{-5} . None of layer weight is frozen.

3.1. Appearance-based

GEINet [44] directly learns gait representation features from GEIs and then corresponds to identities. As shown in Table 5, the network of the GEINet has 4 layers, consisting of 2 convolution and 2 Fully-Connected (FC) layers. Softmax loss is adopted for optimization, and output from the last FC is utilized to calculate a distance between probe and gallery.

Table 5: Overview of adopted baselines, including input data type, number of network layers, dimensions of embedding feature, and loss. N in #embedding of the GEINet means #training identities.

Baseline	Input	#Layers	#Embed.	Loss
GEINet	GEI	4	N	Softmax
TS-CNN	GEI	6	-	2-cls Cross-entropy
GaitSet	Sil.	10	15,872	Batch All triplet
GaitPart	Sil.	10	4,096	Batch All triplet
PoseGait	3D Pose	22	512	Softmax&Center
GaitGraph	2D Pose	44	256	Contrastive

TS-CNN [65] framework adopts two-stream CNN architecture which learns similarities between GEIs pair for gait recognition. MT architecture setting is utilized in this paper, which matches mid-level features at the top layer. TS-CNN also takes GEIs as input and has 6 layers. 2-class Cross-entropy loss is used for training, while classifier indicates probability of two subjects whether they are the same one during inference.

GaitSet [4] uses several convolution and pooling layers to extract convolutional templates on unordered silhouettes set. Batch All triplet loss [17] is adopted for optimizing, and 15,872- d embedding features are utilized for recognition during inference. Following the OU-MVLP training setting, we use more channels convolutional layers and 250K iterations with 2 learning rate schedule.

GaitPart [9] proposes a part-based network design focusing on fine-grained representation and micro-motion in different parts of the human body. Training and testing on the GREW benchmark follow most GaitSet settings.

3.2. Model-based

PoseGait [32] explores 3D human pose as gait recognition input which is estimated by [5]. And 2D pose extracted from [47] is utilized to obtain 3D pose information. For the gait feature part, a 22-layers (20 convolution and 2 FC) CNN with 512- d embedding is trained for extraction, which is optimized by Softmax and Center losses.

GaitGraph [53] is a recent model-based gait recognition approach with a promising result on CASIA-B. This work combines 2D human pose input and graph convolutional network to achieve gait recognition. Supervised Contrastive loss is utilized to optimize the graph network, and we strictly follow its augmentation and training details. During evaluation, the 256- d feature vector is extracted for calculating distance between probe and gallery.

4. Experiments

In experiments, we perform extensive baselines and analyses on the proposed GREW dataset. Firstly, main baseline results of 6 approaches are reported. Then we investigate the influence of the scale including increasing training and testing identities, distractor set size. Thirdly, performance on different human attributes is compared,

Table 6: Rank-1, Rank-5, Rank-10, Rank-20 (%) of baselines. Trained on the GREW `train` set and evaluated on `test` set.

Baseline	Rank-1	Rank-5	Rank-10	Rank-20
GEINet	6.82	13.42	16.97	21.01
TS-CNN	13.55	24.55	30.15	37.01
GaitSet	46.28	63.58	70.26	76.82
GaitPart	44.01	60.68	67.25	73.47
PoseGait	0.23	1.05	2.23	4.28
GaitGraph	1.31	3.46	5.08	7.51

consisting of accuracy on gender, age group, carrying condition and dressing style. Fourthly, we showcase the effectiveness of our dataset for pre-training, and time analyses for practical applications. Last comes sample results on successes and failures of gait recognition.

