

Unconstrained Body Recognition at Altitude and Range: Comparing Four Approaches

Blake A Myers, Matthew Q Hill, Veda Nandan Gandhi, Thomas M Metz, Alice J O'Toole
School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, Texas

Abstract—This study presents an investigation of four distinct approaches to long-term person identification using body shape. Unlike short-term re-identification systems that rely on temporary features (e.g., clothing), we focus on learning persistent body shape characteristics that remain stable over time. We introduce a body identification model based on a Vision Transformer (ViT) (Body Identification from Diverse Datasets, BIDDs) and on a Swin-ViT model (Swin-BIDDs). We also expand on previous approaches [26] based on the Linguistic and Non-linguistic Core ResNet Identity Models (LCRIM and NLCRIM), but with improved training. All models are trained on a large and diverse dataset of over 1.9 million images of approximately 5k identities across 9 databases. Performance was evaluated on standard re-identification benchmark datasets (MARS [42], MSMT17 [34], Outdoor Gait [31], DeepChange [35]) and on an unconstrained dataset [3] that includes images at a distance (from close-range to 1000m), at altitude (from an unmanned aerial vehicle, UAV), and with clothing change. A comparative analysis across these models provides insights into how different backbone architectures and input image sizes impact long-term body identification performance across real-world conditions.

I. INTRODUCTION

Face recognition algorithms are highly accurate at establishing the unique identity of individuals (e.g., [5], [19], [33]). In natural viewing conditions, however, facial identity information is commonly degraded or obscured (e.g., viewing from a far distance or at an extreme angle). When the face is unusable or inaccessible, information about the shape of the body can constrain identity decisions. Body shape can contribute to person identification by supporting/vetoing uncertain face identifications and/or by establishing a plausible identity match to a gallery image. As such, it can serve as a valuable biometric, even if it provides information that does not uniquely identify an individual.

Early attempts to use the body for identification focused on re-identification in a closed-world setting, which aims to track a person in a constrained environment like an airport or train station (for a review, see [39]). In a closed-world environment, algorithms can rely on cues such as clothing that make the problem easily tractable with deep learning. In an open-world, clothing change makes the problem more difficult [1], [7], [14], [16], [22], [23], [26], [36], [38], [26], [23], [16]. As algorithms capable of overcoming clothing change have matured (for a review, see [39]), long-term body identification models have aimed more broadly at identification in highly challenging scenarios [16], [18], [17], [26]. This emphasis coincides with the development of the BRIAR dataset, which contains whole person images, cap-

tured at long-range (e.g., 300+ meters), through atmospheric turbulence, and/or from elevated sensor platforms [3]. Images and videos in this dataset are taken across multiple views (yaw, pitch) and with clothing changes that sideline short-term cues that are useful in re-identification scenarios [40].

Long-term body identification with unconstrained data presents a number of unique challenges. Body shape can deform via the movements of the limbs (e.g., arms up or down, leg extended) and/or by changes in posture (e.g., bending, reaching). Texture and albedo information, which are critically important for face identification, have only limited value for clothed bodies. Perhaps most challenging is the complicating factor of clothing change that alters the overall shape of the body (e.g., pants/skirt, running gear/winter coat). Despite the difficulty of the task, there is ample evidence that humans use body information for person identification when the face is unavailable or insufficiently resolved for identification [8], [29], [41].

The goal of the present study was to compare four machine learning approaches to real-world (longer-term) body identification. We evaluated two Vision Transformer models (ViT) [6], [24] and two ResNet neural networks [10]. All four models were trained to identify bodies from input images, using a very large dataset compiled from 9 feeder datasets. The use of a common dataset for training allowed us to compare the models on an equal footing.

In the first part of the work, we compared the models on four common re-ID datasets. In the second part of the work, we tested the models on the BRIAR test set (BTS) [3]. We also test performance on subsets of data that measure body identification with face included and restricted, at long range, and from overhead. In the third part of the work, we dissect the advantage of the best model to determine whether architecture or image size accounts for its superiority.

The contributions of the paper are:

- We show that ViT models are superior to equivalently-trained ResNet models for body identification.
- We show that a Swin-ViT model is superior to the other tested models across metrics. This was true for the benchmark datasets and the unconstrained dataset.
- Linguistic pre-training in a ResNet model showed only a small performance advantage over an equivalent non-linguistically trained model.
- We show that both architecture and image size contributed to the superior performance of the Swin-ViT, but that image size was the critical factor in its high performance.

A. Related Work

Long-term body identification models can be categorized according to the approach they take to representing body shape information.

1) *2D Body Shape from Images*: The most direct approach is to learn a mapping from variable images of bodies (view, clothing, illumination, distance) to identity [9], [26]. The greatest challenge in this approach has been the limited availability of training data with sufficient variability (especially clothing sets) to learn the task of long-term body identification. Using a ResNet-50 model pretrained on ImageNet [30], the Clothing-Change Feature Augmentation (CCFA) approach [9] augments model training to form meaningful clothing variations in the feature space. The augmented features maximize the change of clothing and minimize the change of identity by adversarial learning. The effectiveness of CCFA was demonstrated with two standard CC-ReID datasets (PRCC-ReID [36] and LTCC-ReID [28]).



