2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 979-8-3503-1892-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/WACV57701.2024.00611

FarSight: A Physics-Driven Whole-Body Biometric System at Large Distance and Altitude

Feng Liu[1], Ryan Ashbaugh[1], Nicholas Chimitt[5], Najmul Hassan[4], Ali Hassani[3], Ajay Jaiswal[2], Minchul Kim[1], Zhiyuan Mao[5], Christopher Perry[1], Zhiyuan Ren[1], Yiyang Su[1], Pegah Varghaei[1], Kai Wang[3], Xingguang Zhang[5], Stanley Chan[5], Arun Ross[1], Humphrey Shi[3], Zhangyang Wang[2], Anil Jain[1] and Xiaoming Liu[1]

[1] Michigan State University, East Lansing MI 48824, USA
[2] University of Texas at Austin, Austin TX 78712, USA
[3] Georgia Tech, Atlanta GA 30332, USA
[4] University of Oregon, Eugene OR 97403, USA
[5] Purdue University, West Lafayette IN 47907, USA

Abstract

Whole-body biometric recognition is an important area of research due to its vast applications in law enforcement, border security, and surveillance. This paper presents the end-to-end design, development and evaluation of FarSight, an innovative software system designed for whole-body (fusion of face, gait and body shape) biometric recognition. FarSight accepts videos from elevated platforms and drones as input and outputs a candidate list of identities from a gallery. The system is designed to address several challenges, including (i) low-quality imagery, (ii) large yaw and pitch angles, (iii) robust feature extraction to accommodate large intra-person variabilities and large inter-person similarities, and (iv) the large domain gap between training and test sets. FarSight combines the physics of imaging and deep learning models to enhance image restoration and biometric feature encoding. We test FarSight's effectiveness using the newly acquired IARPA Biometric Recognition and Identification at Altitude and Range (BRIAR) dataset. Notably, FarSight demonstrated a substantial performance increase on the BRIAR dataset, with gains of $+11.82\%$ *Rank-20 identification and* $+11.30\%$ *TAR@1% FAR.*

1. Introduction

The aim of whole-body biometric recognition is to develop a person recognition system that will surpass the performance of state-of-the-art (SoTA) recognition of the face, gait, and body shape alone, specifically in the challenging, unregulated conditions present in full-motion videos (*e.g.*, aerial surveillance). It encompasses functionalities such as person detection, tracking, image enhancement, the mitigation of atmospheric turbulence, robust biometric feature encoding, and multi-modal fusion and matching. The wide-ranging applications of whole-body recognition in fields like law enforcement, homeland security and surveillance, further underscore its importance [16, 48, 50, 66].

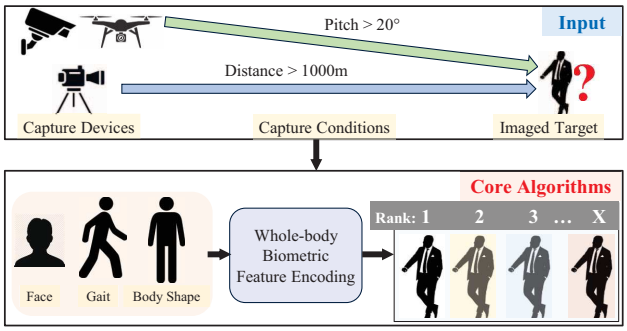


Figure 1. **FarSight** is a person recognition system that implements and fuses SoTA face, gait and body shape recognition modules in challenging conditions presented by full-motion videos.

To achieve these goals, we design, prototype and evaluate a software system called **FarSight** for whole-body (face, gait and body shape) biometric recognition. As illustrated in Fig. 1, FarSight accepts as input a video captured at long-range and from elevated platforms, such as drones, and outputs a candidate list of identities present in the input video.

The design of FarSight confronts a number of novel challenges that have not been adequately addressed in existing literature: i) Low-quality video frames due to long-range capture (hundreds of meters) and atmospheric turbulence

2642-9381/24/$31.00 ©2024 IEEE
DOI 10.1109/WACV57701.2024.00611

(with the refractive index structure parameter C_n^2 in ranges of 10^{-17} to 10^{-14} $m^{-2/3}$ [52]). ii) Large yaw and pitch angles ($>$ 20 degrees) due to elevated platforms (altitudes of up to 400m). iii) Degraded feature sets due to low visual quality (the pixel range for Inter-Pupillary Distance is around $15-100$). iv) Limited domain and paucity of training data due to diversity in the operating environments resulting in a large domain gap between training and test sets.

To address these challenges, the design of FarSight heavily relies on modeling the *underlying physics* of image formation, image degradation and human body models throughout the recognition pipeline. Further, we integrate the learned physics knowledge into the deep learning models for feature encoding. The four key modules of FarSight are 1) image restoration, 2) detection and tracking, 3) biometric feature encoding, and 4) multi-modal fusion.

- Image restoration: Video streams captured from long distances suffer from atmospheric turbulence, platform vibration, and systematic aberrations. Unlike most SoTA approaches that rely on deep learning, we directly model the physics of turbulence. This model not only provides better understanding of imaging limits and turbulence parameters but also enables the creation of datasets for training restoration modules. Consequently, our approach ensures improved explainability and requires fewer labeled samples, leading to superior generalization in unseen environments.
- Detection and tracking: We develop a joint body and face detection module, which is able to associate face and body bounding boxes. Detected bounding boxes can then be fed into an appropriate feature extractor (embedding) without requiring a post-processing stage to match face and body bounding boxes.
- Biometric (face, gait and body shape) feature encoding. (i) Face: We leverage adaptive loss function, two-stage feature fusion, and controllable face synthesis models to effectively manage image quality variation, frame-level feature consolidation, and domain gap. (ii) Gait: We extract both local features and global correlations to improve identification in diverse scenarios. (iii) Body shape: We learn a robust 3D shape representation that is invariant to clothing and body pose variations, leading to improvements in body matching.
- Multi-modal fusion: This module performs score-level fusion and score imputation in case of missing data (when no features could be extracted for one or more biometric modalities), which does occur due to the challenging nature of long range and high angle of inclination videos.