4.1. Main Baseline Results

The Rank- k accuracy of 6 baselines are illustrated in Figure 7 and summarized in Table 6. The GREW `train` and `test` set are utilized for training and evaluation, respectively. Results indicate that GaitSet [4] and GaitPart [9] are superior approaches for gait recognition in the wild, consistent with the performance on constrained CASIA-B [74] and OU-MVLP [51]. More specifically, GaitSet and GaitPart score 46.28% and 44.01% in terms of Rank-1 metric, respectively. Both of them exceed 60% and 70% for Rank-5 and Rank-20 criteria. Since TS-CNN [65] and GEINet [44] take GEIs as input and have relatively fewer layers, they achieve much lower accuracy on the GREW benchmark. GEIs lose some useful temporal information, which may be important for unconstrained gait recognition. Comparing TS-CNN with GEINet, the former adopts two-stream metric learning, thus suffers less from the over-fitting problem and obtains higher accuracy. Model-based PoseGait [32] and GaitGraph [53] baselines result in inferior performance compared with appearance-based ones, indicating that gait recognition in the wild is very challenging for human pose input.

Considering that the GREW is the first unconstrained gait benchmark, we compare the result with that on CASIA-B and OU-MVLP. For top-performed GaitSet and GaitPart, Rank-1 scores on CASIA-B and OU-MVLP exceed 80%. Due to more challenging factors on the GREW dataset such as diverse view, carrying and dressing variations, they only successfully recognize 46.28% and 44.01% sequences in terms of Rank-1 criteria. When the distractor set is added to the gallery, the best accuracy decreases to 41.97%, showing the difficulty of real-world gait recognition. Results indicate that GREW is essential and effective for unconstrained gait recognition, and there is a lot of room for improvement.

4.2. Influence of the Scale

In the deep learning era, large-scale labelled data plays an significant role for bench-marking various vision tasks

Figure 7: Rank- k result (%) of baselines. Trained on the GREW `train` and evaluated on `test` set. Legend shows Rank-1 / Rank-20 accuracy.

[41, 33, 13, 81]. In this section, we investigate the data scale influence for training and testing on the GREW.

Accuracy with Increasing Training Identities In this experiment, we demonstrate gait recognition accuracy with increasing training identities. 6 different subset sizes are prepared, including 1K, 2K, 4K, 8K, 16K and the maximum 20K. The first 5 training subsets are randomly chosen but fixed for different algorithms. The evaluation is performed on whole GREW `test` set.

As presented in Figure 8, for state-of-the-art GaitSet and GaitPart, the Rank-1 on `test` set grows stably with more training identities. Therefore, the 20K size of the whole training set achieves the highest Rank-1 accuracy. Specifically, GaitSet increases the Rank-1 from 28.0% on 1K training subjects to 46.28% on 20K subjects. The results clearly show that large-scale GREW training data is helpful for future gait recognition research.

For GEINet baseline, the scale of training data does not obviously influence the performance. The reason may be that the network architecture in GEINet has limited capability to learn from large data. TS-CNN uses a two-stream metric learning network structure and takes pairs of GEIs as inputs, which may be less suffered from over-fitting. Therefore, its Rank-1 accuracy slightly increases from 9.50% to 13.55%. Model-based baselines are not sensitive to training data scales due to inferior accuracy.

Figure 8: Rank-1 accuracy (%) on `test` set with increasing training identities. Legend shows performance changes from 1K to 20K data.

Accuracy with Increasing Test Identities A sufficient test set is essential for evaluating the performance of the gait recognizer. In this experiment, we study the relationship between the search space scale and the Rank-1 accuracy as shown in Figure 9. When the test identities increase from 1K to 6K, almost all approaches suffer from accuracy degradation. More specifically, GaitSet scores 57.45% Rank-1 on 1K test identities but decreases to 49.83% when test size is doubled. When the subjects increase to 6K, precision degradations of both GaitSet and GaitPart are more than 10%. With increasing identities in the gallery, the possibility of inter-subject appearance similarity becomes higher, so recognizing certain identity by top retrieval is more challenging. Evaluation results on other baselines come to the same conclusion.

Figure 9: Rank-1 accuracy (%) with different identities in test set. Legend shows performance changes from 1K to 6K test data.

Accuracy with Distractor Set In gait applications, the gallery size may be very large considering numerous unrelated identities. We add the constructed distractor set into the gallery to investigate this practical setting. As shown in Figure 10, by enlarging the gallery with distractor set, most approaches obtain lower recognition scores. When all 233K distractor sequences are involved, Rank-1 of the best baseline GaitSet decreases to 41.97%. Accuracy with distractor set shows the necessity of the GREW benchmark again.