Fig. 1. Example body images from the BTS dataset [3]. Subject consented to publication.

The Non-linguistic Core ResNet Identity Model (NL-CRIM) [26] was built on a ResNet-101 backbone pretrained with ImageNet [30]. NL-CRIM was trained to map body images to identities using the BRIAR Research Set (BRS) [3]. It was evaluated with the BRIAR Test Set (BTS), which contained identities viewed at multiple distances (up to 1000 meters) that varied widely in yaw and pitch. Extreme pitch conditions were captured from unmanned aerial vehicles (UAVs). All test items included a change of clothing. (See Figure 1 for image examples). NL-CRIM performed well across all probe distance/pitch conditions. An improved version of this model, with enhanced training and substantially more training data, is tested in the present study.

A similar direct approach to learning a mapping between whole body images and identity was taken in [15]. A ResNet-50 model was trained from scratch with BRS data. This body

encoder was embedded in an end-to-end system that included a trained detector model. The combined model performed well on the unconstrained BTS data.

The causality-based autointervention model (AIM1) was proposed to mitigate clothing bias for robust clothes-changing person ReID (CC-ReID) [37]. Specifically, the effect of clothing on model inference was analyzed. A dual-branch structure of clothing and ID was utilized to simulate the causal intervention process and was penalized by a causality loss. Progressively, clothing bias was automatically eliminated with model training, as AIM learned more discriminative identity clues that are independent of clothing. The superiority of the AIM approach over other approaches was demonstrated with two standard CC-ReID datasets (PRCC-ReID [36] and LTCC-ReID [28]).

2) *3D Body Shape Features*: To overcome reliance on short-term cues in body images, several models have attempted to reconstruct 3D body shapes for identification ([1], [22]). In the 3D Shape Learning (3DSL) approach, a texture-insensitive 3D shape embedding is extracted from a 2D image by adding 3D body reconstruction as an auxiliary task and regularization [1]. The use of the 3D reconstruction regularization forces a decoupling of the 3D body shape from the visual texture, enabling the model to acquire discriminative 3D shape ReID features. An adversarial self-supervised projection (ASSP) model is used to provide a 3D shape ground truth. The effectiveness of the approach was demonstrated with common person ReID datasets (e.g., Market1501 [43]) and clothes-changing datasets (e.g., PRCC-ReID [36] and LTCC-ReID [28]).

In other work, the 3DInvarReID model [22] begins by disentangling identity from non-identity components (pose, clothing shape, and texture) of 3D clothed humans. Next, accurate 3D clothed body shapes are reconstructed, and discriminative features of naked body shapes for person ReID are learned. The model was found to be effective for disentangling identity and non-identity features in 3D clothed body shapes, using a dataset (CCDA [22]) that contains a wide variety of human activities and clothing changes.

3) *Linguistic Models*: Body models based on linguistic descriptors (e.g., “curvy,” “long-legged”) encode shape via the complex myriad of features captured by single and small groups of words [26]. Work in psychology [13] and computer graphics [32] has demonstrated that a linear mapping can be learned from human-generated body descriptions (27 words) to the coefficients of a PCA trained with 3D body scans [25]. Motivated by this finding, the Linguistic Core ResNet Identity Model (LCRIM) was developed using an ImageNet pretrained ResNet augmented with linguistic annotation pretraining. This linguistic core was then trained to map images to identity [26]. Although the LCRIM model performed at a level similar to NL-CRIM, the fusion of the two models performed substantially better than either model alone. This suggests that the two models encode complementary information about body shape.

In related work, linguistic body descriptions were leveraged for ReID in CLIP3DReID [23]. This was done by

integrating human descriptions with visual perception using a pretrained CLIP model. CLIP was used to automatically label body shapes with linguistic descriptors. A student model’s local visual features were then aligned with shape-aware tokens derived from CLIP’s linguistic output. The CLIP image encoder and the 3D SMPL [25] identity spaces were used in combination to align the global visual features. The effectiveness of CLIP3DReID was demonstrated using PRCC-ReID [36] and LTCC-ReID [28].

II. METHODS

A. Model Architectures

We examined four distinct approaches to body shape recognition: two vision transformer models (BIDDS and Swin-BIDDS) and two ResNet-based models (LCRIM and NLCRIM) [26]. Each architecture employs a unique strategy for capturing body shape features across varying conditions.

1) *Vision Transformer Models*: The BIDDS model is built on a Vision Transformer architecture. We used a ViT-B/16 pre-trained on ImageNet-1k. The core model processes 224×224 sized images with patch size 16. We modified the original ViT architecture by replacing the classification head with a custom fully connected layer that maps to a 2048-dimensional embedding. This embedding space is designed to capture essential body shape features crucial for person identification. Following core training, we fine-tune the model on the BRS1–5 datasets (cf. [3]), increasing image size to 384×384 to capture more detailed features of the fine-tuning BRIAR data, while maintaining the same architectural structure.

