

GaitFace: A Multimodal Dataset for Long-Range Person Identification

Alain Komaty Luis S. Luevano Vidit Vidit Zeina Al Amine Anjith George Sébastien Marcel
Idiap Research Institute, Martigny, Switzerland

{akomaty, luis.luevano, vidit.vidit, anjith.george, zeina.alamine, sebastien.marcel}@idiap.ch

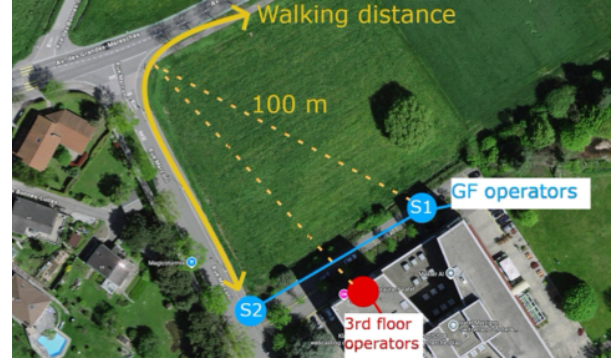
Abstract

Efficient border control is becoming a significant global challenge, mainly due to severe congestion and extended passenger waiting times. To mitigate these bottlenecks and facilitate passenger flow, biometric technologies are increasingly deployed to streamline identity verification and enhance crossing efficiency. Technical limitations frequently impede biometric identification, particularly in long-range surveillance, where systems must deal with adverse atmospheric conditions and degraded image quality. While high-quality frameworks like BRIAR exist, they are frequently restricted to specific government agencies. This paper introduces GaitFace, a new public dataset that contains face and gait data captured at long distances. To ensure that the research reflects authentic border scenarios, we use Pre-Enrollment data, where a traveler registers via a mobile device, and "In-the-Wild" captures, which records individuals at a distance across multiple viewing angles and different cameras. Benchmarking SOTA face and gait models reveals that current architectures fail under low-resolution and elevated viewpoints despite success with optical assistance. GaitFace exposes these critical vulnerabilities, providing a rigorous public benchmark to drive more robust, unconstrained biometric research. Available at <https://idiap.ch/paper/gaitface>

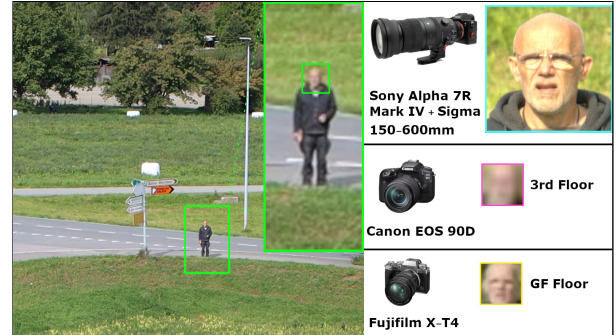
1. Introduction

Biometric recognition at long distances is an essential capability for automated border crossing [23, 21], surveillance [29], and forensic [34] applications. Unlike traditional close-range biometrics in controlled scenarios, long-distance person recognition presents scenarios with severe image degradation caused by low resolution, motion blur, changing viewpoints, atmospheric effects, and varying illumination and weather conditions.

Face and gait data are particularly relevant in this setting because both can be used to provide automatic biometric recognition. These two modalities also provide complementary information: face recognition is highly discriminative when sufficient facial detail is available, while gait can



(a) Map showing positions of recording teams and walking paths.



(b) Visual comparison of facial resolution across different cameras/heights.



(c) "Walking Towards Camera" (WTC) scenario.

Figure 1. Experimental overview: (a) station layout and path; (b) hardware comparison; and (c) sample frames showing long- to near-range capture.

remain informative at long distances.

Despite recent progress in face and gait recognition, existing datasets only partially cover the conditions required for realistic long-distance person recognition. Large-scale face datasets collected in the wild have enabled the development of highly-accurate face recognition models[30], but they are driven by high-quality face imagery and do not show the same performance with native low-quality surveillance probes[27]. Low quality and surveillance face datasets provide more challenging probes, but often lack controlled high-quality enrollment data [3, 4], long-distance optical-zoom captures[12], or temporal variability across repeated sessions[20]. Similarly, most gait datasets [40, 41, 24, 14] include controlled or short-to-medium range sequences, while long-distance gait probes with realistic viewpoint, clothing, and accessory changes remain scarce. As a result, current benchmarks encounter limitations in designing person recognition protocols. Another important constraint is dataset availability. Some of the most relevant resources for long-range and surveillance biometrics, such as IJB-S[20] and BRIAR[5], are highly restricted, limiting reproducibility and excluding many academic, international, and commercial users. At the same time, biometric dataset collection must address personal privacy, informed consent, and comply with data governance regulations [6, 35].

To address these gaps, we introduce **GaitFace**, a multimodal dataset for long-range person identification designed around a border-crossing-inspired acquisition protocol. GaitFace combines controlled pre-enrollment face captures from mobile devices with outdoor long-distance observations, including high-resolution optical-zoom face images, native low-resolution face video, and long-distance gait recordings from multiple viewpoints. The dataset contains 70 participants recorded across two sessions on different days, introducing temporal variability in clothing, accessories, hairstyle, subject appearance, and environmental conditions. This design enables the study of high-quality reference to long-distance probe matching, low-resolution face recognition, long-distance gait recognition, and multimodal face-gait analysis. The main contributions in this paper are:

- We introduce **GaitFace**, a multimodal long-distance person recognition dataset combining controlled mobile pre-enrollment, high-resolution optical-zoom face captures, native low-resolution face video, and long-distance gait recordings.
- We design a realistic multi-session acquisition protocol with multiple viewpoints, subject motion, clothing changes, accessories, phone usage, and natural environmental variability.