The innovations of **FarSight** system are as follows:

$\diamond$ Explicitly modeling the physics of imaging through turbulence and image degradation and integrating physics-based models into deep learning for image restoration.

$\diamond$ Utilizing a joint body and face detection approach, easily integrated with upstream and downstream tasks.

$\diamond$ An effective feature encoding for face, gait and body shape, along with a novel multimodal feature fusion approach, enabling superior recognition performance.

$\diamond$ Utilizing the Biometric Recognition and Identification at Altitude and Range (BRIAR) dataset [10], we demonstrate the superior performance of the proposed FarSight system, and its robustness and effectiveness in whole-body biometric recognition under challenging conditions.

2. Related Work

Whole-Body Biometrics Recognition. Whole-body biometric recognition merges multiple physical traits, specifically face, gait, and body shape, to bolster identification accuracy, especially in challenging scenarios. Unlike traditional biometric systems focusing on a single trait [9, 12, 14, 17, 22, 26, 35, 61, 64], this comprehensive approach can mitigate inherent weaknesses and exploit the strengths of each individual trait, leading to enhanced recognition performance. For example, while face recognition might struggle with varying poses and lighting, gait can be affected by walking speed and attire. Body shape remains a consistent identifier, though it can vary with clothing and posture. Recent literature [18, 25] have increasingly embraced this multi-faceted approach, but many do not provide comprehensive solutions that include image restoration, detection, tracking, and fusion of modalities. This gap indicates potential for further development in holistic biometric systems, ensuring robust recognition in challenging video conditions.

Physics Modeling of Imaging through Turbulence. Turbulence is modeled as a stochastic phenomenon with its modern form largely based on Kolmogorov [28]. The atmosphere can be modeled as a turbulent volume that perturbs light propagating through it [47, 54]. Since the atmosphere is a stochastic phenomenon, its effect on an image is also stochastic. Drawing realizations from this distribution requires a simulator. Simulating these effects most often comes in the form of mirroring nature: a wave is numerically propagated through a simulated atmosphere. Methods that utilize numerical wave propagation in this manner are referred to as split-step propagation [4, 19, 20, 52]. Alternative methods combine empirical understanding and analysis [30, 43, 45, 46] with some recent modification and improvement [39, 40]. Given the scarcity of open-source tools, we introduce a unique modeling approach.

Image Restoration. Successful biometric recognition relies upon robust feature extraction from sensed imagery [23]. With poor-quality imagery, image restoration serves as a way to extract robust and salient features and potentially boost recognition accuracy. However, restoration methods may *change* the person's identity based on recon-

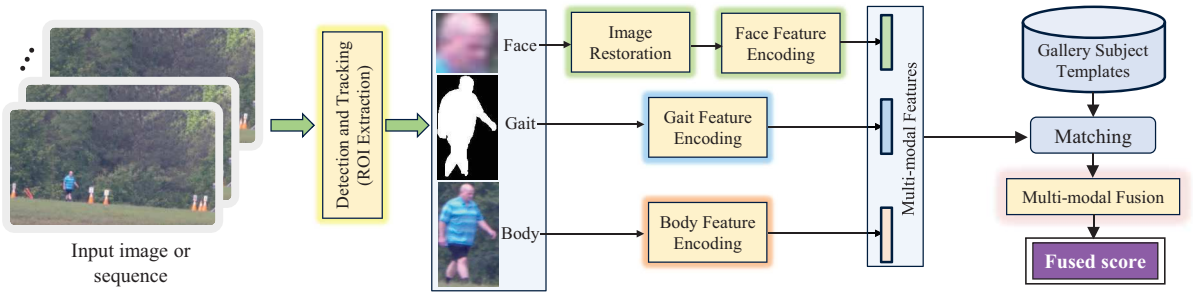


Figure 2. The proposed FarSight system incorporates six components: *detection and tracking, image restoration, face, gait, and body shape feature extraction, and multi-modal biometric fusion*.

structed features as shown in attack-based work [37]. Thus, reconstruction in this biometric context is slightly different. We prefer a reconstructed image that improves downstream recognition performance. Face deblurring in the presence of invariant blur has been shown to have positive results on downstream classification [53]. Furthermore, some efforts in restoration [29,41,59] have suggested that reconstruction may indeed help in the case of atmospheric turbulence degraded images. These methods, however, rely only on single frames, therefore, in the FarSight system we use multi-frame fusion to improve the quality of degraded images.

Detection and Tracking. Face detection has been extensively studied in the field of computer vision, with numerous endeavors aimed at detecting faces across a diverse array of scenes. Various methodologies, as presented in [11, 31, 70], have successfully employed different approaches for detecting faces in unconstrained settings. Building upon this, pedestrian tracking is another significant module in biometrics. A multitude of strategies have been developed to improve both the efficiency and effectiveness of tracking. Among them, tracking by detection paradigms has emerged as the leading approach due to its adaptability and superior performance. Motion-based methods [3, 63, 69] employ spatiotemporal information to enhance object association and improve tracking accuracy. Appearance-based methods [56, 57, 62] introduce various appearance features to facilitate accurate object matching.

Multi-Modal Biometric Fusion. Fusion relies on leveraging encoded biometric features or scores from multiple matchers. An example of a score-level fusion method is the sum rule, where normalized scores are weighted and summed to generate the fused score to be used for performance evaluation [21,49].

3. FarSight: System Architecture

3.1. Overview of FarSight

As illustrated in Fig. 2, FarSight operates through six modules: detection and tracking, image restoration, face, gait, and body shape feature extraction, and multi-modal fusion. These modules work within a scalable testing framework, optimizing GPU usage via adaptable batch sizes. An API utility facilitates communication between the framework and external systems, transmitting video sequences from configuration files to the framework via Google RPC calls. Essential features extracted from these sequences are stored in HDF5 files for performance evaluation.