Swin-BIDDS is based on the hierarchical vision transformer, which uses shifted windows (Swin Transformer, [24]). This type of transformer was developed to better adapt transformers from the language domain to the vision domain, by accommodating large variations in the scale of visual entities. The shifted windowing scheme of the Swin Transformer is more efficient than a standard ViT, because it limits self-attention computation to non-overlapping local windows, while supporting cross-window connections. This hierarchical structure progressively merges patches and is well-suited to modeling at various scales.

2) *ResNet-Based Models*: These models leverage the ResNet architecture with different core training strategies. Both are pre-trained with ImageNet-1k [30]. Additionally, LCRIM incorporates semantic body descriptors into its training process (See Section II-C.1 for details). Its architecture consists of a ResNet-50 base augmented with an encoder-decoder structure that maps to a linguistic feature space before the final identification layers. The encoder pathway compresses the representation ($2048 \rightarrow 512 \rightarrow 64 \rightarrow 16$), while the decoder pathway ($16 \rightarrow 24 \rightarrow 30$) reconstructs linguistic body attributes. NLCRIM is identical to LCRIM, but without linguistic training.

By comparison to the published version of NLCRIM and LCRIM [26], there were three changes: 1.) a new training regime with hard triplet mining was added [12]; 2.) there was a substantial increase in the quantity of training data;

and 3.) the ResNet-101 was replaced by a ResNet-50 (see below for details).

B. Training Methods

All models employed hard triplet loss with negative mining [12]. This operates on image triplets: an anchor image, a positive sample (same identity), and a negative sample (different identity). The loss calculation measures the Euclidean distances between the anchor and positive samples and between the anchor and negative samples. We selected the most challenging negative samples (i.e., those closest to the anchor in the embedding space) within each batch. This hard negative mining encourages the model to learn features that effectively differentiate between similar body shapes. We also ensured that each batch included pairs or small sets of images from the same person. All four models use the Adam optimizer and incorporate dynamic sampling, whereby triplet selection is adapted based on the current state of the embeddings. This ensures that the models continuously encounter challenging examples throughout training. The training process employs a low learning rate (10^{-5}) and weight decay (10^{-6}) to prevent over-fitting while maintaining stability. We applied standard augmentations during training, including random horizontal flip, color jitter, random grayscale, and gaussian blur.

TABLE I
TRAINING AND TESTING DATASETS

Dataset	Images	IDs	Clothes Change
UAV-Human [21]	41,290	119	no
MSMT17 [34]	29,204	930	no
Market1501 [43]	17,874	1,170	no
MARS [42]	509,914	625	no
STR-BRC 1	156,688	224	yes
P-DESTRE [20]	214,950	124	no
PRCC [36]	17,896	150	yes
DeepChange [35]	28,1731	451	yes
BRS 1–5 [3]	697,348	995	yes
Total Training	1,966,895	4,788	
MSMT17 [34]	82,510	2,697	no
MARS [42]	509,966	634	no
Outdoor Gait [31]	402,009	138	no
DeepChange [35]	103,324	671	yes
Total Testing	1,097,809	4,140	

C. Training Data

An important feature of our approach is the use of a large and diverse collection of training datasets (see Table I). The datasets include over 1.9 million images across 4,788 identities. To benchmark the models for the experiments (see Section III), a subset of test data were withheld from three of the training sets (MSMT17 [34], MARS [42], and DeepChange [35]), and a fourth set was designated solely for testing purposes (Outdoor Gait [31]), with none of its data included in the training phase. This diverse collection spans multiple scenarios, from ground-level views to aerial perspectives. The training images were primarily derived from video files, with bodies cropped and processed to

maintain their aspect ratios while being placed on a 224×224 (384×384 for Swin-BIDS) black background. Some of the datasets include clothing change and some do not (see Table I).

1) *Additional Pre-training*: Only LCRIM had additional pre-training. This comprised a specialized linguistic pre-training phase using the HumanID [27] and MEVA [4] datasets. The HumanID dataset provides diverse viewing scenarios of 297 identities, including approach sequences, walking perpendicular to the camera, and elevated viewpoints. Each identity was annotated by 20 human observers using 30 standardized body descriptors, with the final descriptors averaged across annotators (cf., [26]). The MEVA dataset, comprising over 9,300 hours of video across varied activities and scenarios, contributed an additional 158 identities. Images from these datasets were used to train LCRIM’s initial ability to map between visual features and linguistic body descriptions. The model was subsequently tuned for mapping image to identity using the datasets listed in Table I.

2) *Additional Fine-tuning*: Subsequent to the large-scale training using the datasets in Table I, both the BIDS and Swin-BIDS were fine-tuned using the BRIAR training data (BRS1–5), which contained 697,348 images of 995 unique identities. Note: this training data was included in the large scale training and repeated in the fine-tune stage. During this fine-tuning, BIDS processes images at an increased size of 384×384 , allowing for more detailed feature extraction. The Swin-BIDS model used the 384×384 images for both the large-scale training and the fine-tuning.

