SapiensID: Foundation for Human Recognition

Minchul Kim[1], Dingqiang Ye[1], Yiyang Su[1], Feng Liu[2], Xiaoming Liu[1]
[1] Department of Computer Science and Engineering, Michigan State University
[2] Department of Computer Science, Drexel University
[1]{kimminc2, yedingqi, suyiyan1, liuxm}@msu.edu, [2]fl397@drexel.edu

Figure 1. SapiensID is a human recognition model trained on a large-scale dataset of human images featuring varied poses and visible body parts. For the first time, a *single* model performs effectively across diverse face and body benchmarks [25, 56, 71, 85]. This marks a significant improvement over previous body recognition models, which were often limited to one specific camera setup or image alignments for one model, with worse performance in in-the-wild scenarios. Additionally, we introduce a large-scale, cross-pose and cross-scale training and evaluation set designed to facilitate further research in this area. — The name SapiensID pertains to the ability to recognize humans.

Abstract

Existing human recognition systems often rely on separate, specialized models for face and body analysis, limiting their effectiveness in real-world scenarios where pose, visibility, and context vary widely. This paper introduces SapiensID, a unified model that bridges this gap, achieving robust performance across diverse settings. SapiensID introduces (i) Retina Patch (RP), a dynamic patch generation scheme that adapts to subject scale and ensures consistent tokenization of regions of interest, (ii) a masked recognition model (MRM) that learns from variable token length, and (iii) Semantic Attention Head (SAH), an module that learns pose-invariant representations by pooling features around key body parts. To facilitate training, we introduce WebBody4M, a large-scale dataset capturing diverse poses and scale variations. Extensive experiments demonstrate that SapiensID achieves state-of-the-art results on various body ReID benchmarks, outperforming specialized models in both short-term and long-term scenarios while remaining competitive with dedicated face recognition systems. Furthermore, SapiensID establishes a strong baseline for the newly introduced challenge of Cross Pose-Scale ReID, demonstrating its ability to generalize to complex, real-world conditions. Project Link

1. Introduction

Human recognition has traditionally been approached through domain-specific models focused exclusively on either face [13, 28, 29, 34–36, 38, 47, 63, 64, 68, 76] or body [20, 30, 42, 44, 46, 71] recognition (or ReID). Each of these modalities relies heavily on specific dataset alignments, where face recognition models are optimized for tightly cropped, aligned facial images [1, 14, 21, 86], and body recognition models are designed to process full-body images of standing individuals [56, 67, 71, 82].

Despite advances in face and body recognition, no single model has yet effectively managed to handle a diverse range of poses and visible area simultaneously. However, in real-world settings, human recognition often requires harnessing the full spectrum of available clues, integrating both face and body information. Typically, multiple models are fused at the feature or score level [23, 45] to mitigate this issue. In other words, no single model can handle both face and body images as robustly as modality-specific models. Therefore, a unified model would mark a significant advance in human recognition, allowing reliable identification across varied poses and scales of body parts. As in Fig. 2, current models relies heavily on in-domain datasets, fail to generalize effectively to other datasets.

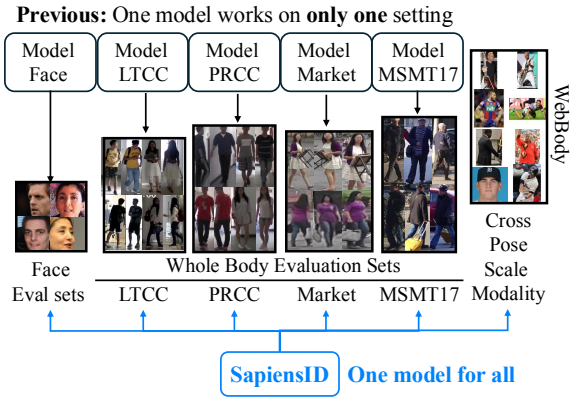

Figure 2. Conventionally, face and body recognition were handled independently. Also body models are trained on one specific dataset without the ability to generalize to other datasets. SapiensID model for the first time generalizes across modalities and different body poses and camera settings.

Addressing this gap is important for several reasons. In real-world applications, human recognition systems should operate across a variety of poses (sitting vs standing) and visible contextual areas (upper torso vs whole body) [73]. For instance, IJB-S [32] contains face gallery images and whole body probe videos. Furthermore, a unified model simplifies model deployment and usage for downstream tasks by eliminating the need for preprocessing steps such as face alignment [14] or dependency on camera setups [56, 71].

However, addressing this problem is not trivial. First, it requires a large-scale labeled human image dataset that captures a wide range of poses and visibility variations. Secondly, even with such a dataset, the model must be capable of managing the substantial variability in scale and pose that human images naturally show. As in Fig. 1, close-up portraits show a large face, while full-body shots display it much smaller. Modality-specific models have eliminated the scale inconsistency problem with some form of pre-alignment stage. For instance, body recognition models assume consistent camera setup [56, 71] and face recognition models assume the images are aligned with 5 facial landmarks to a canonical position [14, 81]. Such transformations of input reduce irrelevant variability in recognizing a person, making training easier. However, models fail to generalize when the preprocessing step fails [37].

To this end, we propose **SapiensID**, one model capable of handling the complexities of human recognition in diverse settings. Our contributions are

- **Model Innovations:** We introduce three major improvements over conventional specialized recognition models:
  1. **Retina Patch** addresses scale variations often encountered in human images by dynamically allocating more patches to important regions.
  2. **Masked Recognition Model** reduces the number of tokens, achieving $8\times$ speed up in ViT during training.
  3. **Semantic Attention Head** addresses pose variations by learning to pool features around keypoints.
- **Data Contribution:** To aid the development and evaluation of **SapiensID**, we release **WebBody4M** (Fig. 1), a large-scale dataset specifically designed for comprehensive human recognition across different poses and scales.
- **Performance:** SapiensID achieves state-of-the-art results across multiple ReID benchmarks and establishes a strong baseline for the novel Cross Pose-Scale ReID task.

Our approach is a paradigm shift human recognition, laying the groundwork for research that bridges the gap between specialized models and holistic recognition systems.