The workflow starts with input video sequences undergoing detection and tracking. Regions of interest (RoI) are identified and forwarded to gait and body modules, with face images undergoing restoration. Gait and body modules produce unique feature vectors via average pooling, while the face module, using CAFace [27], consolidates features across sequences. A probe comprises a single video segment per subject, while gallery enrollments – multiple video sequences and stills – are merged into a singular feature vector for each modality.

3.2. Challenges in FarSight

The FarSight system faces distinct challenges. Captured videos often suffer from poor quality due to long-range capture and atmospheric turbulence. Elevated platforms introduce large yaw and pitch angles, making data analysis more challenging. Extracting identity features is affected by low visual quality, and the training data's limited domain further complicates the learning task. Further, the lack of transparency in deep learning models poses a significant issue. Fig. 3 illustrates these challenges with examples from close-range, mid-range (100-500m), and UAV-captured scenarios.

3.3. Physics Modeling of Turbulence

Atmospheric turbulence is an unavoidable degradation when imaging at range. It is often computationally modeled by splitting the continuous propagation paths into segments via phase screens as illustrated in Fig. 4. While accurate, the spatially varying nature of the propagation makes this a computationally demanding process [19,20,52].

More recent works have explored the possibility of *propagation-free* models where the turbulence effects are implemented as random sampling at the *aperture* [7,8,38]. As shown in Fig. 4, every pixel on the aperture is associated with a random phase function which has a linear rep-

Figure 3. Example frames in the BRIAR dataset [10] showing the same subject (identity) under various conditions, including different standoff distances, clothing, and image quality due to the turbulence effect. The columns represent different scenarios: controlled conditions, close range, 100m-set1, 100m-set2, 200m, 400m, 500m, and UAV capture, respectively.

resentation using the Zernike polynomials [42]. By constructing the covariance matrix of the random process, we can draw samples of the Zernike coefficients to enforce spatial and modal correlations. Propagation-free simulation has enabled 1000× speed up compared to the split-step propagation methods while maintaining accuracy. Therefore, we adopt this simulation approach in our system.

For the generation of training data, realistic optical and turbulence parameters significantly influence the appearance of the generated defects. Therefore, our datasets are synthesized according to the metadata of various long-range optical systems. Our training dataset also consists of both dynamic and static scenes [24, 51, 68].

3.4. Detection and Tracking

Our detection module, based on [55], uses a two-stage R-CNN detector [44] with a modified ResNet50 backbone to associate face and body bounding boxes [55]. This is done using associative embeddings to match faces and bodies, learned via **pulling** and **pushing** loss functions [13]. The pulling loss brings embeddings of the same subject closer in the presence of intra-subject variations, calculated as body-to-body, face-to-face, and face-to-body pairs. These are combined using a weighted sum of body-to-face loss, and the sum of face-to-face and body-to-body losses. Pushing loss, in contrast, pushes away bounding boxes assigned to different subjects to account for inter-subject variations. It is divided into three losses between pairs of body boxes, pairs of face boxes, and body-face pairs. These losses are combined by a weighted sum. The final associative embedding loss used to optimize these embeddings is a weighted sum of the pulling and pushing losses.

The module also predicts "head hook" coordinates for every subject to improve body and face association. The

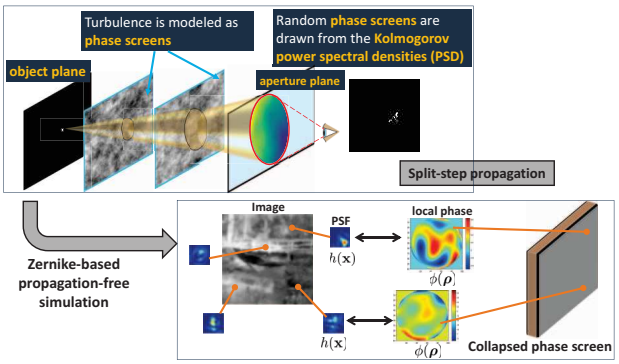


Figure 4. Turbulence modeling. Comparing split-step [5, 19] and Zernike-based simulations [7, 8, 38].

head hook loss is a weighted sum of the Smooth L1 loss [15] and a scale-invariant angular loss. The final association between body and face bounding boxes is based on similarity metrics, including embedding distance, head hook distance, and confidence scores. The RBF kernel is used for both the embedding distance and head hook distance. The confidence scores factor directly into the association loss to mitigate associating low-confidence bounding boxes with high-confidence ones. Finally, all these metrics are integrated into a final association metric. If a face prediction's maximum similarity score with any body is below a set threshold, it is concluded that the subject's face is not visible.

3.5. Image Restoration

Image restoration aims to reverse the image formation process, as described by the equation [6]

$$I(\mathbf{x}) = [\mathcal{B} \circ \mathcal{T}](J(\mathbf{x})), \tag{1}$$

where, $\mathcal{T}$ is the tilt operator and $\mathcal{B}$ represents the blur operation, with $J(\mathbf{x})$ and $I(\mathbf{x})$ as the input and output images, indexed by position $\mathbf{x}$, respectively. In this work, we have considered a single-frame image restoration method as well as a multi-frame method, both aiming to invert $\mathcal{T}$ and $\mathcal{B}$.

Our restoration methods for biometrics focus on preserving identity, using lightweight, real-time techniques. These are divided into single-frame and multi-frame restorations. The former provides lower throughput but relies on strong priors without altering the subject's identity. Multi-frame restoration, on the other hand, utilizes temporal cues, allowing weaker priors but requiring larger throughput.

Our multi-frame approach uses the Recurrent Turbulence Mitigation network (RTM), a bi-directional, multi-scale convolutional recurrent network with a novel Multi-head Temporal Channel self-attention (MTCSA) layer (Fig. 5).