In summary, the strategy across models combines specialized linguistic pre-training, extensive foundation-model training, and targeted fine-tuning to fully exploit the capabilities of each architectural approach. The processing of video-derived images and standardization of input sizes ensures consistent training conditions across the models.

III. EXPERIMENTS

A. Benchmark Dataset Tests

1) *Methods*: The models were evaluated first with the test data from four benchmark Re-ID datasets (MSMT17 [34], MARS [42], Outdoor Gait [31], and DeepChange [35]) (See Table I). DeepChange is a clothes change database; MSMT17, MARS, and Outdoor Gait are not. For MSMT17, MARS, and Outdoor Gait the test data were split into a gallery (half of the items) and a probe set (remaining items). Because the data are derived from video, the split was made to assure minimally similar gallery and probe items. Images were processed by the models, and identity templates were formed for gallery items by averaging embeddings of the images for each identity. Identification was measured by comparing probe image embeddings to the gallery templates.

For DeepChange, all identities had multiple clothing sets. The original DeepChange dataset, however, used similar clothing for each identity across both the probe and gallery sets. Thus, to ensure that clothing did not become a dominant cue, we restructured the partitioning of the probe and gallery sets. Specifically, we designated a single clothing set for all

probe instances and then ensured that each identity’s gallery templates had different clothing.

2) *Results*: Table II summarizes the verification (TAR@FAR) and identification (Rank) performance for the benchmark datasets. The Swin-BIDS model performed best on all metrics and for all datasets. At a general level, the transformer-based models (BIDS and Swin-BIDS) performed more accurately than the ResNet-based models (NLCRIM and LCRIM). This was consistent across all metrics and for all datasets. Although comparisons with benchmarks in the literature are not always possible (or transparent), the rank 1 performance of NLCRIM, BIDS and Swin-BIDS exceeded the state-of-the-art (SOTA) for MARS (cf. previous Rank 1 SOTA: MARS (.908) [11]). The Swin-BIDS exceeded the SOTA for MSMT17 (cf. previous Rank 1 SOTA: MSMT (.917) [2]) and DeepChange (previous Rank 1 SOTA (.48) [35]).

It is worth noting that our models were trained on multiple datasets in which clothing was not a reliable cue to identity. The strong performance of BIDS and Swin-BIDS on the no-clothes-change datasets indicates that the models utilize body shape cues, and other identifying information not linked to clothing (head structure). The strong performance of the Swin-BIDS model on DeepChange (clothes-change) is consistent with this conclusion.

B. Identification in Unconstrained Datasets

1) *Methods*: Next, the models were evaluated using the most challenging of the datasets. Specifically, we used the BRIAR Test Set (BTS) summarized in Table III. The first test was conducted on the entire dataset and subsequent tests were done on targeted partitions of the data into probe items. These partitions included: a.) face-included items; b.) face-restricted items, c.) long-range items taken at distance, and d.) items captured from overhead using an UAV.

To test identification, gallery embedding templates were formed by averaging the embeddings across all still images for each identity. Probe embedding templates were derived from video segments, indicating the specific frames to be used from the video. The embedding for each probe was computed by averaging the frame-level embeddings across this subset of frames. As a result, the videos for a given identity each contributed multiple probe embeddings (one per segmented clip).

2) *Results*: Table IV shows that Swin-BIDS performed substantially better than the other models on nearly all metrics. The table also shows the consistency of this advantage across the test partitions. As for the benchmark datasets, the ViT-based models (BIDS and Swin-BIDS) were clearly superior to the ResNet models. Although it is difficult to compare across partitions, especially given the different numbers of items in each set, Rank 1 and Rank 20 performance suggest that Swin-BIDS provides consistently strong identity information for probes with and without a visible face, probes at a distance, and probes taken from overhead (UAV). Moreover, despite differences in the overall performance of the four models, none collapsed on

TABLE II
TEST SET PERFORMANCE.

Dataset	Model	AUC	TAR@FAR 10^{-3}	TAR@FAR 10^{-4}	Rank 1	Rank 20
MARS	NLCRIM	0.9942	0.8059	0.4621	0.8868	0.9774
	LCRIM	0.9962	0.9235	0.7468	0.9147	0.9785
	BIDS	0.9958	0.9509	0.8562	0.9476	0.9824
	Swin-BIDS	0.9984	0.9698	0.8803	0.9666	0.9906
MSMT 17	NLCRIM	0.9909	0.5979	0.3114	0.5428	0.8604
	LCRIM	0.9909	0.6306	0.3609	0.5256	0.8510
	BIDS	0.9955	0.8426	0.6466	0.7640	0.9406
	Swin-BIDS	0.9993	0.9750	0.9082	0.9445	0.9897
Outdoor Gait	NLCRIM	0.9840	0.3671	0.1513	0.8435	0.9889
	LCRIM	0.9816	0.4021	0.1550	0.8399	0.9835
	BIDS	0.9964	0.8295	0.4672	0.9465	0.9951
	Swin-BIDS	0.9992	0.9657	0.7003	0.9802	0.9978
DeepChange	NLCRIM	0.8709	0.0720	0.0176	0.1539	0.4786
	LCRIM	0.8551	0.0902	0.0276	0.1337	0.4444
	BIDS	0.8869	0.2021	0.0897	0.2276	0.5504
	Swin-BIDS	0.9861	0.4227	0.1527	0.5028	0.8836