2. Related Works

2.1. Face Recognition

Face Recognition (FR) matches query images to an enrolled identity database. State-of-the-art (SoTA) FR models are trained on large-scale datasets [13, 21, 86] with margin-based softmax losses [13, 29, 34, 47, 64]. FR performance is evaluated on a set of benchmarks, *e.g.* LFW [25], CFP-FP [55], CPLFW [84], AgeDB [52], CALFW [85], and IJB-B,C [50, 69]. They are designed to assess the model's robustness to factors such as pose variations and age differences. Models trained on large datasets, *e.g.* WebFace260M, achieve over 97% verification accuracy on these benchmarks [34]. FR in low-quality imagery is substantially harder and TinyFace [11] and IJB-S [32] are popular benchmarks.

Face recognition is often accompanied by facial landmark prediction [6, 39, 60, 81] so that input faces are aligned and tightly cropped around the facial region. However, when alignment fails, FR models perform poorly [37]. Eliminating alignment would not only simplify the pipeline but also enhance robustness in conditions where alignments are prone to fail. We propose an *alignment-free* paradigm capable of handling any human image with or without a visible face.

2.2. Body Recognition

Body recognition, *a.k.a.* Person Re-identification (ReID), seeks to identify individuals across different times, locations, or camera settings. Prior works [18, 19, 40, 41, 43, 65, 77, 80, 82] focus on short-term scenarios where subjects generally end up with the same attire. Removing this assumption has led to long-term, cloth-changing ReID [8, 20, 24, 30, 42, 58, 62, 71, 78], on datasets like PRCC [71], LTCC [56], CCDA [44] and CelebReID [26, 27].

All of these datasets are composed primarily of whole-body images, where the subjects are fully visible from head to toe, with poses generally limited to walking or standing. While this format has been valuable in the development of person ReID models for controlled environments, it lacks the scale and visibility variety often encountered in real-world applications. To address these limitations, we propose a

model capable of handling diverse and complex poses and visible areas. Further, to facilitate the training and evaluation of these models, we introduce a new large-scale, labeled dataset that significantly broadens pose-scale diversity.

2.3. Patch Generation for Vision Transformers

In Vision Transformer (ViT) [15], an image is divided into patches, with each transformed into a token via linear projection. This patch-based approach transforms images to an unordered set of tokens for sequence-to-sequence modeling [61], processing images in a scalable and flexible way in downstream tasks. Typically, patches are created by dividing an image into a grid with a specific number of patches.

Several works explore how the patchifying process helps ViT capture multi-scale objects in images [66]. For instance, [12] predefines patch counts without resizing the input, retaining the image's aspect ratio and scale. [5] randomizes patch sizes in training for generalization across image scales, enhancing efficiency while sometimes reducing accuracy. Importantly, the representation quality of specific regions, such as face or hand, depends on **the number of tokens** allocated to those areas. A smaller face within a constant patch size, for example, generates fewer tokens and thus captures less detail than a larger face. To address this, we propose to maintain a consistent number of tokens for regions of interest while ensuring full, non-overlapping coverage across the image in line with grid-based tokenization principles.

3. Proposed Method

A human recognition model is formulated as a metric learning task such that images of the same subject are closer in feature space than those of different subjects, satisfying

$$d(\mathbf{f}_A^i, \mathbf{f}_A^j) < d(\mathbf{f}_A^i, \mathbf{f}_B^k), \tag{1}$$

where $\mathbf{f}_A^i$ and $\mathbf{f}_A^j$ denote the feature vectors of two different images i and j of the same subject A, while $\mathbf{f}_B^k$ represents the feature vector of an image of a different subject B. Notably, the subjects A and B are not observed during training. Following established research on margin-based techniques for enhancing intra-class compactness in the feature space [13, 34, 47, 51, 64], we utilize a margin-based softmax loss [34] to train our model on a labeled dataset. We collect a large-scale web-collected human image training dataset which will be discussed in Sec. 3.4.

The key challenge that sets this apart from prior work on a separate face [13, 47] or body [42, 71] recognition task is that the input image can be highly varying in 1) scale and 2) body pose. To tackle these challenges, we propose a new architecture, which will be discussed in the subsections.

3.1. Retina Patch (RP)

To address the issue of varying scale in human images, we propose a novel **Retina Patch** mechanism inspired by the

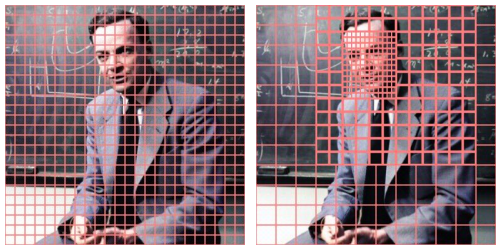

Figure 3. Comparison between the standard grid patch scheme of Vision Transformers (ViT) and our Retina Patch. While maintaining the same or lower computational budget (number of tokens), Retina Patch dynamically allocates more patches to critical regions (e.g., face and upper torso) in an image. This allocation enhances the model's ability to capture fine-grained details in important regions, and to handle varying scales more effectively than fixed grid patch.

human eye's ability to adapt focus dynamically to regions of interest (ROIs) within a scene. In natural images, subjects can appear in diverse poses and with varying visibility of the face and body, leading to substantial differences in scale across regions. For instance, in a full-body image, a face may be a small portion, whereas in a close-up, it dominates. To account for these variations, our Retina Patch dynamically assigns more patches to critical regions within the image.

Assume we have an input image i and a set of image-dependent regions of interest, $\{\text{ROI}_r^i \mid r = 0, 1, \ldots, R\}$, each defined by a bounding box. There are R ROIs per image. Details on how ROIs are computed will be discussed later. We also let ROI_0^i be the whole image. For each ROI_r^i, we set a specific number of patches m_r and an order z_r, both controlling how many patches can come from each ROI_r^i.

To obtain patches, we may perform a grid patching operation on each ROI independently. However, this would naturally result in overlapping patches with redundant feature extraction. Our aim is to cover the whole image with patches *without any overlap*. To avoid redundancy, overlapping patches between regions with a lower order (*e.g.*, order $z = 1$) and those with a higher order (*e.g.*, order $z = 2$) are excluded from the patch set of the low-order regions. This selective inclusion process ensures that each patch belongs uniquely to the ROI with the highest priority, as indicated by the order. Specifically,

$$\mathrm{P}^i = \bigcup_{r_1=0}^{R} \left(\mathrm{P}^i_{\text{ROI}_{r_1}} - \bigcup_{r_2=r_1+1}^{R} \mathrm{P}^i_{\text{ROI}_{r_2}} \right), \tag{2}$$

where $\mathrm{P}_{\text{ROI}_r^i}$ represents the set of patches for region ROI_r of image i, and r denotes the index of each ROI, ordered by their respective priorities for patch inclusion.