- We define evaluation protocols for long-distance face recognition, long-distance gait recognition, and cross-session matching under realistic covariate changes.
- We provide benchmark results using state-of-the-art face and gait recognition methods, highlighting the difficulty of the proposed setting and establishing baselines for future research.

The paper is organized as follows: Section 2 reviews related work in long-distance face and gait recognition. Section 3 presents our proposed GaitFace dataset. Section 4 describes the experimental setup and evaluation protocols. Section 5 reports benchmark results for face and gait recognition. Finally, we give our concluding remarks in Section 6.

2. Related Work

In this section, we review prior work across long-distance face and gait recognition, highlighting the critical gaps in publicly available multimodal benchmarks to motivate our proposed dataset.

Unimodal Recognition Limitations

Modern face recognition relies heavily on large-scale, high-resolution datasets (e.g., LFW [18], CASIA-WebFace [38], MS-Celeb-1M [15], WebFace260M [42], Glint360K [1]) and benchmarks like IJB-C [31], but models trained on these suffer severe performance drops when applied to low-resolution surveillance scenarios (e.g., SCFace [12], UCCS [13], DroneFace [17], TinyFace [3], SurvFace [4]). Similarly, unimodal gait recognition methods are typically evaluated on datasets restricted to short-to-medium ranges (e.g., CASIA-B [40], Gait3D [41]).

Multimodal Face and Gait Datasets

Because facial features degrade at a distance while gait remains informative, multimodal evaluation is critical for realistic border control and surveillance. However, publicly available datasets rarely provide synchronized face and gait captures at extreme distances. While datasets like IJB-S [20] and BRIAR [5, 26, 36] capture long-range surveillance scenarios, their access is highly restricted, limiting broad research and commercial reuse. To address this gap, GaitFace introduces a public benchmark combining controlled mobile pre-enrollment with long-distance multimodal outdoor captures. Table 1 compares GaitFace against the limited existing multimodal datasets from the literature.

3. GaitFace: Long-distance face & gait dataset

Overview The dataset was ethically and legally acquired from 70 consenting participants (ID 0–69) under a protocol approved by the relevant institutional review boards, with all individuals providing explicit informed consent for the collection and processing of their biometric data. Recorded across two acquisition sessions separated by several weeks,

Dataset Name	Dist	Env	#Subjs	Public
BRIAR [5, 26, 36]	<1,000m	Both	~1,760	No
NIST Database [19]	10-15m	Outdoor	30	Yes
OU-ISIR Database [32]	~8m	Indoor	1,912	Yes
FOCS [28]	<8.3m	Outdoor	437	Yes
NLPR Gait [10]	5-12m	Outdoor	20	Yes
CASIA-B-Gait-Face [9]	5-15m	Indoor	124	Yes
Walking Video [2]	<25m	Outdoor	25	No
GaitFace (Ours)	<100m	Both	70	Yes

Table 1. Comparison of Multimodal Face and Gait Datasets

the dataset includes associated demographic metadata and amounts to approximately 2.7 TB of multi-modal data, comprising RGB high-resolution facial images in RAW format and multi-view gait recordings. The collection protocol incorporates repeated measurements to capture natural temporal variations in appearance and utilizes a long-distance acquisition design—capturing biometrics at distances of up to 100 meters—to reflect realistic border surveillance conditions. By recording face and gait data simultaneously from multiple viewpoints, the dataset enables consistent multi-modal analysis for each participant, and it will be publicly released upon acceptance.

3.1. Capture conditions and devices

Data were collected outdoors, ensuring realistic variability in illumination (natural daylight variations) and weather conditions (sunny, cloudy, windy, and rainy). To introduce controlled variations in appearance and behavior, participants were recorded under four standardized scenarios: normal walking, walking while simulating a phone call, walking with an accessory (hand bag or backpack), and walking with a different jacket. All recordings followed a predefined walking trajectory (Figure 1a) shared across scenarios and sessions: participants started from a fixed point, walked along the path to a designated stop location where they briefly paused and faced the third floor camera (Figure 1b), continued to an end point, turned, and returned to the starting point. This round-trip path was repeated for each scenario. The acquisition geometry includes a ground-level viewpoint (eye-level perspective) and an elevated viewpoint (approximately 9–10 meters), simulating typical surveillance camera positions.

3.2. Data Capture Modalities

We describe the multiple modalities considered in GaitFace which are: pre-enrollment face captures, outdoor face captures, outdoor gait captures, and relevant subject meta-data.

Pre-enrollment Face Captures. Face enrollment (Figure 2a) was performed indoors using two mobile devices (iPhone 12 and Samsung Galaxy S9), where each participant recorded facial sequences for 15 seconds, including

static (first 5 seconds of the recording) and controlled horizontal and vertical head movements (last 10 seconds of the recording) using the phones’ frontal camera. These data serve as high-quality ground-truth references for subsequent analysis. This setup mimics a real-world remote pre-enrollment process.



Figure 2. Data samples: (a) Enroll, (b) Probe HQ, (c) 100m 3rd Floor (3F), (d) 100m Ground Floor (GF), (e) 3m GF.

Outdoor Face Captures Facial data were collected in an outdoor environment to simulate realistic long-distance surveillance conditions under natural illumination. Participants followed the predefined path and were instructed to stop at a designated point and face the sensor for a short duration (5–10 seconds), enabling the acquisition of high-quality facial images using a long-range zoom camera (Figure 1b) positioned at the third-floor level (Figure 2b). The captured images were stored in RAW format to preserve fine-grained facial details. Simultaneously, the faces were recorded from two cameras (Canon EOS 90D and Fujifilm X-T4 shown in Figure 1b) from the 3rd floor (Figure 2 (c)) and Ground floor (Figure 2 (d),(e)). The acquisition was conducted under varying weather conditions, including sunny, cloudy, and rainy environments, introducing natural variability in lighting and visibility. This setup provides high-quality reference imagery, enabling robust evaluation of face recognition systems under challenging conditions.