3.6. Multi-Modal Biometric Feature Encoding

We describe here our methods for obtaining biometric features from the face, gait and body shape, as well as

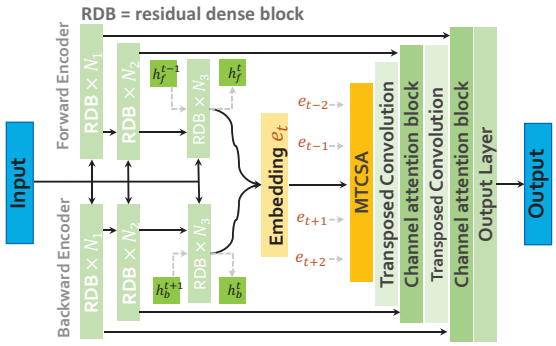


Figure 5. Multi-frame image restoration by the recurrent network for turbulence mitigation (RTM).

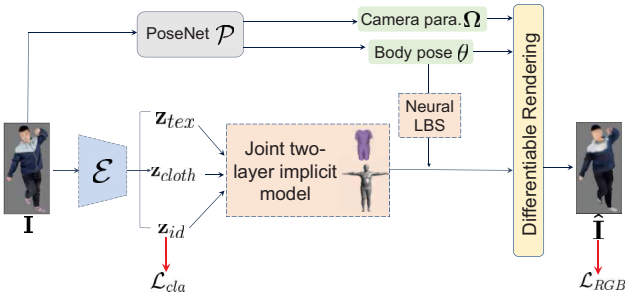


Figure 6. Overview of the proposed body shape feature encoding framework (3DInvarReID [33]). In the body matching process, the identity shape features $\mathbf{z}_{id}$ are utilized for matching.

the multi-modal fusion technique applied to generate fused scores for evaluation on the metrics described in Sec. 4.

3.6.1 Face

Our face recognition pipeline integrates the techniques of Adaptive Margin Function (AdaFace [26]), Cluster and Aggregate (CAFace [27]), and Controllable Face Synthesis Model (CFSM [34]), addressing the challenge of recognizing faces across variable image qualities and media types.

Initially, AdaFace [26], an adaptive loss function strategy, helps manage low-quality face datasets. It adjusts the emphasis on misclassified samples based on image quality, effectively dealing with a wide range of image quality levels. Next, CAFace [27], a two-stage feature fusion technique, is crucial for integrating features from multiple frames. By grouping inputs to a few global cluster centers and subsequently fusing these features, CAFace maintains order invariance while combining multiple frames. Lastly, CFSM [34] helps bridge domain gaps between training and testing scenarios. It replicates the target datasets' distribution in a style latent space, generating synthetic face images similar to the target evaluation datasets, thereby reconciling the disparity between high-quality training data and lower-quality surveillance images. The combination of AdaFace, CAFace, and CFSM effectively navigates the challenges of face recognition across diverse image qualities, leveraging feature extraction, feature integration, and synthetic image generation to improve face recognition performance.

3.6.2 Gait

We propose an innovative framework, GlobalGait, to address the limitations of existing gait recognition models that mainly focus on local features and often overlook vital global correlations. GlobalGait enriches these local features by factoring in global correlations across a gait sequence, thereby boosting recognition accuracy.

Given an input sequence, GlobalGait uses a CNN backbone to extract local spatiotemporal features, and then divides them into source and target features. These feature maps are projected into tokens for each joint, using sampling around each 2D joint. We employ a stack of multi-head self-attention layers to model the sequences' spatial and temporal correlations. Further, GlobalGait attempts to reconstruct target frame pixels based on source sequences and to choose the correct target sequence from a set of candidates. This approach harnesses the spatial and temporal correlations in gait recognition, with these supervisory signals guiding the model to learn more distinct gait features.

3.6.3 Body Shape

Our method (3DInvarReID [33]) for encoding body features harnesses the power of Person Re-ID [2,32,60,65], with the primary aim to effectively capture static body features. We posit that the most reliable cue for body matching is the naked 3D body shape, despite the considerable challenges in reconstructing it from a 2D image. Taking cues from advancements in 3D feature learning, we introduce a pipeline to disentangle identity (naked body) from non-identity components (pose, clothing shape and texture) of 3D clothed humans. The core of our approach lies in a novel joint two-layer neural implicit function that disentangles these components in latent representations.

As illustrated in Fig. 6, given a training set of T images $\{\mathbf{I}_i\}_{i=1}^{T}$ and the corresponding identity labels $\{l_i\}_{i=1}^{T}$, the image encoder $\mathcal{E}(\mathbf{I}) : \mathbf{I} \rightarrow (\mathbf{z}_{id}, \mathbf{z}_{cloth}, \mathbf{z}_{tex})$ predicts the identity shape code of naked body $\mathbf{z}_{id} \in \mathbb{R}^{L_{id}}$, clothed shape code $\mathbf{z}_{cloth} \in \mathbb{R}^{L_{cloth}}$ and texture code $\mathbf{z}_{tex} \in \mathbb{R}^{L_{tex}}$. A joint two-layer implicit model decodes the latent codes to identity shape, clothing shape, and texture components, respectively. Additionally, PoseNet $\mathcal{P}$ predicts the camera projection $\mathbf{\Omega}$ and SMPL body pose θ: $(\mathbf{\Omega}, \theta) = \mathcal{P}(\mathbf{I})$. Mathematically, the learning objective is defined as:

$$\underset{\mathcal{E},\mathcal{F},\mathcal{C},\mathcal{T}}{\arg\min} \sum_{i=1}^{T} \left(\left| \hat{\mathbf{I}}_i - \mathbf{I}_i \right|_1 + \mathcal{L}_{cla}(\mathbf{z}_{id}, l_i) \right), \qquad (2)$$

where $\mathcal{L}_{cla}$ is the classification loss. $\hat{\mathbf{I}}$ is the rendered image. This objective enables us to jointly learn accurate 3D clothed shape and discriminative shape for the naked body.

We utilize CAPE [36] and THuman2.0 [67] datasets to train our model, generating individual identity shape code, clothing shape code, and texture code for each training sample. For inference, the encoder processes body images to extract identity shape features $\mathbf{z}_{id}$. The Cosine similarity of two $\mathbf{z}_{id}$ determines if two images belong to the same person. This method, excluding the explicit 3D reconstruction during inference, is highly efficient.