4.3. Performance on Different Attributes

This section investigates the performance variations of gait recognition between different attributes, including gender, age group, carrying and dressing. We adopt GaitSet [4] as the recognition approach since it performs best in the baseline experiments.

The Rank-1 accuracy for gender and age group is illustrated in Figure 11. According to the results, for most age groups, gait recognition performance on females is always better than that on males. We argue that females contain more different variations such as wearing and hairstyle, which may be helpful for individual recognition by gait silhouettes. For results on different age groups, one can find

Figure 10: Rank-1 accuracy (%) with increasing gallery size. Different distractor scales are added. Legend shows performance changes from test to test + distractor.

that performance on the children is worse than other groups because of walking mode immaturity. Besides, recognition accuracy on elders is slightly lower than on adults due to physical degeneration. The attribute results on carrying and dressing are shown in Table 7. Compared with normal walking (*i.e. None*), various carryings always decrease gait recognition accuracy. More specifically, *Lift-stuff* is most difficult since it contains more diversity. For dressing styles, results indicate that *Skirt* is more challenging to recognize by silhouettes in GaitSet.

Figure 11: Rank-1 accuracy (%) on gender and age group attributes.

Table 7: Rank-1 accuracy (%) on carrying and dressing attributes. Subsets of probe (sequences with the corresponding attribute) are chosen to perform gait recognition. For evaluation with dressing, *All* means gait probe and gallery are paired without attention to any clothing style. *Short/Long* refers to short/long-wearing in both upper and lower body.

Carrying	Rank-1	Dressing	Rank-1
None	52.36	All	46.28
Backpack	48.83	Short	48.16
Shoulder bag	46.68	Long	44.92
Handbag	47.02	Skirt	44.30
Lift-stuff	45.66	-	-

4.4. GREW for Pre-training

To validate the effectiveness of pre-trained models using the GREW dataset, we conduct a cross-dataset experiment in this section. Original (both training and testing on CASIA-B), direct cross-dataset evaluation (training on GREW, testing on CASIA-B), and fine-tuning (pre-trained

by the GREW, finetuning and evaluating on CASIA-B) performance of the GaitSet are compared. Specifically, GaitSet obtains 83.64%, 45.14%, 84.48% on CASIA-B via three settings. The accuracy of the second configuration is inferior because of the obvious domain gap. With fine-tuning on the target domain, gait recognition accuracy significantly outperforms the original ones by 0.84%, which shows the superior capacity of our dataset for pre-training.

4.5. Times

Apart from accuracy, speed is also a crucial factor for practical gait recognition, which is always neglected in previous literature. In this section, we compare inference time of different baselines, including pre-processing, gait feature extraction, and searching in the gallery. Times are roughly measured on the GREW test set by averaging all sequences duration. As shown in Table 8, for a sequence with 157 frames on average, pre-processing (*i.e.* detection, segmentation, pose estimation *et al.*) takes most of the time. Gait feature extraction (main network inference) and searching procedure are relatively faster. FLOPs and parameters of gait networks are also calculated for comparison. In summary, current gait recognition pipelines need to be optimized for real-world applications.

Table 8: Inference time, FLOPs and parameters of baselines (with single 2080TI GPU). Since TS-CNN needs multiple forward steps for a certain sequence, it is not compared.

Baseline	Pre-process	Feature	Search	Total	FLOPs	Params
GEINet	45.62s	0.03s	0.00066s	45.65s	0.02G	7.68M
GaitSet	45.62s	2.89s	0.00058s	48.51s	1.06G	6.31M
GaitPart	45.62s	3.09s	0.00234s	48.71s	0.92G	6.01M
PoseGait	54.69s	0.18s	0.00046s	54.87s	0.08G	7.74M
GaitGraph	53.59s	0.05s	0.00041s	53.64s	0.06G	527.95K

4.6. Sample results

Figure 12 provides several sample results on the GREW test set, which are performed by GaitSet baseline. For the first probe, GaitSet successfully retrieves the subject in Rank-1 result, with changed clothes and different walking directions. For the second probe, the results of Rank-1 are incorrect due to similar skirt dressing, while following two retrievals, with carrying and partial occlusion, are true positive.