Figure 3. Face captures acquired using the GF camera, showing the subject’s face evolution while approaching the camera at different timestamps (Top left: first frame, bottom right: last frame).

Outdoor Gait Captures Gait data were collected in an outdoor environment using a multi-view acquisition setup designed to capture full-body motion from complementary perspectives. Participants followed the predefined walking trajectory, ensuring consistent spatial alignment across recordings. Data were acquired simultaneously from

ground-level cameras and an elevated viewpoint (third-floor level), providing both frontal/lateral and top-down observations of motion patterns. The recordings from the third-floor level were consistently captured across both sessions using the same cameras setup. At ground level, gait videos were recorded using a Fujifilm X-T4 camera, with the camera position intentionally varied between the two sessions to introduce additional viewpoint variation. Examples of the scenario of session 2 *Walking Towards the Camera* (WTC) are given in Figure 1c for gait and Figure 3 for face. The recordings were performed under natural illumination and varying environmental conditions, introducing realistic variability in appearance and visibility. This multi-view configuration allows us to capture different gait dynamics under unconstrained outdoor settings.

Quantitative Resolution Analysis. To quantitatively characterize the dataset’s image degradation, we measured the average face crop resolution and Inter-Pupillary Distance (IPD) across configurations. At the maximum standoff distance of 100m without optical assistance (LQ probes from GF and 3F cameras), the average face resolution is 18×22 pixels with an IPD of 4 ± 1.2 pixels, representing severe spatial degradation. Conversely, the 100m optically-zoomed HQ probes provide an average face resolution of 210×250 pixels with an IPD of 55 ± 3.5 pixels. For the WTC protocol, the face resolution scales dynamically from the baseline 18×22 pixels at 100m up to 450×520 pixels (IPD of 115 ± 6.2 pixels) at the 3m terminal range.

Protocol	Enroll (# Frames, Criteria)	Probe (# Frames, Criteria)
HQ-[1, 3]	1, 3 (Frontal)	1 (HQ)
HQ-[10, 20]	10, 20 (Mix: Frontal/Tilted)	1 (HQ)
GF-3	3 (Frontal)	[1, 10, 20, 40] (LQ)
GF-10	10 (Mix: Frontal/Tilted)	[1, 10, 20, 40] (LQ)
3F	3 (Frontal)	[1, 20, 40] (LQ)
WTC	3 (Frontal)	1 (HQ/LQ)

Table 2. FR benchmarking protocols. HQ = High Quality (100m zoom), LQ = Low Quality (no zoom), GF = Ground Floor, 3F = 3rd Floor, WTC = Walking Towards Camera. All enrollment data were captured using the frontal cameras of mobile devices, imitating a real pre-enrollment scenario.

3.3. Metadata and demographic distribution

Along with visual data, this dataset includes participant metadata (age, gender, and skin color) to enable future benchmarking for algorithmic fairness and demographic robustness. The demographic distribution of the dataset is balanced in terms of age and gender, with 45.7% of female and 54.3% of male volunteers. However, due to local population characteristics at the acquisition location, the dataset predominantly represents individuals with lighter skin tones, as illustrated in Figure 4, which present the distributions of age, gender, skin color, and Fitzpatrick scale [16].

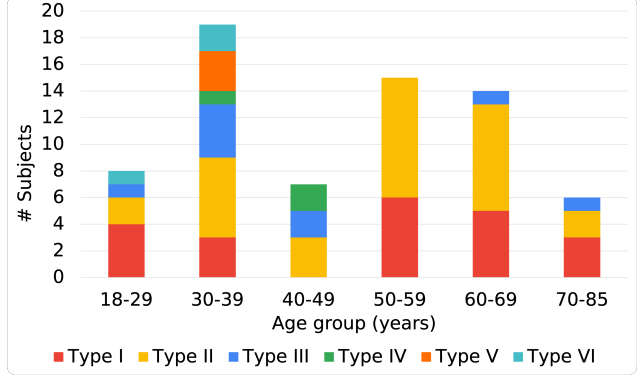


Figure 4. Demographic distributions of the dataset for Age and skin color (Fitzpatrick scale [16])



Figure 5. Gait analysis data: RGB images (top) and generated silhouettes (bottom). Left four columns show walking direction variations; right four columns show walking style and appearance variations.

4. Experimental Setup

4.1. Face Recognition

ISO/IEC 19795 Standardized Metrics To ensure scientific rigor and operational relevance, the performance of the LDFR system is characterized using standardized biometric metrics. These metrics allow for a precise analysis of the trade-off between security (False Match Rate (FMR)) and convenience (False Non-Match Rate (FNMR)).

Protocol definition Focusing on border scenarios, we enroll using mobile phone data (Fig. 2a). Probes are categorized as High-Quality (HQ, 100m with optical zoom) or Low-Quality (LQ, default lens, no optical assistance). LQ probes vary by camera position (Ground-Floor [GF] vs. 3rd Floor [3F]) and subject movement (Walking Towards Camera [WTC] from 100m to 3m). Table 2 details the specific enrollment and probe configurations for each protocol.

4.2. Gait Recognition

The challenge in long distance gait recognition is that the motion of different parts get less detailed. Therefore, with

our dataset we want to study the efficacy of gait recognition models under these difficult scenarios. Similar to other datasets [40, 14, 24], we record sequences with variations like change of piece of clothing, carrying a bag, and gesture variation, i.e. normal walk vs walking while talking on the phone. There are two separate sessions to further include diversity in the scene and person’s appearance.