3.6.4 Multi-Modal Biometric Fusion

To produce a comprehensive probe-gallery score from multiple biometric modalities, we initially calculate per-modality scores for each probe-gallery pair. For the face, gait, and body, we create a singular subject-level feature using CAFace (Sec. 3.6.1), mean fusion on video-only gallery features, and mean fusion on whole-body media, excluding face-only images, respectively. This exclusion is necessary due to the prevalence of face-only gallery images and the unsuitability of gait recognition on single images. Probe features are then compared to gallery features, and an equal-weighted sum score fusion is employed to generate a single score from the cosine similarity scores of the three modalities. When feature extraction fails for one or more modalities, we impute missing scores to the middle of the score range, which is zero for the cosine similarity metric used in generating probe-gallery scores. This imputation method was chosen after evaluating alternative techniques, with this approach showing the least bias and greatest stability.

4. Experimental Results

All modules are run together in a configurable container environment on PyTorch version 1.13.1. We perform experiments on 8 Nvidia RTX A6000s, with 48 GiB of VRAM, over the course of 48 hours on 2 dual-socket servers with either AMD EPYC 7713 64-Core or Intel Xeon Silver 4314 32-Core processors.

BRIAR Datasets [1] and Protocols. The IARPA BRIAR dataset [10], comprises two collections—BRIAR Government Collections 1 (BGC1) and 2 (BGC2), is a pioneering initiative to support whole-body biometric research. It addresses the necessity for broader and richer data repositories for training and evaluating biometric systems in challenging scenarios. BRIAR consists of over $350,000$ images and $1,300$ hours of videos from $1,055$ subjects in outdoor settings. The dataset, with its focus on long-range and elevated angle recognition, provides a fertile ground for algorithm development and evaluation in biometrics.

The dataset, in accordance with Protocol V2.0.1, has been partitioned into a training subset (BRS, 411 subjects) and a testing subset (BTS, 644 subjects), with non-overlapping subjects. Regarding the test subjects, we utilize the controlled images and videos as gallery, and the field-collected data as probe. The protocol provides for 644 subjects for closed-set search and includes two subsets of 544 subjects each for open-set search, both containing 444 distractors who lack corresponding probe subjects. The probes, totaling $20,432$ templates, are categorized into FaceIncluded and FaceRestricted. FaceIncluded ensures the face is discernible, with at least 20 pixels in head height. FaceRestricted contains data with challenges like occlusions and low resolution.

Metrics. We employ BRIAR Program Target Metrics [1] to measure FarSight's performance across multiple modalities and their fusion: verification (TAR@1% FAR), closed-set identification (Rank-20 accuracy), and open-set identification (FNIR@1% FPIR), allowing for a thorough examination of its performance across various settings.

Baselines. In our study, we utilize established benchmarks for each biometric modality to ensure a comprehensive comparison: For facial recognition, we utilize AdaFace coupled with an average feature aggregation strategy, a popular approach known for its excellent performance [26]. For gait recognition, we adopt GaitBase [14], a solution known for its efficacy. For body shape modality, we employ CAL [17], a SoTA cloth-changing person re-identification method. These benchmarks provide an excellent basis to fairly evaluate our proposed method.

4.1. Evaluation and Analysis

In Tab. 1, we provide a thorough comparison of our approaches and the baselines for each modality. The detailed comparison analysis clearly highlights the superior performance of our proposed FarSight system across all performance metrics when compared to the baselines. For each modality, our module outperforms the baselines by a significant margin. For instance, in the verification metric (TAR@1% FAR) on FaceIncluded sets, FarSight (Face) sees an increase of 11.81%. For gait, there's an improvement of 13.65%, and for body shape, we see an improvement of 2.13%. Further, upon fusion, we gain an additional improvement of 16.78% ($69.15\% \rightarrow 85.93\%$).

The FarSight system's effectiveness across various modalities and distances is evident in Tab. 2, displaying each modality's distinct robustness at different ranges. Especially noteworthy is the integrated FarSight model, exhibiting an outstanding accuracy consistently above 88% across all investigated ranges. The observed increase in face recognition accuracy with distance is tied to the growing similarity between sensors used in training and testing data. As this sensor alignment increases with distance, it reduces the domain gap, leading to enhanced performance. This finding underscores the critical role of sensor type and domain adaptation in optimizing biometric recognition.

[1] All human data is collected in accordance with ethical standards and received approval from IRB.

Method	Verification (1:1) TAR@1% FAR ↑		Rank Retrieval (1:N) Rank-20, Closed Search ↑		Open Search (1:N) FNIR@1% FPIR ↓	
	FaceRestricted	FaceIncluded	FaceRestricted	FaceIncluded	FaceRestricted	FaceIncluded
Baseline-AdaFace [26]	9.61	66.20	14.97	73.85	96.22	70.64
FarSight (Face)	25.04	78.01	31.78	84.12	92.11	57.39
Baseline-GaitBase [14]	44.33	45.55	64.90	68.03	98.53	98.79
FarSight (Gait)	56.23	59.20	72.55	74.64	95.24	95.31
Baseline-CAL [17]	48.58	51.87	66.27	71.18	96.98	96.17
FarSight (Body)	51.02	54.00	69.18	72.91	96.95	96.23
FarSight (Face+Gait)	57.30	83.98	75.15	91.19	**87.64**	**54.55**
FarSight (Face+Body)	54.68	**85.93**	73.97	**93.13**	89.57	58.99
FarSight (Gait+Body)	58.91	62.08	73.06	75.57	94.86	94.74
AdaFace+GaitBase+CAL	51.70	69.15	65.57	80.19	94.92	67.53
FarSight	**63.00**	81.88	**77.39**	91.74	90.66	67.77

Table 1. Whole body biometric recognition results on the BRIAR dataset (N=644 in retrieval and 544 in open-set search).