TABLE III
BTS DATASET AND PARTITIONS

Dataset Partition	IDs	Media Files
Test Sets		
Gallery	858	164,638
Probe	367	9,215
Probe Subsets		
Face Included	367	5,749
Face Restricted	367	1,893
Long-range Body	362	2,832
Unmanned aerial vehicle	139	834

the partition tests, highlighting the diversity and quantity of the training data in the success of the models.

C. Ablation Experiments: Architecture vs. Input Size?

The Swin-BIDS model performed best on the benchmark datasets and the challenging BRIAR data. It was also consistently best on all partitions of the BRIAR data. In these experiments, we test factors that might account for the superior performance of Swin-BIDS over its closest competitor, BIDS. The models differed in two ways. The backbone architecture changed from a ViT model (BIDS) to a Swin-ViT (Swin-BIDS) and the input image size for the core model training changed from 224×224 sized images (BIDS) to 384×384 sized images (Swin-BIDS). Both models were fine-tuned with 384×384 images.

Technically, changes in architecture and image size are independent. However, a simultaneous change in both is not uncommon, due to the fact that the Swin-ViT scales far more efficiently than ViT with increasing input image size. It does this by implementing a hierarchical structure of shifting attention windows, giving Swin-ViT the desirable feature of linear complexity as a function of image size [24]. Thus, a change from a ViT to Swin-ViT is often undertaken with the goal of minimizing the computational resources required for an increase in image size.

1) *Methods*: To tease apart whether architecture, image size, or both were responsible for the performance boost, we trained additional models as comparators. To test directly

for the effects of image size independent of architecture, we compared two Swin-ViT models. The first was the Swin-BIDS model tested in the previous experiments. This is a Swin-ViT model that used image size 384×384 for core training and fine-tuning. For clarity, we refer to this as Swin-BIDS($384,384$). For the comparison model, we trained Swin-BIDS($224,224$). This is a Swin-ViT model that used 224×224 images for core training and fine-tuning.

To determine the impact of the image size independent of architecture we compared a ViT model with a Swin-ViT model, keeping image size constant. Specifically, the Swin-BIDS($224,224$) model was compared with a BIDS($224,224$). This latter is a version of the BIDS($224,384$) model used in previous experiments, but with the image size for fine-tuning lowered to 224.

2) *Results*: Plots showing the CMC and ROC of these models appear in Figure 2. Both plots indicate that the performance of the Swin-BIDS($384,384$) surpasses the other models primarily based on its processing of the larger image size. There is also a smaller contribution of the Swin-ViT architecture, which is seen more clearly in Table V. Metrics for each model appear in the top rows. At the bottom of the table, the first row shows the effects of changing architecture from the ViT to Swin-ViT. This is computed as the difference between the two architecture comparator networks. All 4 metrics increase with that change. The final row of Table V shows the effects of changing image size from the smaller ($224, 224$) size to larger ($384, 384$) size. This is computed as the difference between the two image size comparator networks. Again, all 4 metrics increase, but by a substantially larger margin.

As a final check on the consistency of image size as the critical factor in the superior performance of Swin-BIDS over BIDS, we conducted the same comparisons on the benchmark datasets. The results appear in Figure 3 and show again that image size is the driving factor in the superior performance of Swin-BIDS over BIDS. In these benchmark datasets, changing the backbone architecture had variable effects for different datasets.

TABLE IV
BRIAR TEST SET PERFORMANCE.

Probe Set	Model	AUC	TAR@FAR 10^{-3}	TAR@FAR 10^{-4}	Rank 1	Rank 20
All Probes	NLCRIM	0.9403	0.0698	0.0167	0.1467	0.5825
	LCRIM	0.9275	0.0740	0.0188	0.1615	0.5731
	BIDDS	0.9745	0.2926	0.1207	0.3342	0.7816
	Swin-BIDDS	0.9802	0.3575	0.1523	0.3909	0.8228
Face Included	NLCRIM	0.9394	0.0658	0.0183	0.1372	0.569
	LCRIM	0.9258	0.0699	0.0184	0.1571	0.5655
	BIDDS	0.9736	0.2767	0.1089	0.3248	0.7779
	Swin-BIDDS	0.9797	0.3451	0.1388	0.3799	0.8201
Face Restricted	NLCRIM	0.9326	0.0666	0.0116	0.1337	0.5473
	LCRIM	0.9173	0.0666	0.0127	0.1432	0.5346
	BIDDS	0.969	0.2573	0.1051	0.2948	0.7517
	Swin-BIDDS	0.9745	0.3133	0.1352	0.3391	0.7855
Long-range Body	NLCRIM	0.9323	0.0639	0.018	0.1063	0.5035
	LCRIM	0.9123	0.0749	0.018	0.119	0.4905
	BIDDS	0.9625	0.2331	0.0964	0.2549	0.7023
	Swin-BIDDS	0.9685	0.2669	0.1095	0.2832	0.7376
UAV	NLCRIM	0.9295	0.0779	0.0108	0.1367	0.5468
	LCRIM	0.9253	0.1031	0.0228	0.1823	0.5408
	BIDDS	0.9779	0.3549	0.1835	0.3573	0.7926
	Swin-BIDDS	0.9777	0.3765	0.2338	0.4053	0.801

TABLE V
ABLATION RESULTS ON BRIAR TEST SET, ALL PROBES.