Protocol and Metrics The walking pattern is broken down into 4 directions(**V1-V4** in fig. 5) where the camera sees mainly the frontal or back part of the person or the two sides. This setup is helpful to analyze models behavior when the probe and gallery sequences are from different views. As shown in fig. 1a, we record the same walking pattern from two camera levels, the third floor of a building(**G1**) and the ground floor(**G2**). The two ground floor sessions are recorded from a different location. This leads to a drastic change of viewpoint between sessions. The gallery consists of only one session of the normal walking sequence per id. This choice reflects authentic operational constraints: because high-quality personal gait enrollment is practically unfeasible at remote registration, the system must utilize standard baseline surveillance captures as the gallery template. The probe set is entirely made of the second session. The probe includes all the other variations that include per id (a) normal walking sequence from the second session.(b) walking with a phone. (c) carrying a bag.(d) change of clothing. The probe set is entirely made of the second session. For each sequence, we do a cross evaluation of the walking direction, for e.g. V1 direction is only compared with V2-V4 directions. Rank-1 accuracy is used to evaluate performance for a closed id set.

5. Benchmarking Results

5.1. Long-Distance Face Recognition

To evaluate robustness against Long Distance Face Recognition(LDFR)-induced degradation, we benchmark four SOTA models (Table 3) representing a computational spectrum: EdgeFace [11] (18.22M parameters) trained on WebFace12M, for edge-deployment, AdaFace IR50/IR101 (43–65M) [22] trained on WebFace4M and WebFace12M respectively, for quality-adaptive robustness, and LV-Face [39] (140M+) trained on Glint360K, for maximum accuracy. This selection enables a comparative analysis of how diverse architectures, ranging from lightweight hybrids to large-scale Vision Transformers, utilize specialized loss functions to mitigate the effects of atmospheric turbulence and resolution loss inherent in long-distance sensing.

The HQ faces using optical zoom (HQ-1, HQ-3, HQ-10 and HQ-20): All these protocols use the same set of probes, that is the HQ faces captured using an optical zoom

Model	Arch.	Params	Innovation	Target
IR50	ResNet	43.6M	Adaptive Margin	Robustness/Latency
IR101	ResNet	65.2M	Deeper Backbone	High Accuracy
EdgeFace	Hybrid	18.2M	LoRaLin/STDA Encoder	Edge/IoT
LV-Face	ViT	140M+	Hybrid Loss	Max Accuracy

Table 3. Comparison of the selected Face Recognition models.

Model	HQ-1		HQ-3		HQ-10		HQ-20	
	1%	0.1%	1%	0.1%	1%	0.1%	1%	0.1%
IR50	1.3	1.8	1.1	1.6	0.9	1.2	0.9	1.1
IR101	0.9	1.3	0.9	1.1	0.8	1.0	0.8	0.9
EdgeFace	1.8	2.5	1.5	2.1	1.3	1.8	1.3	1.8
LVFace	0.9	1.2	0.5	0.7	0.4	0.6	0.4	0.6

Table 4. FNMR (%) evaluation results for HQ protocols. Model thresholds were calibrated on IJB-C at FMR = 1% and 0.1%.

attached to the camera. Table 4 shows the FNMR performance metrics for the four FR systems. The operational thresholds used were calibrated on IJB-C at FMR = 1% and 0.1%. We can notice that all the four system achieve very good recognition results, with slight advantage for LVFace and IR101. This indicates that when baseline optical resolution is maintained by the camera lens, larger models effectively leverage the preserved fine-grained facial features. Performance improves slightly as enrollment samples increase; LVFace’s FNMR drops from 0.9% (HQ-1) to 0.4% (HQ-20). Notably, using only three samples (HQ-3) yields results comparable to using twenty (HQ-20), suggesting a performance cap.

Face captured from the Ground Floor (GF) at 100m distance (GF-3, GF-10): In this protocol, we extracted probe frames from a video sequence. We used 1, 10, 20 and 40 frames to investigate how increasing the number of probe images would affect the performance. The overall results on this protocol were not as good as the ones from HQ protocol due to the difference of face resolution as shown in the example in figure 2-d.

Figure 6 shows the impact of probe frame count and enrollment strategy on long-distance face recognition. The plot illustrates the True Match Rate (TMR) at a pre-calibrated operational threshold (FMR=1% on IJB-C) as the number of low-quality probe frames increases from 1 to 40. Notably, as shown in the performance trends, increasing the number of fused probe frames and utilizing a richer enrollment consistently improves recognition accuracy across all evaluated architectures. We observe that the best overall performance is achieved by the IR101 and EdgeFace models under the optimal condition of a 10-Mix enrollment combined with 40 fused probe frames. Interestingly, EdgeFace outperformed the heavier LV-Face. This suggests a critical domain gap; while SOTA models excel when provided sufficient optical detail, lightweight architectures may exhibit greater robustness against extreme res-

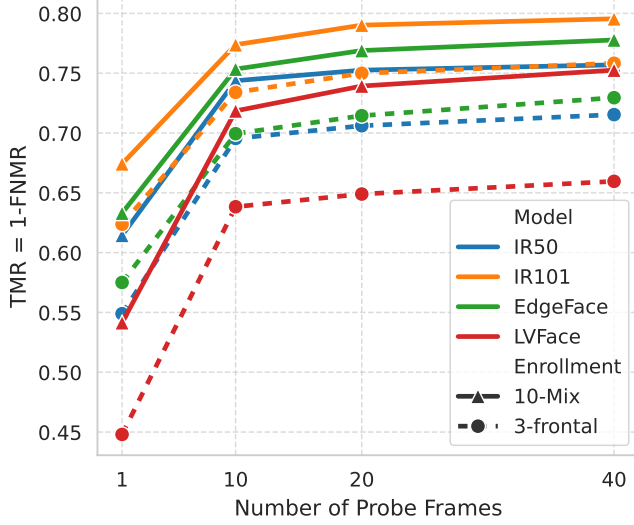


Figure 6. **Protocol GF variants (GF-3, GF-10)**: Robustness to low-quality probes through multi-frame fusion. Colors distinguish the four evaluated models (IR50, IR101, EdgeFace, LVFace). Line styles and markers denote the enrollment protocol: solid lines with triangles indicate the rich 10-Mix strategy, while dotted lines with circular markers represent the “3-frontal” strategy.

olution loss, likely avoiding the reliance on high-frequency details that hinders larger models under severe degradation. Despite this configuration yielding the highest accuracy, the peak True Match Rate (TMR) only approximates to 80%.