Probe	Close range	100m	200m	400m	500m	UAV
FarSight (Face)	68.57	66.07	89.47	90.78	86.32	72.51
FarSight (Gait)	75.25	73.49	76.53	74.23	71.41	72.89
FarSight (Body)	72.68	73.25	75.79	77.40	73.91	73.90
FarSight	88.55	88.01	93.26	93.92	91.81	88.15

Table 2. Rank-20 (%) on BRIAR at different altitudes and ranges.

FaceIncluded	TAR@ 1% FAR	Rank-20	FNIR@ 1% FPIR
AdaFace [26]	66.20	73.85	70.64
+ CFSM [34]	67.38	77.22	68.51
+ CAFace [27]	71.54	78.57	61.77
+BRS1 **FarSight (Face)**	78.01	84.12	57.39

Table 3. Ablation of different parts in face recognition pipeline.

TAR@1% FAR	FaceIncluded
Face w/o Restoration	72.39
Face w/ Restoration	**72.57**

Table 4. Face recognition with and without image restoration.

4.1.1 Face

The efficacy of including various modules in the face recognition pipeline is shown in Tab. 3. We initially use the combination of AdaFace IR101 backbone with the average feature aggregation which has shown good performance in low-quality imagery [26]. CFSM [34] adds performance improvement by adopting training data to a low-quality image dataset WiderFace [58] (+1.18 in TAR@1% FAR). CAFace [27] is a feature fusion method that improves upon the basic average pooling (+4.16). Lastly, finetuning the model on the BGC1 training dataset further improves the performance (+6.47). The inclusion of an RTM-based image restoration model, as demonstrated in Table 4, leads to noticeable performance enhancements

4.1.2 Gait

In our gait recognition experiments, we observe consistent improvements compared to GaitBase [14], our baseline, across all four metrics. Our findings demonstrate significant enhancements in the model's ability to accurately verify individuals, with the TAR@1% FAR reaching an impressive improvement of 11.90% in FaceRestricted verification and 13.65% in FaceIncluded verification. Further, the rank-20 metric exhibits notable advancement, showcasing a remarkable increase of 6.61%. Lastly, our model showcases improved performance in open-set search, achieving a noteworthy reduction of 3.29% in FNIR@1% FPIR. These promising outcomes reaffirm the efficacy of FarSight (Gait) to extract more discriminative features based on global features and highlight its potential for reliable and robust biometric identification in real-world applications.

4.1.3 Body

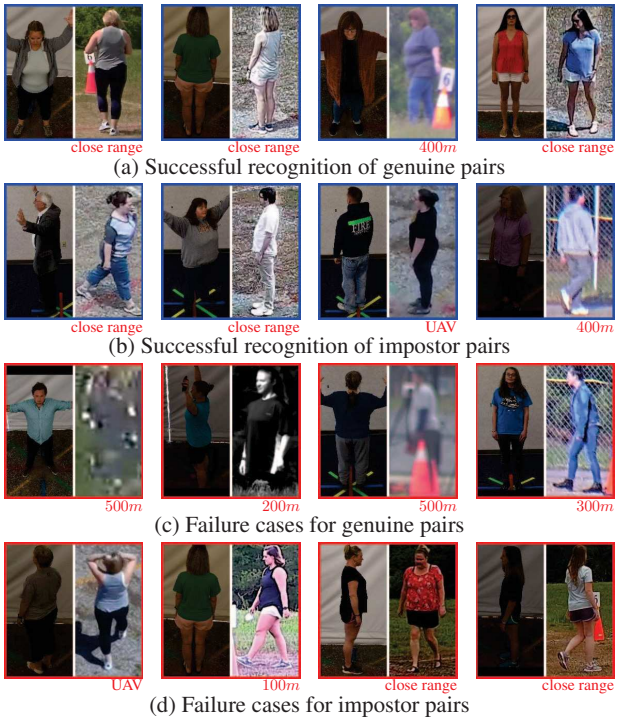
Tab. 1 clearly demonstrates that our FarSight (body) consistently outperforms the CAL baseline on both FaceRestricted and FaceIncluded sets, as evidenced in both verification and Rank retrieval metrics. In Fig. 7, we show successful and failed matches in body matching. Our method copes well with clothing differences, but struggles with motion blur, turbulence, or hairstyle changes. Misidentifications in impostor pairs often happen due to similar body shapes.

4.1.4 Multi-Modal Fusion

As seen in Tab. 1, the fusion of three modalities improves over the next best-performing algorithm in the FaceRestricted condition (+11.30 in TAR@1% FAR and +11.82 in Rank-20). We also see the strength of combining the face and body modalities in the FaceIncluded condition, where face and body fusion excels in both verification and rank retrieval (+1.95 TAR@1% FAR and +1.94 Rank-20) over the next best algorithm. The open search metric performs best when fusing face and gait, scoring 87.64% and 54.55% in FNIR@1% FPIR for both the FaceRestricted and FaceIncluded conditions, which is in part due to the challenge that single body and gait modalities on open-set search.

4.2. System Efficiency

Template Size. Feature vectors for face, gait and body are of sizes 512, 8704 and 6144. Multiplying these values by



(a) Successful recognition of genuine pairs

(b) Successful recognition of impostor pairs

(c) Failure cases for genuine pairs

(d) Failure cases for impostor pairs

Figure 7. Successful and failure examples of body matching.

Module	1080p	4K	Average Combined
Detection & Tracking	20.0	34.7	24.9
Restoration	6.1	5.3	5.9
Face	2.6	2.2	2.5
Gait	3.3	2.5	3.0
Body	3.7	3.1	3.5
FarSight System (fps)	8.4	6.3	7.8

Table 5. FarSight module processing times (sec.) and system efficiency (fps) for 1080p (1920x1080) and 4k (3840x2160) probes.

8 and dividing by 1024 provides the template size: 4KB, 68KB and 48KB, respectively, and 120KB in total.