This approach allows us to dynamically allocate critical regions with more patches while ensuring that the entire image is represented by patches without repetition. Also, the scale inconsistency is mitigated as long as the ROIs are

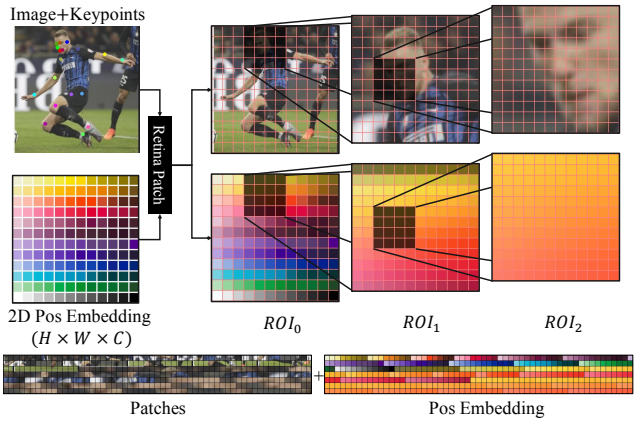

Figure 4. Illustration of Retina Patch and Position Encoding computation. **Top:** It shows three different ROIs generating patches at various scales (e.g., full image, upper torso, face). It also shows the corresponding position encodings sampled from the same spatial locations as the patches, allowing ViT to infer spatial context and understand where each patch originated within the image. **Bottom:** patches and position embedding created by Retina Patch.

semantically defined (*e.g., face, upper torso*). The number of patches within each ROI is kept consistent across images, ensuring that each patch covers a similar scale within its designated ROI. Fig. 3 uses an example to compare the vanilla grid patch of ViT with our proposed Retina Patch.

Computing ROI. Retina Patch is a generic algorithm that can work for any class of images by designing ROIs for the particular domain. In this paper, for recognizing a subject from a human image, we set the ROIs in 3 parts: 1) whole image, 2) upper torso and 3) face. The upper torso and face ROIs are computed using the off-the-shelf body keypoint detector [7]. Details on transforming the keypoints into a bounding box can be found in Supp.

Tokenization. The input to ViT's transformer block is a set of tokens or feature vectors. Since each patch's size is dependent on both the ROI size and the number of patches m_r, the size of each patch may not be the same across ROIs. We simply resize all patches to be the size of patches from the whole image ROI_0^i. We then use a linear layer to map each patch to the desired dimension, as in ViT.

Position Embedding. Since Transformer operates on sets of tokens without inherent order, Position Embedding (PE) is crucial for informing ViT of the spatial origin of each patch within the original image. For tokens of Retina Patch, we cannot use a traditional PE as the patch's source location is dynamic. Thus, we propose a Region-Sampled PE.

Let $\text{PE} \in \mathbb{R}^{C \times H \times W}$ be the fixed 2D sin-cosine position embedding [4, 10] for the whole image. Given a normalized region of interest $\text{ROI}_r^i = (x_r^i, y_r^i, h_r^i, w_r^i)$ with values between 0 and 1, we define a sampling grid $\text{Grid}_{\text{ROI}_r^i}$ over the region $[x_r^i, x_r^i + w_r^i]$ and $[y_r^i, y_r^i + h_r^i]$ within the position embedding PE. Let (h'_r, w'_r) be the target output shape for $\text{PE}_{\text{ROI}_r^i}$, such that $h'_r \cdot w'_r = m_r$, the desired number of patches for ROI_r^i. The Region Sampled PE, $\text{PE}_{\text{ROI}_r^i}$ is then obtained by bilinearly interpolating PE at the points in $\text{Grid}_{\text{ROI}_r^i}$ to match the shape (h'_r, w'_r):

$$\text{PE}_{\text{ROI}_r^i} = \text{GridSample}(\text{PE}, \text{Grid}_{\text{ROI}_r^i}, (h'_r, w'_r)) + v_r. \quad (3)$$

We add a leanable parameter $v_r \in \mathbb{R}^c$ to $\text{PE}_{\text{ROI}_r^i}$ to indicate ROI level. In summary, we create region-specific position embeddings to differentiate between patches from distinct areas of the image. An example is shown in Fig. 4.

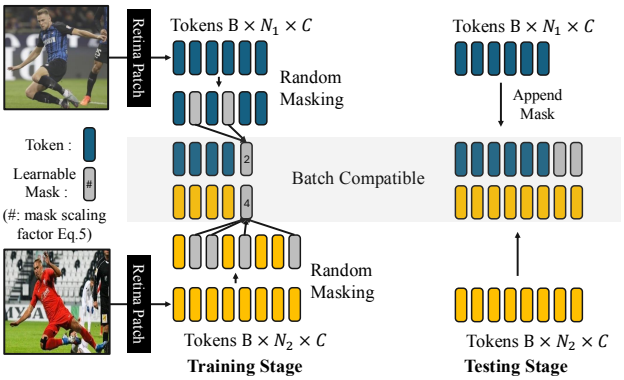

Figure 5. Illustration of Masked Recognition Backbone with masking and attention scaling trick for batched input during training. In testing, we pad with mask tokens to make the length the same.

3.2. Masked Recognition Model (MRM)

For each image, Retina Patch results in different numbers of tokens because different ROIs create different areas of intersection. For example, the number of patches from ROI_0 in Fig. 4 is 12×12 but the upper torso ROI_1 subtracts 4×4 patches from ROI_0 to avoid overlap. This operation leads to a different number of tokens per image, which prevents us from training and testing with batched inputs. To address the token inconsistency, we propose the Masked Recognition Model (MRM), introducing two key techniques: (1) masking with attention scaling and (2) a variable masking rate.

Masking with Attention Scaling. During training, we select tokens to keep. Unlike MAE [22], which discards the masked tokens, we replace them with a learnable mask token. We do this because (i) the mask token will be used during testing for padding the input, and (ii) this allows the model to explicitly know *how many* tokens are masked. Yet, since all masked tokens share the same value, we can reduce computation by applying the Attention Scaling Trick.