Face captured from the 3rd Floor (3F): Extreme resolution loss at 100m causes severe degradation across all FR models. As shown in Fig. 7, ROC curves collapse toward the chance line, indicating a systemic inability to distinguish genuines from zero-effort impostors. This failure is reflected in high EERs: 36.7% (IR101), 37.7% (IR50), 40.2% (EdgeFace), and 43.9% (LVFace). These results demonstrate that standard architectures are practically ineffective at this range, establishing a challenging benchmark for future long-distance recognition research.

Walking towards the camera (WTC): This protocol evaluates the effect of the distance on selected face recognition algorithm. We will use the same four recognition systems listed in table 4. For each video used in this experiment, the first frames correspond when the person was at around 100 meters away from the camera, while the last frames correspond to when the person was close to the camera (less than 4 meters) (examples in Figures 1c and 3).

Figure 8 illustrates a monotonic increase in genuine similarity scores as subjects approach the camera (moving from Frame 1 to Frame 100). At extreme distances, the genuine interquartile ranges (IQR) and the impostor 90% bands severely overlap, but as the subject closes the distance, these distributions successfully separate. This

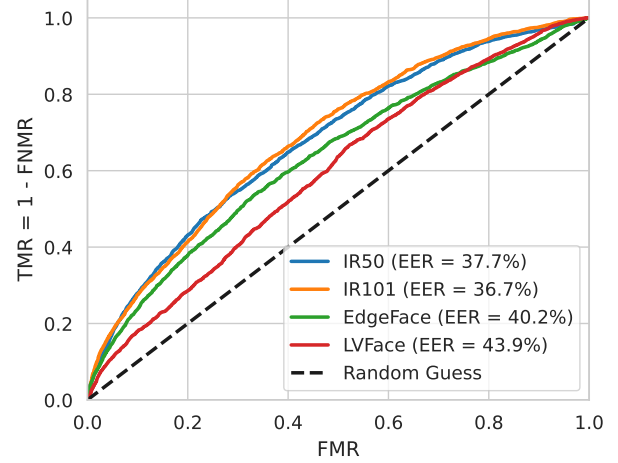


Figure 7. **Protocol 3F**: ROC curves for four face recognition systems. All curves follow the chance line ($AUC \approx 0.5$), demonstrating failure to separate genuine and zero-effort impostor scores.

physical improvement in image quality and discriminability translates directly into enhanced verification accuracy, as demonstrated in Figure 9. While all systems achieve near-perfect True Match Rates (TMR) at close range, their robustness degrades at farther distances and is highly sensitive to the chosen operational security level; imposing stricter thresholds (e.g., $FMR = 0.01\%$) triggers an earlier collapse in TMR. Notably, while the heavier architectures (AdaFace-IR101 and IR50) demonstrate the greatest resilience within this performance corridor, EdgeFace, despite being a lightweight model, consistently outperforms the heavier LVFace architecture across all distances and thresholds. Nevertheless, all evaluated models ultimately fail to sustain reliable recognition at the farthest captured frames, underscoring the profound challenge of long-distance facial recognition.

5.2. Long-Distance Gait Recognition

We proceed to show baselines using state-of-the-art models on gait recognition and discuss the findings.

Baselines. We compare the performance of silhouette-based methods like GaitBase [8], DeepGaitV2 [7], and GaitGL[25]. GaitBase and DeepGaitV2 both are trained on a large GREW [14] dataset. GaitBase [8] uses a ResNet-like (88M params) backbone and creates temporal and spatial features for a given sequence of silhouette frames. Although it was a simple baseline, it improved gait recognition on standard benchmarks. DeepGaitV2 [7](95M params) proposed novel residual layers to increase the depth of the network and improve the recognition performance. GaitGL[25] on the other hand is trained with a smaller CASIA-B[40] dataset. GaitGL architecture is designed to address the limitations of traditional

Method	Train	Condition	Rank-1 Acc. on G1					Rank-1 Acc. on G2				
			V1	V2	V3	V4	Mean	V1	V2	V3	V4	Mean
GaitBase	GREW	Normal	28.10	20.10	29.05	29.95	26.80	14.36	26.15	25.64	17.95	21.02
		Phone	20.00	17.65	23.33	26.57	21.89	8.99	27.96	24.87	13.23	18.76
		Bag	25.24	18.57	20.95	28.57	23.33	11.11	25.25	25.25	18.18	19.95
		Clothes	23.67	24.15	21.26	27.05	24.03	7.58	20.00	23.74	12.12	15.86
DeepGaitV2	GREW	Normal	27.62	19.12	21.91	29.47	24.53	25.13	28.21	36.92	24.62	28.72
		Phone	27.14	15.69	19.52	28.02	22.59	19.05	27.96	31.75	27.52	26.57
		Bag	22.86	14.29	14.76	26.19	19.52	21.72	24.24	34.34	19.70	25.00
		Clothes	24.64	15.46	15.46	28.50	21.01	19.19	26.15	29.29	20.71	23.84
GaitGL	CASIA-B	Normal	10.47	8.33	15.72	14.49	12.25	5.13	14.36	11.80	2.57	8.46
		Phone	10.00	9.80	17.14	13.04	12.50	4.76	9.68	8.99	4.76	7.05
		Bag	9.52	7.14	13.81	13.81	11.07	3.03	11.62	10.61	3.03	7.07
		Clothes	9.66	9.18	14.49	14.98	12.08	5.05	11.28	8.59	4.55	7.37
BiggerGait	CCPG	Normal	6.67	6.86	11.43	10.14	8.78	1.93	3.98	3.92	2.90	3.18
		Phone	10.48	2.94	8.57	8.70	7.67	1.99	2.53	3.98	3.49	3.00
		Bag	5.24	8.09	13.33	11.90	9.64	2.94	3.92	5.39	1.96	3.55
		Clothes	10.63	9.18	11.90	10.95	10.67	1.93	5.32	4.83	3.38	3.86

Table 5. Rank-1 accuracy (%) on G1 and G2 gait protocol. Probe sequences: Normal, Phone, Bag, Clothes. Gallery: Normal. V1-V4 denote the four camera views. GaitBase [8] (88M) and DeepGaitV2 [7] (95M) are trained on GREW [14], GaitGL [25](3.1M) on CASIA-B [40], and BiggerGait [37](140M) on CCPG [24].