Processing Speed. The speed of our FarSight system, as outlined in Tab. 5, is examined under stringent conditions to gauge both the efficiency of individual components and the overall pipeline. This system operates asynchronously and concurrently, similar to the actual deployment conditions. To precisely measure efficiency, the components are assessed in a serialized manner, even though they typically run in parallel. We conduct this assessment using representative sample videos, encompassing 2400 frames of 1080p and 1200 frames of 4K video, each set originating from four distinct subjects. The restoration process is primarily directed towards detected faces, which implies that any instances of undetected faces would naturally lead to reduced restoration and face module processing times. A notable observation is that our system can successfully detect bodies in 95% of all frames and faces in 26% of frames.

5. Future Research

Image restoration. We plan to expand our optical simulation tool to handle higher levels of distortion and explore "simulation-in-the-loop" techniques. Our goal is also to balance fidelity and perceptual quality by integrating generative and discriminative restoration methods.

Detection and tracking. We plan to refine our current detector or shift to YOLO-based detectors. We are also considering using separate face detectors on subject bounding boxes to reduce latency.

Biometric feature encoding. In our face module, we are exploring the potential of adaptive restoration based on the available information from given frames, to avoid any negative impact on performance. For our gait module, our goal is to delve further into the usage of 3D body shape and pose information, which is currently under-explored in gait recognition. This involves combining shape parameters with global features to generate 3D-aware shape features and enriching local features with 3D pose information. For body analysis, we aim to refine 3D body reconstructions using multiple frames and assess the value of 3D poses compared to 2D imagery. Future research will encompass additional baselines, including face, gait, and body shape.

Multi-modal fusion. We plan to further enhance our technique for fusing face, gait, and body features, to better exploit the strengths of each modality and alleviate challenges from the long tail of body and gait scores in the non-match open search distributions.

6. Conclusion

We develop and prototype an end-to-end whole-body person recognition system, **FarSight**. Our solution attempts to overcome hurdles such as low-quality video frames, large yaw and pitch angles, and the domain gap between training and test sets by utilizing the physics of imaging in harmony with deep learning models. This innovative approach has led to superior recognition performance, as demonstrated in tests using the BRIAR dataset. With the far-reaching potential to enhance homeland security and forensic identification, the FarSight system paves the way for the next generation of biometric recognition in challenging scenarios.

Acknowledgments. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2022-21102100004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] IARPA-BAA-20-04. https://govtribe.com/file/government-file/iarpa-baa-20-04-briar-final-12-10-2020-c-dot-pdf. Accessed: 2023-06-25. 6

[2] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 5

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, 2016. 3

[4] J. P. Bos and M. C. Roggemann. Technique for simulating anisoplanatic image formation over long horizontal paths. *Optical Engineering*, 2012. 2

[5] Jeremy P Bos and Michael C Roggemann. Technique for simulating anisoplanatic image formation over long horizontal paths. *Optical Engineering*, 2012. 4

[6] Stanley H. Chan. Tilt-then-blur or blur-then-tilt? clarifying the atmospheric turbulence model. *IEEE Signal Processing Letters*, 2022. 4

[7] N. Chimitt and S. H. Chan. Simulating anisoplanatic turbulence by sampling intermodal and spatially correlated Zernike coefficients. *Optical Engineering*, 2020. 3, 4

[8] Nicholas Chimitt, Xingguang Zhang, Zhiyuan Mao, and Stanley H Chan. Real-time dense field phase-to-space simulation of imaging through atmospheric turbulence. *IEEE Transactions on Computational Imaging*, 2022. 3, 4

[9] Patrick Connor and Arun Ross. Biometric recognition by gait: A survey of modalities and features. *CVIU*, 2018. 2

[10] David Cornett, Joel Brogan, Nell Barber, Deniz Aykac, Seth Baird, Nicholas Burchfield, Carl Dukes, Andrew Duncan, Regina Ferrell, Jim Goddard, et al. Expanding accurate person recognition to new altitudes and ranges: The BRIAR dataset. In *WACV*, 2023. 2, 4, 6

[11] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020. 3

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2

[13] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, 2019. 4

[14] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *CVPR*, 2023. 2, 6, 7

[15] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 4

[16] Shaogang Gong and Tao Xiang. *Person re-identification*. Springer London, 2011. 1

[17] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *CVPR*, 2022. 2, 6, 7

[18] Yuxiang Guo, Cheng Peng, Chun Pong Lau, and Rama Chellappa. Multi-modal human authentication using silhouettes, gait and rgb. In *FG*, 2023. 2

[19] Russell C Hardie, Jonathan D Power, Daniel A LeMaster, Douglas R Droege, Szymon Gladysz, and Santasri Bose-Pillai. Simulation of anisoplanatic imaging through optical turbulence using numerical wave propagation with new validation analysis. *Optical Engineering*, 2017. 2, 3, 4

[20] Russell C. Hardie, Michael A. Rucci, Santasri R. Bose-Pillai, Richard Van Hook, and Barry K. Karch. Modeling and simulation of multispectral imaging through anisoplanatic atmospheric optical turbulence. *Optical Engineering*, 2022. 2, 3

[21] Mingxing He, Shi-Jinn Horng, Pingzhi Fan, Ray-Shine Run, Rong-Jian Chen, Jui-Lin Lai, Muhammad Khurram Khan, and Kevin Octavius Sentosa. Performance evaluation of score level fusion in multimodal biometric systems. *Pattern Recognition*, 2010. 3

[22] Yuge Huang, Pengcheng Shen, Ying Tai, Shaoxin Li, Xiaoming Liu, Jilin Li, Feiyue Huang, and Rongrong Ji. Improving face recognition from hard samples via distribution distillation loss. In *CVPR*, 2020. 2

[23] Anil Jain, Karthik Nandakumar, and Arun Ross. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters*, 2016. 2

[24] D. Jin, Y. Chen, Y. Lu, J. Chen, P. Wang, Z. Liu, S. Guo, and X. Bai. Neutralizing the impact of atmospheric turbulence on complex scene imaging via deep learning. *Nature Machine Intelligence*, 2021. 4