Specifically, although there are multiple masked tokens, we can achieve the same effect with a single mask token by adjusting its attention scores to reflect the total number of masked tokens. Let n_i be the total number of tokens for i-th image, n_k be the number of tokens we keep, and

$n_{m,i} = n_i - n_k$ be the number of masked tokens. We modify the attention computation in the Transformer as:

$$A = \text{softmax}\left(\mathbf{Q}\mathbf{K}^\top / \sqrt{d} + \boldsymbol{\delta}\right), \tag{4}$$

where $\mathbf{Q} \in \mathbb{R}^{(n_k+1)\times d}$ and $\mathbf{K} \in \mathbb{R}^{(n_k+1)\times d}$ are the query and key matrices with tokens to keep and one mask token. d is the embedding dimension. We add a bias matrix $\boldsymbol{\delta} \in \mathbb{R}^{n\times n}$ so that it is mathematically equivalent to repeating the mask tokens $n_{m,i}$ times during attention computation.

$$\boldsymbol{\delta}_{ij} = \begin{cases} \log n_{m,i}, & \text{if } j \text{ is the mask token,} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

In summary, we reduce the number of tokens from n_i to $(n_k + 1)$. Note that $(n_k + 1)$ is fixed and not image dependent. But we adjust the attention to make it equivalent to using n_i tokens where $n_{m,i}$ tokens are replaced by learnable mask tokens (proof in the Supp.). By applying the Attention Scaling Trick, we handle varying token counts in training. Also in practice, n_k is set to be about $1/3$ of n_i, masking 66% of tokens for the speed gain. During testing, we simply find the longest token length and pad the others with the mask token to batchify the inputs. An illustration is in Fig. 5.

Variable Masking Rate. As we view masked training as a form of augmentation, we randomize n_k during training and adjust the batch size correspondingly. For each batch, let $\hat{n}_k$ be the sampled number of tokens to keep,

$$\hat{n}_k = n_k + (n_i - n_k) \cdot e^{-\lambda \cdot U(0,1)}. \tag{6}$$

λ is a scaling factor, and $U(0, 1)$ denotes a random uniform distribution between 0 and 1. In short, $\hat{n}_k$ is sampled from a distribution that peaks at n_k and exhibits an exponential decay in probability toward n_i (see Supp. for its visualization).

With a randomized token length n_k, we adjust the batch size B based on the relationship $n_k^2 \propto \frac{1}{B}$, where increasing n_k would require decreasing B to maintain the same GPU memory and FLOP. And we adjust the learning rate according to the effective batch size $\mathcal{L}_{\text{adj}} = \mathcal{L}_{\hat{n}_k} \times B_{\hat{n}_k}/B_{n_k}$ to maintain consistent gradient magnitudes per sample.

The effect of (1) masking with attention scaling and (2) variable masking rate is ablated in Tab 5. While (1) and (2) are both helpful, the effect of (2) is more pronounced.

3.3. Semantic Attention Head (SAH)

In biometric recognition, the head module is key for converting the backbone's output feature map into a compact feature vector for recognition. Face recognition models flatten the feature map and apply linear layers [13, 34], while body recognition models use horizontal pooling [7, 74]. However, these approaches rely on input image alignment (aligned face or standing body) which fails when there are large pose variations. To tackle this, we introduce a Semantic Attention Head (SAH) that extracts semantic part features from key body parts, making the representation less sensitive to pose.

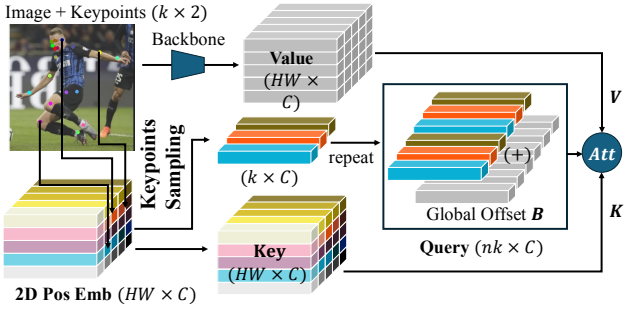

Figure 6. Illustration of semantic pooling in Semantic Attention Head. Keypoints (e.g., nose, feet) are used to grid-sample position embeddings (PE), forming queries that repeat n times and added with a global offset bias $\mathbb{B}$. This setup enables attention to slightly varied locations around each keypoint. Value comes from ViT backbone and Key is the PE. Result is a learned pooling mechanism.

Our method uses keypoints (*e.g.*, nose, hip) for capturing semantic parts. But instead of sampling features only at keypoints, which may miss the surrounding context, SAH *learns* to pool features around each keypoint. We construct a semantic query $\mathbf{Q}^i_{kp}$ (*e.g.*, nose) using 2D position embeddings (PE) from the backbone, sampled at keypoint locations:

$$\mathbf{Q}^i_{kp} = \text{GridSample}(\text{PE}, \text{kp}^i) + \mathbf{B}, \tag{7}$$

where PE is the fixed 2D image position embedding. $\text{kp}_i \in \mathbb{R}^{nk\times 2}$ is the image-specific predicted keypoints [7]. We duplicate keypoints n times and add shared bias $\mathbf{B} \in \mathbb{R}^{nk\times C}$. The purpose of $\mathbf{B}$ is to learn to offset the center of attention so that it learns to pool from diverse locations around keypoints. Key in attention is the fixed PE. Value is the backbone's feature map. The attenton [75] with $\mathbf{Q}^i_{kp}$ captures the neighborhood of the backbone feature map around keypoints:

$$\mathbf{O}^i_{\text{part}} = \text{Attention}\left(\mathbf{Q}^i_{kp}, \text{PE}, \text{backbone}(\mathbf{X}^i)\right). \tag{8}$$

The $\mathbf{O}^i_{\text{part}} \in \mathbb{R}^{B\times k\times C}$ contains semantic part features corresponding to k keypoints. Finally, applying a multi-layer perceptron (MLP) to the flattened $\mathbf{O}^i_{\text{part}}$ produces a feature,

$$f^i = \text{MLP}(\text{flatten}(\mathbf{O}^i_{\text{part}})). \tag{9}$$

By learning to pool features adaptively around each keypoint, this attention mechanism enables pose-invariant recognition that goes beyond conventional alignment-dependent methods. Fig. 6 illustrates the attention pooling.