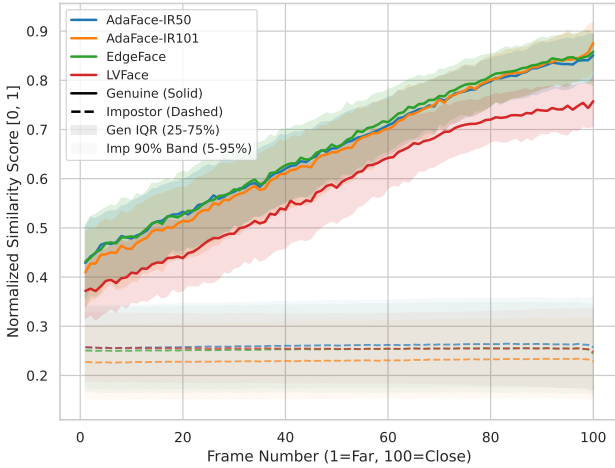


Figure 8. **Protocol WTC**: Normalized similarity score decay across distance (Frame 1 = Far, 100 = Close). Solid and dashed lines denote genuine and impostor medians, respectively, while shaded regions represent the genuine interquartile range (IQR: 25–75%) and the impostor 90% band (5–95%).

gait recognition by simultaneously extracting both coarse-grained global features and fine-grained local details from silhouette sequences. This model is significantly smaller number of parameters than GaitBase i.e 3.1M. Additionally, we benchmark BiggerGait [37] trained on RGB-images from CCPG [24] dataset. BiggerGait uses the layer-wise features of foundation models such as DinoV2[33] to fine-tune a shallow gait encoder network. In comparison to

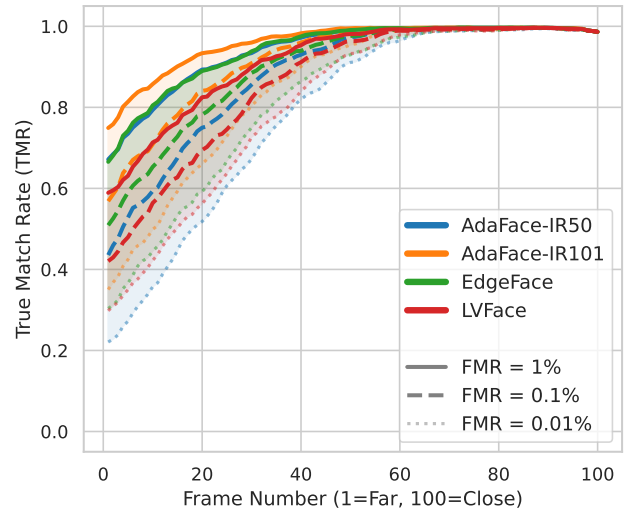


Figure 9. **Protocol WTC**: True Match Rate (TMR) degradation as a function of distance across three False Match Rate (FMR) thresholds. The shaded performance corridors bound the system accuracy from the most lenient (FMR=1%, solid top line) to the strictest (FMR=0.01%, dotted bottom line) operational settings. Thresholds were calibrated on IJB-C dataset.

silhouette-based methods, they achieve better performance on standard short distance gait benchmarks [24, 40] but with a parameter size of 120M. We use OpenGait[8] framework to evaluate these methods on our dataset and use their publicly available trained weights.

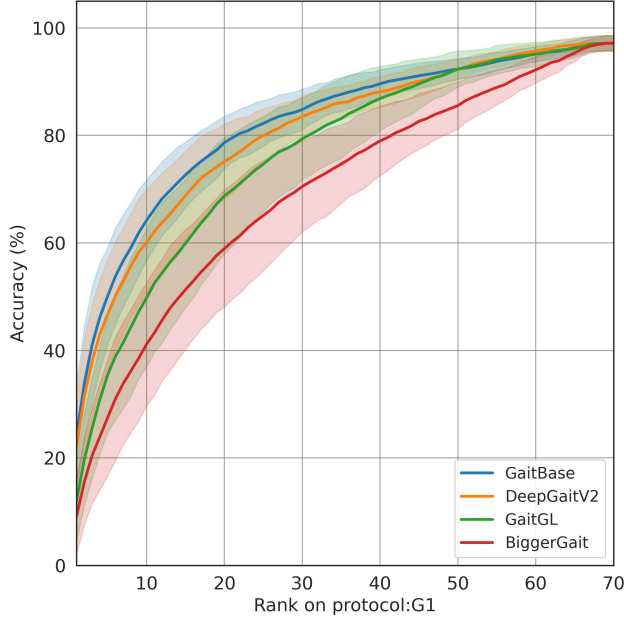


Figure 10. CMC plot for gait recognition protocol G1: Here GaitBase [8] performs the best.