5. Discussion and Conclusion

Discussion During construction of the GREW benchmark, *privacy and bias problems are our first concern*. To protect privacy, *only silhouettes, flow and human poses would be utilized and released*, which do not reveal any personal visual information. We will *provide strict access for applicants who sign the license, and try our best to guarantee it for research purposes only*. For dataset bias, the

Figure 12: Sample results on the GREW with GaitSet. Left part with blue boxes shows probes (3 frames belong to the same sequence), while results with green and red boxes are true positive and false positive respectively. Note that only silhouettes are used for gait recognition, and RGB images are just for visualization.

GREW has balanced gender distribution, while some attributes (*e.g.* race, age group, dressing) are inevitably biased due to capture location and time. Since our dataset is large-scale and diverse, one can sample balanced data to train models with less bias. Besides, recent de-bias researches in the biometrics community [59, 11, 58] may also alleviate this problem.

Conclusion This paper makes the first step to large-scale gait recognition in the wild, to the best of our knowledge. Firstly, the GREW dataset contains 128K sequences of 26K subjects with rich attribute variations from flexible data. Secondly, we manually annotate thousands of hours streams from hundreds of cameras, resulting in 14M boxes with automatic silhouettes and human poses. Moreover, 233K distractor set sequences are collected for practical evaluation. Lastly, comprehensive baselines are conducted to quantitatively analysis the challenges in unconstrained gait recognition, deriving in-depth and constructive insights. Future work will further investigate open problems for gait recognition, *e.g.* influence of pre-processing, deeper and modern network, disentanglement, soft-biometric recognition, un/semi/self-supervised learning.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 61822603, Grant U1813218, and Grant U1713214, in part by a grant from the Beijing Academy of Artificial Intelligence (BAAI), and in part by a grant from the Institute for Guo Qiang, Tsinghua University.

References

- [1] <https://github.com/NVlabs/Flownet2-pytorch/>. 4
- [2] 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. *TBIOM*, 2020. 2, 3
- [3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *FG*, 2018. 2
- [4] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. GaitSet: Regarding gait as a set for cross-view gait recognition. In *AAAI*, 2019. 1, 2, 3, 5, 6, 7
- [5] Ching-Hang Chen and Deva Ramanan. 3D human pose estimation = 2D pose estimation + matching. In *CVPR*, 2017. 4, 5
- [6] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiao Xiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 3, 4
- [7] Naresh Cuntoor, Amit Kale, and Rama Chellappa. Combining multiple evidences for gait recognition. In *ICASSP*, 2003. 1
- [8] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2
- [9] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. GaitPart: Temporal part-based model for gait recognition. In *CVPR*, 2020. 1, 3, 5, 6
- [10] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. Horizontal pyramid matching for person re-identification. In *AAAI*, 2019. 2
- [11] Sixue Gong, Xiaoming Liu, and Anil K Jain. Jointly debiasing face recognition and demographic attribute estimation. In *ECCV*, 2020. 8
- [12] Ralph Gross and Jianbo Shi. The CMU motion of body (MoBo) database. 2001. 2
- [13] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 2, 6
- [14] Ju Han and Bir Bhanu. Individual recognition using gait energy image. *TPAMI*, 2006. 2
- [15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 4
- [16] Yiwei He, Junping Zhang, Hongming Shan, and Liang Wang. Multi-task GANs for view-specific feature learning in gait recognition. *TIFS*, 2019. 1
- [17] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv:1703.07737*, 2017. 1, 2, 5
- [18] Martin Hofmann, Jürgen Geiger, Sebastian Bachmann, Björn Schuller, and Gerhard Rigoll. The TUM gait from audio, image and depth (GAID) database: Multimodal recognition of subjects and traits. *JVCIR*, 2014. 2
- [19] Md Altab Hossain, Yasushi Makihara, Junqiu Wang, and Yasushi Yagi. Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. *PR*, 2010. 2
- [20] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *ECCV*, 2020. 1, 3
- [21] Yuge Huang, Yuhang Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. CurricularFace: adaptive curriculum learning loss for deep face recognition. In *CVPR*, 2020. 2
- [22] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *CVPR*, 2017. 4
- [23] Haruyuki Iwama, Mayu Okumura, Yasushi Makihara, and Yasushi Yagi. The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition. *TIFS*, 2012. 1, 2
- [24] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The MegaFace benchmark: 1 million faces for recognition at scale. In *CVPR*, 2016. 2
- [25] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *ICCV*, 2019. 3, 4
- [26] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, 2018. 3, 4
- [27] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deep-ReID: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 2
- [28] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Joint intensity transformer network for gait recognition robust against clothing and carrying status. *TIFS*, 2019. 3
- [29] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. In *CVPR*, 2020. 3
- [30] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, Shiqi Yu, and Mingwu Ren. End-to-end model-based gait recognition. In *ACCV*, 2020. 3
- [31] Rijun Liao, Chunshui Cao, Edel B Garcia, Shiqi Yu, and Yongzhen Huang. Pose-based temporal-spatial network (PT-SN) for gait recognition with carrying and clothing variations. In *CCBR*, 2017. 1
- [32] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *PR*, 2020. 3, 5, 6
- [33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 6
- [34] Dan Liu, Mao Ye, Xudong Li, Feng Zhang, and Lan Lin. Memory-based gait recognition. In *BMVC*, 2016. 3
- [35] Jiwen Lu, Gang Wang, and Pierre Moulin. Human identity and gender recognition from gait sequences with arbitrary walking directions. *TIFS*, 2014. 2