Architecture	Core	Fine-Tune	TAR@FAR 10^{-3}	TAR@FAR 10^{-4}	Rank 1	Rank 20
BIDDS	224	224	0.2549	0.0990	0.3072	0.7645
Swin-BIDDS	224	224	0.2776	0.1063	0.3141	0.7865
BIDDS	224	384	0.2926	0.1207	0.3342	0.7816
Swin-BIDDS	384	384	0.3575	0.1523	0.3909	0.8228
Architecture: BIDDS to Swin-BIDDS			+0.0227	+0.0073	+0.0069	+0.0220
Image size: 224 to 384			+0.0799	+0.0460	+0.0768	+0.0363

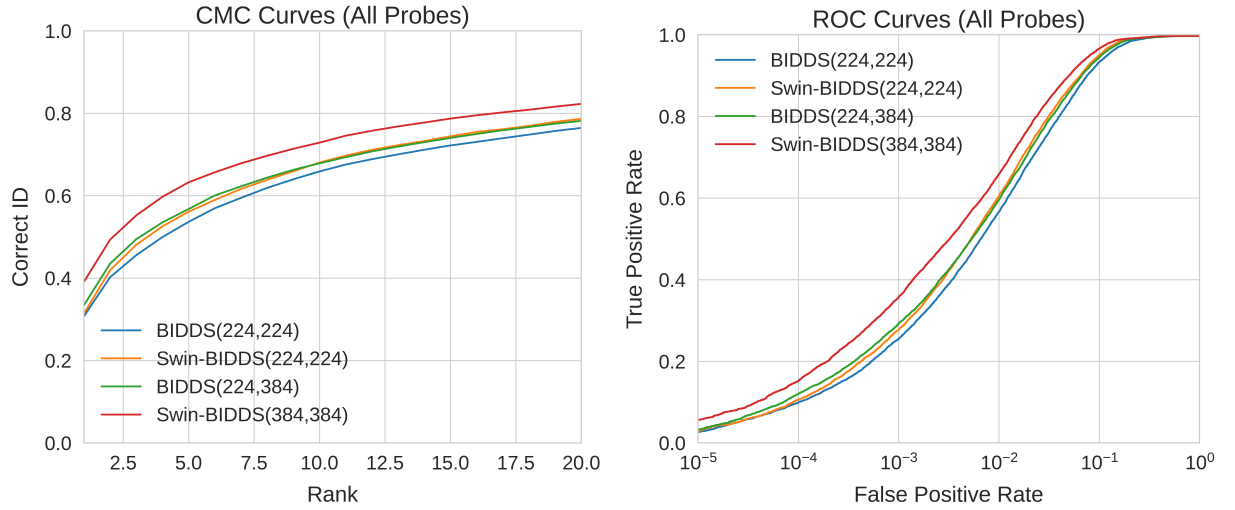


Fig. 2. Ablation: CMC and ROC curves for the architecture and image size comparisons show that image size is the critical factor in the superior performance of the Swin-BIDDS model over the BIDDS model.

IV. CONCLUSIONS AND FUTURE WORKS

We implement four long-term body identification models based on ResNet (LCRIM, NLCRIM) and ViT (BIDDS, Swin-BIDDS) architectures. The models are tested on their ability to identify bodies in benchmark re-identification datasets and in a highly challenging unconstrained dataset

that includes people viewed at a distance, from elevated vantage points, and with clothing variability. An important aspect of our approach is that we train the models on a large-scale, diverse dataset of nearly two million images of nearly 5,000 identities. We showed that vision transformer architectures consistently outperformed ResNet architectures.