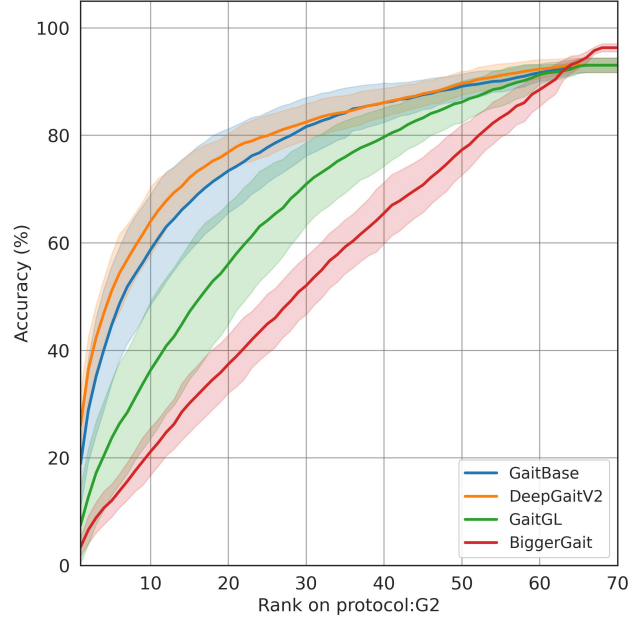


Figure 11. CMC plot for gait recognition protocol G2: Here DeepGaitV2 [7] performs the best. This protocol is difficult than G1 as there is an extreme viewpoint change between two session.

Results. We report in table 5, Rank-1 accuracy across our defined protocols. Overall, both silhouette- and rgb-images-based methods generalize poorly on our dataset, highlighting the challenges of long distance gait recognition. When the camera viewpoint between sessions is small, as in G1, the performance is higher. With respect to covariate changes such as phone, bag, or clothing, performance is worse than the normal walk sequences. However, the impact of clothing change is less in BiggerGait than in silhouette-based methods as it is trained on a large CCPG [24] dataset, which contains large appearance changes. The RGB-based BiggerGait severely underperforms compared to the silhouette based method because of the domain-gap between training CCPG dataset and ours. This difference is less drastic in the binary-mask silhouettes used in GaitBase, DeepGaitV2 and GaitGL. Through, CMC plots figs. 10 and 11, we see that GaitBase [8] performs best under G1 protocol and DeepGaitV2 [7] under challenging G2 protocol. However, all of the models fail to achieve satisfactory generalization on long distance gait recognition. This encourages more research in this direction.

6. Conclusion and Future Work

In this paper, we introduced GaitFace, a novel multimodal dataset explicitly designed to address the challenges of long-range person identification in realistic border crossing and surveillance scenarios. By combining controlled, high-quality mobile pre-enrollment with highly unconstrained outdoor probes, GaitFace exposes the opera-

tional vulnerabilities of current biometric systems. Our extensive benchmarking of state-of-the-art face recognition models demonstrated systemic performance degradation at extreme distances (up to 100 meters) and elevated viewpoints, despite maintaining high accuracy when assisted by optical zoom. Similarly, gait recognition baselines struggled to generalize across the viewpoint shifts, clothing variations, and carried objects introduced in our multi-session protocols. These baseline evaluations underscore that unconstrained long-distance biometric recognition remains an open and highly challenging problem. As future work, we plan to explore the fusion between face and gait modalities to leverage their complementary strengths, aiming to achieve reliable identification even when individual biometric cues are severely degraded by distance or atmospheric interference. We hope that the public release of the GaitFace dataset will stimulate the broader research community to develop more robust biometric solutions.

Acknowledgment The authors would like to thank all the volunteers who participated in the data collection for their continuous support for our research activities. We acknowledge the funding provided by the following projects: Frontex under the Frontex Research Grants Programme, Call 2024/CFP/INNOVATE/01, Grant Agreement No. 2025/280. CarMen project, HORIZON-CL3-2023-BM-01, no. 101168325 and PopEye project, HORIZON-CL3-2023-BM-01, no. 101168317.