Training with Mixed Datasets. While SAH effectively handles pose variations, we hypothesize that key cues for recognition differ between short-term and long-term training datasets. Clothing and hairstyle, for example, are useful

Method	Arch	Train Data	Avg	LTCC (General)		PRCC (SC) [71]		CCVID (General)		Market1501		MSMT17 [67]	
				top1	mAP	top1	mAP	top1	mAP	top1	mAP	top1	mAP
CAL [20]	R50	LTCC	48.64	74.04	40.84	99.51	95.64	75.63	28.08	35.60	16.11	15.92	5.06
CAL [20]	R50	PRCC	35.07	20.69	6.19	100.00	99.76	74.48	20.86	18.97	6.47	2.56	0.69
CAL [20]	R50	LTCC+PRCC	49.69	72.41	38.12	99.54	99.01	74.83	29.43	43.65	21.03	14.48	4.44
CLIP3DReID [46]	R50	LTCC	50.89	**75.66**	**45.15**	99.43	96.43	77.28	30.01	41.66	20.33	17.45	5.50
CLIP3DReID [46]	R50	PRCC	35.14	21.30	6.19	100.00	**99.84**	71.73	19.81	20.93	7.49	3.28	0.85
SOLDIER [9]	Swin-Base	LU4M+Market1501	64.85	73.83	36.28	99.51	99.53	40.27	36.56	**97.03**	**94.04**	48.64	22.77
SOLDIER [9]	Swin-Base	LU4M+MSMT17	70.19	74.44	36.74	99.30	98.71	32.73	27.76	89.85	73.20	**91.12**	**78.01**
HAP [79]	ViT-Base	LU4M+LTCC	45.71	65.11	29.02	95.53	86.44	44.16	30.43	51.63	27.29	20.89	6.56
HAP [79]	ViT-Base	LU4M+PRCC	54.09	63.29	29.36	98.84	98.38	49.15	37.73	73.49	50.11	29.61	10.99
HAP [79]	ViT-Base	LU4M+Market1501	66.61	73.02	35.97	99.30	98.45	54.74	45.14	96.23	92.20	48.01	23.02
HAP [79]	ViT-Base	LU4M+MSMT17	66.64	67.95	32.07	99.15	96.50	37.81	30.52	80.37	57.07	89.13	75.85
HAP [79]	ViT-Base	WebBody4M (Ours)	61.49	56.80	25.88	99.72	98.26	89.00	71.65	66.18	42.41	43.61	21.42
SapiensID (Ours)	ViT-Base	WebBody4M (Ours)	**73.05**	72.01	34.56	**100.00**	98.79	**92.57**	**77.82**	88.18	68.26	67.25	31.02

(a) Short-Term ReID

Method	Arch	Train Data	Avg	LTCC (CC) [56]		PRCC (CC)		CCVID (CC) [20]		CCDA [44]		Celeb-ReID [26]	
				top1	mAP	top1	mAP	top1	mAP	top1	mAP	top1	mAP
CAL [20]	R50	LTCC	28.40	38.01	18.84	37.00	35.20	74.97	25.08	3.91	9.67	37.42	3.92
CAL [20]	R50	PRCC	24.71	6.38	3.14	55.69	55.64	71.61	17.40	2.85	8.61	23.59	2.20
CAL [20]	R50	LTCC+PRCC	29.46	33.16	16.27	45.39	45.42	73.89	26.65	3.74	9.14	37.11	3.81
CLIP3DReID [46]	R50	LTCC	30.24	41.84	**22.58**	40.81	38.38	76.28	26.69	4.31	10.18	37.31	4.02
CLIP3DReID [46]	R50	PRCC	25.79	6.63	3.17	62.40	61.97	69.32	16.38	3.17	8.89	23.82	2.17
SOLDIER [9]	Swin-Base	LU4M+Market1501	24.84	25.00	12.18	26.87	32.12	39.61	35.48	8.62	16.48	46.37	5.66
SOLDIER [9]	Swin-Base	LU4M+MSMT17	22.17	26.02	11.33	22.27	25.36	31.85	26.48	8.79	15.54	47.95	6.14
HAP [79]	ViT-Base	LU4M+LTCC	20.21	25.00	11.63	26.14	22.34	41.64	25.77	4.56	11.18	30.28	3.54
HAP [79]	ViT-Base	LU4M+PRCC	26.12	29.08	12.52	38.05	41.94	45.73	33.12	5.13	13.40	37.79	4.48
HAP [79]	ViT-Base	LU4M+Market1501	27.49	24.74	11.71	33.90	37.00	52.37	41.33	8.30	16.02	44.38	5.20
HAP [79]	ViT-Base	LU4M+MSMT17	21.61	23.47	10.74	23.82	25.00	34.54	26.81	6.27	13.33	46.37	5.77
HAP [79]	ViT-Base	WebBody4M (Ours)	44.90	22.70	9.96	54.93	49.38	88.34	68.66	28.80	41.49	65.78	18.93
SapiensID (Ours)	ViT-Base	WebBody4M (Ours)	**66.30**	**42.35**	17.79	**78.75**	**72.60**	**88.72**	**72.22**	**61.84**	**69.08**	**92.80**	**66.92**

(b) Long-Term ReID

Table 1. Generalization comparison with SoTA ReID models on two settings. "Long-term" refers to clothing change (CC) protocol of LTCC, PRCC, and CCVID datasets, while "short-term" the same clothing (SC) protocol. For other datasets, the data capture characteristics define short or long-term conditions. SapiensID demonstrates superior generalization in both settings. Our WebBody4M dataset shows higher performance in long-term ReID, but not with the dataset alone, as shown in the comparison of HAP vs SapiensID with the same training set. The proposed Retina-Patch and Semantic Attention Head are essential for learning under large pose and scale variations.

in short-term datasets but less reliable in long-term due to possible appearance changes.

To aid learning with mixed datasets which combines short-term and long-term datasets, we introduce one more measure during training. We introduce a learnable scale that controls the importance of individual part features in $(\mathbf{O}^i_{\text{part}})$ for each dataset. It is to allow the model to emphasize features that are most discriminative for each dataset. During testing, however, we can use the average scale because we do not want to utilize the knowledge about the test dataset a priori.

Specifically, let $\mathbf{W}_t \in \mathbb{R}^k$ be a weight for the t-th dataset. For each sample, we choose the weight and apply

$$f^i = \text{MLP}(\text{flatten}(\mathbf{O}^i_{\text{part}} \cdot \sigma(\mathbf{W}_t))), \tag{10}$$

where σ is the Sigmoid function, ensuring weights are between 0 and 1, controlling the influence of each of the k semantic parts. We observe that after training, short-term datasets tend to focus on the clothing and long-term datasets focus on the upper torso. The learned weight is visualized in Supp. The weight is for learning discriminative parts during training but we do not use dataset-specific weights in testing.