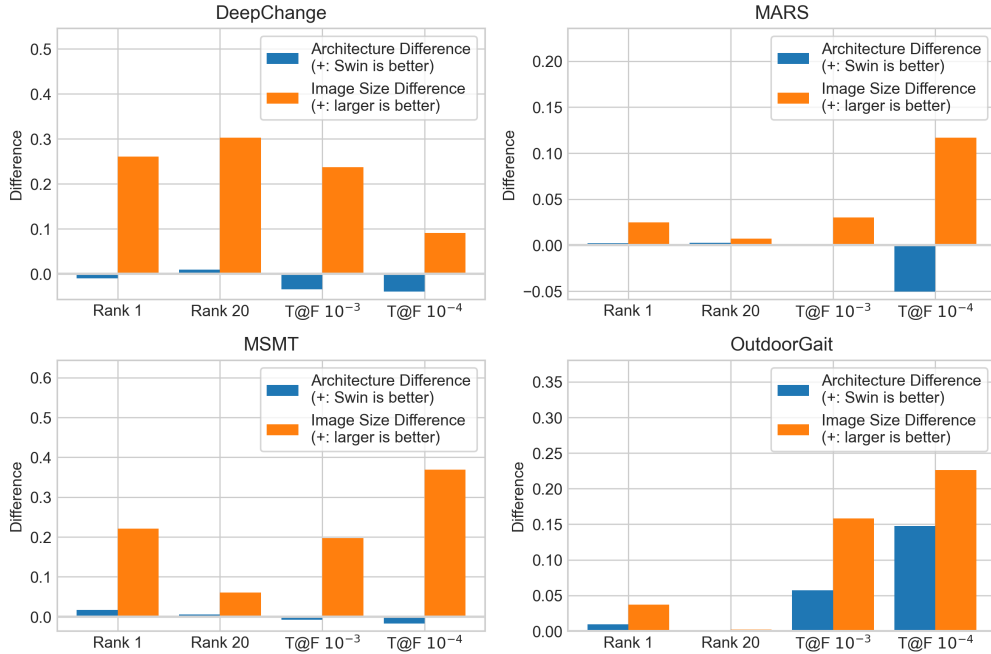


Fig. 3. Difference in performance by model architecture (ViT, Swin-ViT) and input image size (224 px², 384 px²). Comparisons shown for each of four datasets (DeepChange, MARS, MSMT, and Outdoor Gait) on four identification metrics (retrieval at ranks 1 and 20, true accept rate at false accept rates 10⁻³ and 10⁻⁴). Architecture Difference (blue) is defined as the difference between Swin-BIDDS(224,224) and BIDDS(224,224). Image Size Difference (orange) is defined as the difference between Swin-BIDDS(384,384) and Swin-BIDDS(224,224).

Swin-BIDDS was the most accurate model across metrics and models. Its primary advantage over the BIDDS model was its use of a larger input image size. The results underscore the importance of leveraging large image size and hierarchical self-attention in capturing subtle body shape differences. Overall, the Swin-BIDDS model demonstrates strong generalization across benchmark datasets and the BRIAR set, making it a highly robust approach for long-term, real-world body identification.

Promising avenues for future work include the use of larger and more diverse training sets, and new learning paradigms that leverage unlabeled data in complex viewing conditions. Moreover, unsupervised or semi-supervised approaches that exploit partially labeled videos could reduce reliance on fully annotated datasets, potentially handling rare body shapes and challenging occlusions more robustly. Integrating such techniques directly into the current architectures or using them in conjunction with shape-based encoders may open the door to even more accurate and resilient body-identification systems under real-world constraints.

V. 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 [2022-21102100005]. 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.

ETHICAL IMPACT STATEMENT

We have read the guidelines for the Ethical Impact Statement. The development of body identification models does not involve direct contact with human subjects, and therefore does not require approval by an Institutional Review Board. Instead, images/videos of human subjects are incorporated as training and test data for body identification models. We used only datasets (videos and images of people) that have been pre-screened and approved for ethical data collection standards by a United States government funding agency, XXXX. The standards applied for dataset approval require consent from the subjects who are depicted in the images/videos for use in research. Specifically, the standards are set in accordance with Health Services Research and applicable privacy policies, statutes, and federal regulations. Images/videos of subjects who appear in publications require additional consent. We followed these guidelines carefully. Images displayed in the paper have been properly consented and are displayed according to the published instructions for use of the dataset.

The development and study of biometric identification algorithms entails risk to individuals and societies. It is clear that these systems can have negative impacts if they are misused. They can potentially threaten individual privacy and can impinge on freedom of movement and expression in a society. The goal of our work is to better understand how these systems work. The results of this work can have both positive

and negative societal impacts. On the positive side, knowing the types of representations created by body identification networks can help to minimize person identification errors. It can also help to set reasonable performance expectations thereby limiting the scope of use. On the negative side, the knowledge gained can potentially be used to manipulate a system in unethical ways and to create synthetic images that can be misused or misinterpreted.

These risks are mitigated by the potential for positive societal impact. Body identification algorithms can be used to locate missing people (including children). They can also be used in law enforcement to identify individuals implicated in crimes. Legitimate and societally-approved use can protect the general public from harm. Of note, body identification systems can be used in combination with face identification systems to improve identification accuracy, thereby minimizing erroneous identifications.