References

- [1] X. An, X. Zhu, Y. Gao, Y. Xiao, Y. Zhao, Z. Feng, L. Wu, B. Qin, M. Zhang, D. Zhang, and Y. Fu. Partial fc: Training 10 million identities on a single machine. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1445–1449, 2021.
- [2] H. M. L. Aung, C. Pluempitiwiriyaewej, K. Hamamoto, and S. Wangsiripitak. Multimodal biometrics recognition using a deep convolutional neural network with transfer learning in surveillance videos. *Computation*, 10(7), 2022.
- [3] Z. Cheng, X. Zhu, and S. Gong. Low-Resolution Face Recognition. *arXiv preprint arXiv:1811.08965*, pages 1–16, nov 2018.
- [4] Z. Cheng, X. Zhu, and S. Gong. Surveillance face recognition challenge. *arXiv preprint arXiv:1804.09691*, 2018.
- [5] D. Cornett, J. Brogan, N. Barber, D. Aykac, S. Baird, N. Burchfield, C. Dukes, A. Duncan, R. Ferrell, J. Goddard, G. Jager, M. Larson, B. Murphy, C. Johnson, I. Shelley, N. Srinivas, B. Stockwell, L. Thompson, M. Yohe, R. Zhang, S. Dolvin, H. J. Santos-Villalobos, and D. S. Bolme. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 593–602, 2023.
- [6] European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016.
- [7] C. Fan, S. Hou, Y. Huang, and S. Yu. Exploring deep models for practical gait recognition. *arXiv preprint arXiv:2303.03301*, 2023.
- [8] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang, and S. Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9707–9716, 2023.
- [9] H. Fu, W. Kang, Y. Zhang, and M. S. Shakeel. Fusion of gait and face for human identification at the feature level. In *Biometric Recognition: 16th Chinese Conference, CCBR 2022, Beijing, China, November 11–13, 2022, Proceedings*, page 475–483, Berlin, Heidelberg, 2022. Springer-Verlag.
- [10] X. Geng, L. Wang, M. Li, Q. Wu, and K. Smith-Miles. Adaptive fusion of gait and face for human identification in video. In *Proceedings of the 2008 IEEE Workshop on Applications of Computer Vision*, WACV ’08, page 1–6, USA, 2008. IEEE Computer Society.
- [11] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, and S. Marcel. Edgeface: Efficient face recognition model for edge devices. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 6(2):158–168, 2024.
- [12] M. Grgic, K. Delac, and S. Grgic. Sface — surveillance cameras face database. *Multimedia Tools Appl.*, 51(3):863–879, Feb. 2011.
- [13] M. Günther, P. Hu, C. Herrmann, C. H. Chan, M. Jiang, S. Yang, A. R. Dhamija, D. Ramanan, J. Beyerer, J. Kitzler, M. A. Jazaery, M. I. Nouyed, G. Guo, C. Stankiewicz, and T. E. Boulton. Unconstrained face detection and open-set face recognition challenge. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 697–706, 2017.
- [14] X. Guo, Z. Zhu, T. Yang, B. Lin, J. Huang, J. Deng, G. Huang, J. Zhou, and J. Lu. Gait recognition in the wild: A large-scale benchmark and nas-based baseline. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [15] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *ArXiv*, abs/1607.08221, 2016.
- [16] V. Gupta and V. K. Sharma. Skin typing: Fitzpatrick grading and others. *Clinics in Dermatology*, 37(5):430–436, 2019. The Color of Skin.
- [17] H.-J. Hsu and K.-T. Chen. Droneface: An open dataset for drone research. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys’17, page 187–192, New York, NY, USA, 2017. Association for Computing Machinery.
- [18] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Marseille, France, Oct. 2008. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie.
- [19] A. Kale, A. Roychowdhury, and R. Chellappa. Fusion of gait and face for human identification. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages V–901, 2004.
- [20] N. D. Kalka, B. Maze, J. A. Duncan, K. OrConnor, S. Elliott, K. Hebert, J. Bryan, and A. K. Jain. Ijb-s: Iarpa janus surveillance video benchmark. *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9, 2018.
- [21] N. Khan and M. Efthymiou. The use of biometric technology at airports: The case of customs and border protection (cbp). *International Journal of Information Management Data Insights*, 1(2):100049, 2021.
- [22] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022.
- [23] R. D. Labati, A. Genovese, E. Muñoz, V. Piuri, F. Scotti, and G. Sforza. Biometric recognition in automated border control: A survey. *ACM Comput. Surv.*, 49(2), June 2016.
- [24] W. Li, S. Hou, C. Zhang, C. Cao, X. Liu, Y. Huang, and Y. Zhao. An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13824–13833, June 2023.
- [25] B. Lin, S. Zhang, M. Wang, L. Li, and X. Yu. Gaitgl: Learning discriminative global-local feature representations for gait recognition. *arXiv preprint arXiv:2208.01380*, 2022.
- [26] F. Liu, N. Chimitt, L. Guo, J. Jain, A. Kane, M. Kim, W. Robbins, Y. Su, D. Ye, X. Zhang, J. Zhu, S. Satyakam, C. Perry, S. H. Chan, A. Ross, H. Shi, Z. Wang, A. Jain, and X. Liu.

- Person recognition at altitude and range: Fusion of face, body shape and gait, 2025.
- [27] L. S. Luevano, L. Chang, H. Méndez-Vázquez, Y. Martínez-Díaz, and M. González-Mendoza. A study on the performance of unconstrained very low resolution face recognition: Analyzing current trends and new research directions. *IEEE Access*, 9:75470–75493, 2021.
 - [28] S. Maity, M. Abdel-Mottaleb, and S. S. Asfour. Multimodal low resolution face and frontal gait recognition from surveillance video. *Electronics*, 10(9), 2021.
 - [29] Y. Martinez-Diaz, H. Mendez-Vazquez, L. S. Luevano, L. Chang, and M. Gonzalez-Mendoza. Lightweight Low-Resolution Face Recognition for Surveillance Applications. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5421–5428, Los Alamitos, CA, USA, Jan. 2021. IEEE Computer Society.
 - [30] Y. Martínez-Díaz, M. Nicolás-Díaz, H. Méndez-Vázquez, L. S. Luevano, L. Chang, M. Gonzalez-Mendoza, and L. E. Sucar. Benchmarking lightweight face architectures on specific face recognition scenarios. *Artificial Intelligence Review*, 54(8):6201–6244, Dec. 2021.
 - [31] B. Maze et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, 2018.
 - [32] D. Muramatsu, H. Iwama, Y. Makihara, and Y. Yagi. Multi-view multi-modal person authentication from a single walking image sequence. In *2013 International Conference on Biometrics (ICB)*, pages 1–8, 2013.
 - [33] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
 - [34] D. Seckiner, X. Mallett, P. Maynard, D. Meuwly, and C. Roux. Forensic gait analysis — morphometric assessment from surveillance footage. *Forensic Science International*, 296:57–66, 2019.
 - [35] E. D. P. Supervisor. *AI Act Regulation (EU) 2024/1689 - Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)*. Publications Office of the European Union, 2025.
 - [36] Z.-Y. Wang, Z. Shao, A. Nanduri, B. Pal, L. McDaniel, J. Chen, and R. Chellappa. Combo-gait: Unified transformer framework for multi-modal gait recognition and attribute analysis, 2026.
 - [37] D. Ye, C. Fan, Z. Huang, C. Luo, J. Li, S. Yu, and X. Liu. Biggertgait: Unlocking gait recognition with layer-wise representations from large vision models. *arXiv preprint arXiv:2505.18132*, 2025.
 - [38] D. Yi, Z. Lei, S. Liao, and S. Li. Learning face representation from scratch. *ArXiv*, abs/1411.7923, 2014.
 - [39] J. You, S. Li, Y. Sun, J. Wei, M. Guo, C. Feng, and J. Ran. Lvface: Progressive cluster optimization for large vision models in face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11840–11849, 2025.
 - [40] S. Yu, D. Tan, and T. 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. IEEE, 2006.
 - [41] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
 - [42] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021.