[25] Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *CVPR*, 2022. 2

[26] Minchul Kim, Anil K Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *CVPR*, 2022. 2, 5, 6, 7

[27] Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Cluster and aggregate: Face recognition with large probe set. In *NeurIPS*, 2022. 3, 5, 7

[28] A. N. Kolmogorov. The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers. *Doklady Akademii Nauk SSSR*, 1941. 2

[29] Chun Pong Lau, Hossein Souri, and Rama Chellappa. Atfacegan: Single face image restoration and recognition from atmospheric turbulence. In *FG*, 2020. 3

[30] K. R. Leonard, J. Howe, and D. E. Oxford. Simulation of atmospheric turbulence effects and mitigation algorithms on stand-off automatic facial recognition. In *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence VIII*, 2012. 2

[31] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *CVPR*, 2019. 3

[32] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 5

[33] Feng Liu, Minchul Kim, ZiAng Gu, Anil Jian, and Xiaoming Liu. Learning clothing and pose invariant 3D shape representation for long-term person re-identification. In *ICCV*, 2023. 5

[34] Feng Liu, Minchul Kim, Anil Jain, and Xiaoming Liu. Controllable and guided face synthesis for unconstrained face recognition. In *ECCV*, 2022. 5, 7

[35] Feng Liu, Ronghang Zhu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Disentangling features in 3D face shapes for joint face reconstruction and recognition. In *CVPR*, 2018. 2

[36] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3D people in generative clothing. In *CVPR*, 2020. 6

[37] Guangcan Mai, Kai Cao, Pong C. Yuen, and Anil K. Jain. On the reconstruction of face images from deep face templates. *TPAMI*, 2019. 3

[38] Z. Mao, N. Chimitt, and S. H. Chan. Accelerating atmospheric turbulence simulation via learned phase-to-space transform. In *ICCV*, 2021. 3, 4

[39] Kevin J. Miller and Todd Du Bosq. A machine learning approach to improving quality of atmospheric turbulence simulation. In *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXXII*, 2021. 2

[40] Kevin J. Miller, Bradley Preece, Todd W. Du Bosq, and Kevin R. Leonard. A data-constrained algorithm for the emulation of long-range turbulence-degraded video. In *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXX*, 2019. 2

[41] Nithin Gopalakrishnan Nair, Kangfu Mei, and Vishal M. Patel. At-ddpm: Restoring faces degraded by atmospheric turbulence using denoising diffusion probabilistic models. In *WACV*, 2023. 3

[42] R. J. Noll. Zernike polynomials and atmospheric turbulence. *Journal of the Optical Society of America*, 1976. 4

[43] Guy Potvin, J. Luc Forand, and Denis Dion. A simple physical model for simulating turbulent imaging. In *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXII*, 2011. 2

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 4

[45] Endre Repasi and Robert Weiss. Analysis of image distortions by atmospheric turbulence and computer simulation of turbulence effects. In *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XIX*, 2008. 2

[46] Endre Repasi and Robert Weiss. Computer simulation of image degradations by atmospheric turbulence for horizontal views. In *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XXII*, 2011. 2

[47] M. C. Roggemann and B. M. Welsh. *Imaging through Atmospheric Turbulence*. Taylor & Francis, 1996. 2

[48] Arun Ross, Sudipta Banerjee, Cunjian Chen, Anurag Chowdhury, Vahid Mirjalili, Renu Sharma, Thomas Swearingen, and Shivangi Yadav. Some research problems in biometrics: The future beckons. In *ICB*, 2019. 1

[49] Arun Ross and Anil Jain. Information fusion in biometrics. *Pattern recognition letters*, 2003. 3

[50] Arun A Ross, Karthik Nandakumar, and Anil K Jain. *Handbook of multibiometrics*. Springer Science & Business Media, 2006. 1

[51] Seyed Morteza Safdarnejad, Xiaoming Liu, Lalita Udpa, Brooks Andrus, John Wood, and Dean Craven. Sports videos in the wild (svw): A video dataset for sports analysis. In *FG*, 2015. 4

[52] Jason D Schmidt. Numerical simulation of optical wave propagation with examples in MATLAB. *(No Title)*, 2010. 2, 3

[53] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *CVPR*, 2018. 3

[54] V. I. Tatarskii. *Wave Propagation in a Turbulent Medium*. New York: Dover Publications, 1961. 2

[55] Junfeng Wan, Jiangfan Deng, Xiaosong Qiu, and Feng Zhou. Body-face joint detection via embedding and head hook. In *ICCV*, 2021. 4

[56] Qiang Wang, Yun Zheng, Pan Pan, and Yinghui Xu. Multiple object tracking with correlation learning. In *CVPR*, 2021. 3

[57] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 3

[58] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, 2016. 7

[59] Rajeev Yasarla and Vishal M Patel. CNN-based restoration of a single face image degraded by atmospheric turbulence. *TBIOM*, 2022. 3

[60] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *TPAMI*, 2021. 5

[61] Xi Yin, Ying Tai, Yuge Huang, and Xiaoming Liu. Fan: Feature adaptation network for surveillance face recognition and normalization. In *ACCV*, 2020. 2

[62] En Yu, Zhuoling Li, and Shoudong Han. Towards discriminative representation: multi-view trajectory contrastive learning for online multi-object tracking. In *CVPR*, 2022. 3

[63] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, 2022. 3

[64] Ziyuan Zhang, Luan Tran, Feng Liu, and Xiaoming Liu. On learning disentangled representations for gait recognition. *TPAMI*, 2020. 2

[65] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 5

[66] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 1

[67] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3D human reconstruction from a single image. In *ICCV*, 2019. 6

[68] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *TPAMI*, 2017. 4

[69] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. *ECCV*, 2020. 3

[70] Yanjia Zhu, Hongxiang Cai, Shuhan Zhang, Chenhao Wang, and Yichao Xiong. Tinaface: Strong but simple baseline for face detection. *arXiv preprint arXiv:2011.13183*, 2020. 3