3.4. WebBody Dataset

To facilitate the training, we collect a large-scale, labeled human dataset from the web. Specifically, we gather 94 million images with 3.8 million celebrity names. Given the inherent noise in web-sourced name queries, we perform extensive label cleaning. First, we use YOLOv8 [31] to crop the dominant person in each image to a size of 384×384, adding padding to maintain aspect ratio. We then extract facial features using RetinaFace [14] and KP-RPE [37]. Following the approach in [86], we apply DBSCAN [16] clustering to identify the most consistent group of images for each name. By assuming all images stem from a single name query, we relax the similarity threshold beyond conventional face recognition standards. We also exclude any images with face features matching those in validation sets [25, 52, 55, 84, 85].

This process yields a labeled dataset of 4.4 million images from $217,722$ unique subjects. However, as the dataset is labeled based on facial similarity, it lacks images where the face is obscured (*e.g.* back-facing images). Thus, we incorporate additional body ReID training datasets [17, 20, 26, 56, 59, 70, 71, 83], which account for $\sim$10% of the final dataset. The resulting dataset—named WebBody4M—comprises 4.9 million images and $263,920$ subjects in total. WebBody4M is the largest labeled dataset to date with high pose and scale variation. The keypoint visibility distribution of different body parts in Supp. shows a predominance of visible upper body, with visibility decreasing gradually down the body (around 17% visible ankles). An example of the WebBody4M dataset can be seen in Fig. 1.

The dataset collection and label cleaning procedure is

Method	Arch	Train Data	Avg	WebBody Testset top1	mAP
CAL [20]	R50	PRCC	2.47	4.29	0.64
CAL [20]	R50	LTCC	3.79	6.57	1.02
SOLDIER [9]	Swin-Base	Market1501	3.22	5.42	1.02
SOLDIER [9]	Swin-Base	MSMT17	5.96	9.95	1.98
HAP [79]	ViT-Base	LTCC	1.74	2.89	0.58
HAP [79]	ViT-Base	PRCC	2.61	4.37	0.85
HAP [79]	ViT-Base	Market1501	4.31	7.22	1.39
HAP [79]	ViT-Base	MSMT17	4.87	8.22	1.52
HAP [79]	ViT-Base	WebBody4M	47.12	64.36	29.89
SapiensID (Ours)	ViT-Base	WebBody4M	**64.41**	**76.82**	**52.00**

Table 2. ReID Performance on variable pose and scale settings.

similar to WebFace4M dataset [86]. We compare the face-cropped version of WebBody4M with WebFace4M and observe that an FR model trained on WebBody4M-FaceCrop is similar in performance to WebFace4M (see details in the Supp.). Separate from the WebBody4M, we also prepare a test set called WebBody-test to evaluate the cross pose-scale ReID performance. It comprises $96,624$ images of $4,000$ gallery and probe subjects. Examples are shown in Fig. 2.

4. Experiments

Implementation Details. To train SapiensID on Webbody4M, we use AdaFace [34] loss and ViT-Base with KP-RPE as the main backbone [37], following the convention of face recognition model training pipeline. We do not include additional losses such as Triplet Loss [53] since there are a sufficient number of subjects in the training set. Input image size is 384×384 with white padding if the aspect ratio is not 1. We use 3 ROIs (whole image, upper torso, and head) and the grid size per ROI is 12×12 leading to a maximum 144×3 number of patches. With masked recognition training, we replace at most 66% of tokens with mask (Sec. 3.2), leading to ~ 9 times speed up in training. The masking probability and batch size rule are discussed in Supp. We use 7 H100 GPUs to train the whole model in 2 days, starting from scratch.

Whole Body ReID. The task identifies individuals walking or standing in distant camera views, categorized into short or long-term scenarios based on the time gap between captures and the likelihood of clothing changes. Tab. 1 shows our results on the ReID benchmarks. A significant departure from prior works is the use of a single SapiensID model across all evaluation settings, whereas previous methods employ fine-tuned models for each evaluation dataset (one model per dataset). This distinction highlights SapiensID's potential for deployment in diverse, unseen, real-world environments.

SapiensID achieves the highest average mAP of 73.05% across short-term ReID benchmarks. Furthermore, we attain SoTA results on all evaluated long-term ReID datasets. This strong performance underscores the value of the WebBody4M dataset in training a generalizable model. However, this achievement would not have been possible without our SapiensID architecture, which effectively handles variations in pose and visible body areas. A strong baseline (HAP [79]) trained on WebBody4M alone does not achieve comparable results, highlighting the importance of our architectural innovations to leverage the dataset. SapiensID marks a significant advance by being the first single model capable of strong performance across short and long-term ReID tasks.

Method	Training Data	OccludedReID top1	mAP
KPR [57] + SOLDIER	LU4M +OccludedReID	84.80	**82.60**
SapiensID	WebBody4M	**87.30**	75.57

Table 3. Performance in occluded ReID. SapiensID achieves a higher top-1 accuracy, while KPR [57] shows a higher mAP. SapiensID is trained without OccludedReID training data.

Method	AdaFace-ViT [34]	SapiensID (Ours)
Train Data	WebBody4M-FaceCrop	WebBody4M
LFW [25]	99.82	99.82
CPLFW [84]	95.12	94.85
CFPFP [55]	99.19	98.74
CALFW [85]	96.07	95.78
AGEDB [52]	97.97	97.33
Face Avg	**97.63**	97.31
LTCC [56]	21.70	72.01
Market1501 [82]	7.81	88.18
Body Avg	14.76	**80.10**
Combined Avg	56.19	**89.80**

Table 4. Performance on cross-modality setting. Face recognition is evaluated on aligned face recognition datasets and body recognition is evaluated on short-term ReID datasets. LTCC and Market1501 measure top1 of short-term setting.

Cross Pose-Scale ReID. Real-world human recognition can present scenarios where subjects are captured across varying camera viewpoints and exhibit diverse poses, such as sitting, bending, or engaging in activities. For example, a security camera might capture a person standing upright, while a social media photo shows the same individual sitting in a cafe. This poses a challenge for conventional ReID systems. We refer to this setting as Cross Pose-Scale ReID.

To evaluate this setting, we introduce the WebBody-Test dataset, specifically designed to encompass such pose and scale variations. Tab. 2 details the performance comparison on this dataset. Conventional ReID models struggle to generalize to this scenario due to the significant shift in visual appearance caused by pose and scale changes. SapiensID with the highest performance establishes a strong baseline for this research area. Since the task itself is challenging, there is still room for improvement. WebBody dataset demonstrates the potential of SapiensID to address the complexities of Cross Pose-Scale ReID, while offering a valuable starting point for future research in this area.