REFERENCES

- [1] J. Chen, X. Jiang, F. Wang, J. Zhang, F. Zheng, X. Sun, and W.-S. Zheng. Learning 3d shape feature for texture-insensitive person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8146–8155, 2021.
- [2] W. Chen, X. Xu, J. Jia, H. Luo, Y. Wang, F. Wang, R. Jin, and X. Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15050–15061, 2023.
- [3] D. Cornett, J. Brogan, N. Barber, D. Aykac, S. Baird, N. Burchfield, C. Dukes, A. Duncan, R. Ferrell, J. Goddard, et al. Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 593–602, 2023.
- [4] K. Corona, K. Osterdahl, R. Collins, and A. Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1060–1068, January 2021.
- [5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.
- [6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [7] X. Gu, H. Chang, B. Ma, S. Bai, S. Shan, and X. Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1060–1069, 2022.
- [8] C. A. Hahn, A. J. O’Toole, and P. J. Phillips. Dissecting the time course of person recognition in natural viewing environments. *British Journal of Psychology*, 107(1):117–134, 2016.
- [9] K. Han, S. Gong, Y. Huang, L. Wang, and T. Tan. Clothing-change feature augmentation for person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22066–22075, 2023.
- [10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.
- [11] T. He, X. Jin, X. Shen, J. Huang, Z. Chen, and X.-S. Hua. Dense interaction learning for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1490–1501, 2021.
- [12] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.
- [13] M. Q. Hill, S. Streuber, C. A. Hahn, M. J. Black, and A. J. O’Toole. Creating body shapes from verbal descriptions by linking similarity spaces. *Psychological science*, 27(11):1486–1497, 2016.
- [14] P. Hong, T. Wu, A. Wu, X. Han, and W.-S. Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10513–10522, 2021.
- [15] S. Huang, R. P. Kathirvel, Y. Guo, C. P. Lau, and R. Chellappa. Whole-body detection, identification and recognition at altitude and range. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.
- [16] S. Huang, R. P. Kathirvel, C. P. Lau, and R. Chellappa. Whole-body detection, recognition and identification at altitude and range. *arXiv preprint arXiv:2311.05725*, 2023.
- [17] S. Huang, R. Prabhakar, Y. Guo, R. Chellappa, and C. Peng. Vills: Video-image learning to learn semantics for person re-identification, 2024.
- [18] S. Huang, Y. Zhou, R. P. Kathirvel, R. Chellappa, and C. P. Lau. Self-supervised learning of whole and component-based semantic representations for person re-identification, 2023.
- [19] 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.
- [20] S. A. Kumar, E. Yaghoubi, A. Das, B. Harish, and H. Proença. The p-destre: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. *IEEE Transactions on Information Forensics and Security*, 16:1696–1708, 2020.
- [21] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16261–16270, 2021.
- [22] F. Liu, M. Kim, Z. Gu, A. Jain, and X. Liu. Learning clothing and pose invariant 3d shape representation for long-term person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19617–19626, 2023.
- [23] F. Liu, M. Kim, Z. Ren, and X. Liu. Distilling clip with dual guidance for learning discriminative human body shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 256–266, 2024.
- [24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [25] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.
- [26] B. A. Myers, L. Jaggernauth, T. M. Metz, M. Q. Hill, V. N. Gandhi, C. D. Castillo, and A. J. O’Toole. Recognizing people by body shape using deep networks of images and words. *Proceedings of the IEEE: International Joint Conference on Biometrics*, 2023.
- [27] A. J. O’Toole, J. Harms, S. L. Snow, D. R. Hurst, M. R. Pappas, J. H. Ayyad, and H. Abdi. A video database of moving faces and people. *IEEE Transactions on pattern analysis and machine intelligence*, 27(5):812–816, 2005.
- [28] X. Qian, W. Wang, L. Zhang, F. Zhu, Y. Fu, T. Xiang, Y.-G. Jiang, and X. Xue. Long-term cloth-changing person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 2020.
- [29] A. Rice, P. J. Phillips, and A. O’Toole. The role of the face and body in unfamiliar person identification. *Applied Cognitive Psychology*, 27(6):761–768, 2013.
- [30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.
- [31] A. Song, L. Linjie, C. Atalla, and G. Gottrell. Learning to see people like people: Predicting social impressions of faces. *Cognitive Science*, 2017.
- [32] S. Streuber, M. A. Quiros-Ramirez, M. Q. Hill, C. A. Hahn, S. Zuffi, A. O’Toole, and M. J. Black. Body talk: Crowdshaping realistic 3d avatars with words. *ACM Transactions on Graphics (TOG)*, 35(4):1–14, 2016.
- [33] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.
- [34] L. Wei, S. Zhang, W. Gao, and Q. Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018.
- [35] P. Xu and X. Zhu. Deepchange: A long-term person re-identification benchmark with clothes change. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11196–11205, 2023.
- [36] Q. Yang, A. Wu, and W.-S. Zheng. Person re-identification by contour

- sketch under moderate clothing change. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2029–2046, 2019.
- [37] Z. Yang, M. Lin, X. Zhong, Y. Wu, and Z. Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1472–1481, 2023.
 - [38] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.
 - [39] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022.
 - [40] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.
 - [41] G. Yovel and A. J. O’Toole. Recognizing people in motion. *Trends in cognitive sciences*, 20(5):383–395, 2016.
 - [42] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VI 14*, pages 868–884. Springer, 2016.
 - [43] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.