- [36] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR Workshops*, 2019. 2
- [37] Yasushi Makihara, Hidetoshi Mannami, Akira Tsuji, Md Altab Hossain, Kazushige Sugiura, Atsushi Mori, and Yasushi Yagi. The OU-ISIR gait database comprising the treadmill dataset. *CVA*, 2012. 1
- [38] Yasushi Makihara, Hidetoshi Mannami, and Yasushi Yagi. Gait analysis of gender and age using a large-scale multi-view gait database. In *ACCV*, 2010. 2
- [39] Athira Nambiar, Alexandre Bernardino, and Jacinto C Nascimento. Gait-based person re-identification: A survey. *ACM Computing Surveys*, 2019. 1
- [40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5
- [41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *IJCV*, 2015. 6
- [42] Sudeep Sarkar, P Jonathon Phillips, Zongyi Liu, Isidro Robledo Vega, Patrick Grother, and Kevin W Bowyer. The humanID gait challenge problem: Data sets, performance, and analysis. *TPAMI*, 2005. 1, 2
- [43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2
- [44] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. GEINet: View-invariant gait recognition using a convolutional neural network. In *ICB*, 2016. 1, 3, 5, 6
- [45] Jamie D Shutler, Michael G Grant, Mark S Nixon, and John N Carter. On a large sequence-based human gait database. In *Applications and Science in Soft Computing*. 2004. 2
- [46] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *AAAI*, 2018. 3, 4
- [47] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 4, 5
- [48] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 2
- [49] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Liior Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 2
- [50] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. On Input/Output architectures for convolutional neural network-based cross-view gait recognition. *TCSVT*, 2017. 1
- [51] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *CVA*, 2018. 1, 2, 6
- [52] Daoliang Tan, Kaiqi Huang, Shiqi Yu, and Tieniu Tan. Efficient night gait recognition based on template matching. In *ICPR*, 2006. 2
- [53] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. GaitGraph: Graph convolutional network for skeleton-based gait recognition. *arXiv preprint arXiv:2101.11228*, 2021. 3, 5, 6
- [54] Akira Tsuji, Yasushi Makihara, and Yasushi Yagi. Silhouette transformation based on walking speed for gait identification. In *CVPR*, 2010. 2
- [55] Md Zasim Uddin, Thanh Trung Ngo, Yasushi Makihara, Noriko Takemura, Xiang Li, Daigo Muramatsu, and Yasushi Yagi. The OU-ISIR large population gait database with real-life carried object and its performance evaluation. *CVA*, 2018. 2, 3
- [56] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *CVPR*, 2018. 2
- [57] Liang Wang, Tieniu Tan, Huazhong Ning, and Weiming Hu. Silhouette analysis-based gait recognition for human identification. *TPAMI*, 2003. 2
- [58] Mei Wang and Weihong Deng. Mitigate bias in face recognition using skewness-aware reinforcement learning. *arXiv preprint arXiv:1911.10692*, 2019. 8
- [59] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *CVPR*, 2019. 8
- [60] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, 2014. 3, 4
- [61] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *CVPR*, 2018. 2
- [62] Thomas Wolf, Mohammadreza Babaei, and Gerhard Rigoll. Multi-view gait recognition using 3D convolutional neural networks. In *ICIP*, 2016. 1
- [63] Xinhui Wu, Weizhi An, Shiqi Yu, Weiyu Guo, and Edel B García. Spatial-temporal graph attention network for video-based gait recognition. In *ACPR*, 2019. 1
- [64] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *CVPR*, 2018. 3, 4
- [65] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep CNNs. *TPAMI*, 2017. 1, 3, 5, 6
- [66] Qiqi Xiao, Hao Luo, and Chi Zhang. Margin sample mining loss: A deep learning based method for person re-identification. *arXiv:1710.00478*, 2017. 2
- [67] Chi Xu, Yasushi Makihara, Xiang Li, Yasushi Yagi, and Jianfeng Lu. Gait recognition from a single image using a phase-aware gait cycle reconstruction network. In *ECCV*, 2020. 1
- [68] Chi Xu, Yasushi Makihara, Gakuto Ogi, Xiang Li, Yasushi Yagi, and Jianfeng Lu. The OU-ISIR gait database compris-