Occluded ReID. Occlusions, whether due to obstacles in the scene or self-occlusion from the subject's pose, present a further challenge for robust human recognition. We evaluate SapiensID in occluded scenarios on the OccludedReID dataset [87], comparing with KPR [57], a SoTA method designed for occlusion handling. As shown in Tab. 3, SapiensID achieves a competitive performance of top-1

	All	Face	Whole Body ReID	
			Short	Long
(1) ViT	59.54	90.63	56.17	31.81
(2) ViT+RP	66.35	92.93	59.16	46.95
(3) ViT+SAH	71.67	95.84	72.63	46.55
(4) ViT+RP+SAH (SapiensID)	**78.67**	**96.66**	**73.05**	**66.30**
(4) − Learned Mask	76.99	96.08	70.44	64.46
(4) − Variable n_k	74.39	95.95	69.58	57.64

Table 5. Ablation study of SapiensID. Face is the average accuracy of CPLFW, CFPFP, CALFW, and AGEDB. Short and Long Term use the average of the datasets in Tab 1. Results show the necessity and strong complementarity of both RP and SAH in SapiensID.

		LTCC CC		PRCC CC	
		Top1	mAP	Top1	mAP
1	None	0.00	3.56	1.47	4.28
2	1+Nose	25.77	5.78	27.21	21.04
3	2+Eye	30.61	8.87	63.87	55.17
4	3+Mouth	38.01	11.81	73.36	65.05
5	4+Ear	39.80	14.05	77.65	70.45
6	5+Shoulder	41.84	15.82	79.73	73.14
7	6+Elbow	41.07	16.64	**80.55**	**73.54**
8	7+Wrist	41.07	17.16	79.34	73.16
9	8+Hip	40.56	17.50	79.99	73.38
10	9+Knee	42.35	17.73	79.00	72.88
11	10+Ankle (Full)	**42.35**	**17.79**	78.75	72.6

Table 6. Impact of adding body parts on ReID. None means all features are zeroed out. Each row adds features to the previous row.

87.30%, demonstrating its strong ability to handle occlusions even without being explicitly trained on the OccludedReID dataset. This result further underscores the value of our architecture and training dataset in learning representations that are resilient to real-world challenges like occlusions.

Face Recognition. We evaluate on traditional aligned face recognition benchmarks to assess the ability to handle FR tasks. Tab. 4 compares SapiensID with a SoTA FR model, AdaFace [34], both with a ViT-Base backbone. AdaFace is trained on faces aligned and cropped to 112×112 by [14]. AdaFace achieves a slightly higher average accuracy of 97.63% across five benchmarks. This marginal difference is expected, given AdaFace's training on tightly cropped, aligned faces. However, SapiensID's performance remains highly competitive, bridging the gap between specialized face recognition and general human recognition tasks.

While AdaFace excels in FR datasets, its performance degrades when applied to ReID datasets which contain images without visible face region (*e.g.* back of the head). AdaFace is evaluated by cropping faces using [14]. In contrast, SapiensID maintains strong performance across both modalities. More experiments can be found in Supp B.10 and B.11.

Ablation of Components. Tab. 5 ablates SapiensID's key components: Retina Patch (RP) and Semantic Attention Head (SAH). Starting from a simple ViT backbone with AvgMax pooling [20] as a baseline, we progressively incorporate RP and SAH to analyze their individual and combined contributions. Performance is evaluated across face recognition and both short-term and long-term ReID. The results show that both RP and SAH are essential.

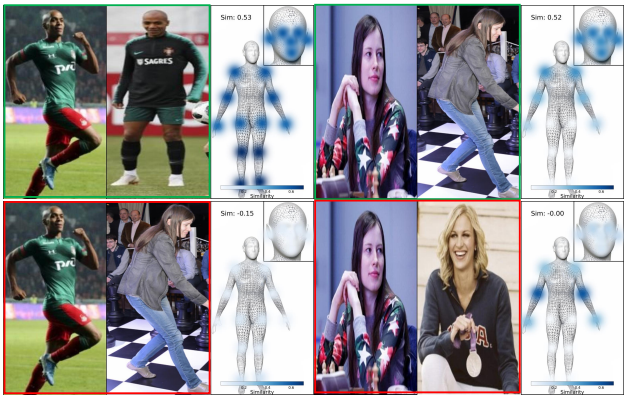

Figure 7. Part Similarity Visualization. Top shows the same subject pairs. Bottom shows different subject pairs. Part features provide some indication of where the similar parts are, but the final similarity is generated through a nonlinear mapping of the part features.

We also show the importance of MRM. (4) - Learned Mask means using MAE [22] to simply drop tokens. (4) - Variable n_k is fixing n_k without sampling. The result shows that learned mask is of some benefit while changing the masking rate during training is of larger benefit.

Analysis of Part Contribution. To see the impact of body parts in recognition, we erase part features by making them zero. Tab. 6 shows a trend of performance gain as more parts are added. For LTCC dataset accuracy increases from 25.77% to 42.35% as body parts from the nose to ankle are incorporated. This suggests that including the full range of body parts aids recognition. In contrast, PRCC achieves high performance by using upper body cues, reaching a top-1 accuracy of 80.55% with parts up to the shoulder and elbow. Lower body features add minimal or even negative value. This analysis implies the benefit of scenario-specific adjustments where relevant body regions can optimize recognition performance. We also visualize the part features similarity with sample images from the test set of WebBody4M in Fig 7. Samples of different scales and poses are visualized.

5. Conclusion

SapiensID presents a paradigm shift in human recognition, moving beyond modality-specific models to a unified architecture capable of identification across diverse poses and body-part scales. Retina Patch, Semantic Attention Head, and Masked Recognition Model combined with WebBody4M dataset, enable SapiensID to achieve SoTA performance across various ReID benchmarks and establish a strong baseline for Cross Pose-Scale ReID. This work marks a step towards holistic human recognition systems. We include an in-depth discussion of the ethical impacts in Supp, ensuring that our approach respects intellectual property, privacy, and responsible data use.