- ing the large population dataset with age and performance evaluation of age estimation. *CVA*, 2017. 2, 3
- [69] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *TPAMI*, 2019. 3, 4
 - [70] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014. 2
 - [71] Shiqi Yu, Haifeng Chen, Edel B Garcia Reyes, and Norman Poh. GaitGAN: Invariant gait feature extraction using generative adversarial networks. In *CVPR Workshops*, 2017. 1
 - [72] Shiqi Yu, Haifeng Chen, Qing Wang, Linlin Shen, and Yongzhen Huang. Invariant feature extraction for gait recognition using only one uniform model. *Neurocomputing*, 2017. 1
 - [73] Shijie Yu, Shihua Li, Dapeng Chen, Rui Zhao, Junjie Yan, and Yu Qiao. Cocas: A large-scale clothes changing person dataset for re-identification. In *CVPR*, 2020. 3, 4
 - [74] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, 2006. 1, 2, 6
 - [75] Kaihao Zhang, Wenhan Luo, Lin Ma, Wei Liu, and Hongdong Li. Learning joint gait representation via quintuplet loss minimization. In *CVPR*, 2019. 1
 - [76] Peng Zhang, Jingsong Xu, Qiang Wu, Yan Huang, and Xianye Ben. Learning spatial-temporal representations over walking tracklet for long-term person re-identification in the wild. *TMM*, 2020. 3, 4
 - [77] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. AlignedReID: Surpassing human-level performance in person re-identification. *arXiv:1711.08184*, 2017. 2
 - [78] Yuqi Zhang, Yongzhen Huang, Shiqi Yu, and Liang Wang. Cross-view gait recognition by discriminative feature learning. *TIP*, 2019. 1
 - [79] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *CVPR*, 2019. 1
 - [80] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. MARS: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 2, 3, 4
 - [81] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 2, 6
 - [82] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv:1610.02984*, 2016. 2
 - [83] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 2
 - [84] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. WebFace260M: A benchmark unveiling the power of million-scale deep face recognition. In *CVPR*, 2021. 2