Acknowledgments. This research is based upon work supported 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] InsightFace. https://github.com/deepinsight/insightface.git. Accessed: 2021-9-1. 1

[2] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *ICCV*, 2021. 1

[3] Romain Beaumont. img2dataset: Easily turn large sets of image urls to an image dataset. https://github.com/rom1504/img2dataset, 2021. 10

[4] Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Better plain vit baselines for imagenet-1k. *arXiv preprint arXiv:2205.01580*, 2022. 4

[5] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *CVPR*, 2023. 3

[6] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *ICCV*, 2017. 2

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *PAMI*, 2019. 4, 5

[8] Jiaxing Chen, Xinyang Jiang, Fudong Wang, Jun Zhang, Feng Zheng, Xing Sun, and Wei-Shi Zheng. Learning 3d shape feature for texture-insensitive person re-identification. In *CVPR*, 2021. 2

[9] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *CVPR*, 2023. 6, 7, 10

[10] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 4

[11] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. Low-resolution face recognition. In *ACCV*, 2018. 2, 5

[12] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. In *NeurIPS*, 2024. 3

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1, 2, 3, 5

[14] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020. 1, 2, 6, 8

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[16] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, 1996. 6

[17] Dengpan Fu, Dongdong Chen, Hao Yang, Jianmin Bao, Lu Yuan, Lei Zhang, Houqiang Li, Fang Wen, and Dong Chen. Large-scale pre-training for person re-identification with noisy labels. *arXiv preprint arXiv:2203.16533*, 2022. 6

[18] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 2

[19] Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, et al. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In *NeurIPS*, 2020. 2

[20] 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. 1, 2, 6, 7, 8, 3

[21] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016. 1, 2

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 4, 8, 1

[23] 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*, 43(5):1789–1800, 2010. 1

[24] Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *CVPR*, 2021. 2

[25] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled Faces in the Wild: A database forstudying face recognition in unconstrained environments. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008. 1, 2, 6, 7

[26] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Celebrities-ReID: A benchmark for clothes variation in long-term person re-identification. In *IJCNN*, 2019. 2, 6

[27] Yan Huang, Jingsong Xu, Qiang Wu, Yi Zhong, Peng Zhang, and Zhaoxiang Zhang. Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *TCSVT*, 2019. 2

[28] 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 *ECCV*, 2020. 1

[29] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang.

CurricularFace: adaptive curriculum learning loss for deep face recognition. In *CVPR*, 2020. 1, 2

[30] 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. 1, 2

[31] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. 6, 4

[32] Nathan D Kalka, Brianna Maze, James A Duncan, Kevin O'Connor, Stephen Elliott, Kaleb Hebert, Julia Bryan, and Anil K Jain. IJB–S: IARPA Janus Surveillance Video Benchmark. In *BTAS*, 2018. 2, 6

[33] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2025. 1

[34] Minchul Kim, Anil K Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *CVPR*, 2022. 1, 2, 3, 5, 7, 8

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

[36] Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. DCFace: Synthetic face generation with dual condition diffusion model. 2023. 1

[37] Minchul Kim, Yiyang Su, Feng Liu, Anil Jain, and Xiaoming Liu. Keypoint relative position encoding for face recognition. In *CVPR*, 2024. 2, 6, 7, 1

[38] Yonghyun Kim, Wonpyo Park, and Jongju Shin. BroadFace: Looking at tens of thousands of people at once for face recognition. In *ECCV*, 2020. 1

[39] Abhinav Kumar, Tim K. Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *CVPR*, 2020. 2

[40] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, 2018. 2

[41] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *PAMI*, 2019. 2

[42] Yu-Jhe Li, Xinshuo Weng, and Kris M Kitani. Learning shape representations for person re-identification under clothing change. In *WACV*, 2021. 1, 2, 3

[43] Yutian Lin, Xuanyi Dong, Liang Zheng, Yan Yan, and Yi Yang. A bottom-up clustering approach to unsupervised person re-identification. In *AAAI*, 2019. 2

[44] Feng Liu, Minchul Kim, ZiAng Gu, Anil Jain, and Xiaoming Liu. Learning clothing and pose invariant 3d shape representation for long-term person re-identification. In *ICCV*, 2023. 1, 2, 6

[45] Feng Liu, Ryan Ashbaugh, Nicholas Chimitt, Najmul Hassan, Ali Hassani, Ajay Jaiswal, Minchul Kim, Zhiyuan Mao, Christopher Perry, Zhiyuan Ren, et al. Farsight: A physics-driven whole-body biometric system at large distance and altitude. In *WACV*, 2024. 1

[46] Feng Liu, Minchul Kim, Zhiyuan Ren, and Xiaoming Liu. Distilling CLIP with dual guidance for learning discriminative human body shape representation. In *CVPR*, 2024. 1, 6

[47] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 1, 2, 3

[48] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[49] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[50] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. IARPA Janus Benchmark-C: Face dataset and protocol. In *ICB*, 2018. 2

[51] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. MagFace: A universal representation for face recognition and quality assessment. In *CVPR*, 2021. 3

[52] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. AGEDB: the first manually collected, in-the-wild age database. In *CVPRW*, 2017. 2, 6, 7

[53] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 7

[54] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 10

[55] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *WACV*, 2016. 2, 6, 7

[56] Xiujun Shu, Xiao Wang, Xianghao Zang, Shiliang Zhang, Yuanqi Chen, Ge Li, and Qi Tian. Large-scale spatio-temporal person re-identification: Algorithms and benchmark. *TCSVT*, 2021. 1, 2, 6, 7

[57] Vladimir Somers, Alexandre Alahi, and Christophe De Vleeschouwer. Keypoint promptable re-identification. In *ECCV*, 2025. 7, 4

[58] Yiyang Su, Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Open-set biometrics: Beyond good closed-set models. In *ECCV*, 2024. 2

[59] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 6

[60] Ying Tai, Yicong Liang, Xiaoming Liu, Lei Duan, Jilin Li, Chengjie Wang, Feiyue Huang, and Yu Chen. Towards highly accurate and stable face alignment for high-resolution videos. In *AAAI*, 2019. 2

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[62] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *CVPRW*, 2020. 2

[63] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. NormFace: L2 hypersphere embedding for face verification. In *ACMMM*, 2017. 1

[64] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *CVPR*, 2018. 1, 2, 3

[65] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *CVPR*, 2018. 2

[66] Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. In *NeurIPS*, 2021. 3

[67] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018. 1, 6

[68] Frederick W Wheeler, Xiaoming Liu, and Peter H Tu. Multi-frame super-resolution for face recognition. In *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2007. 1

[69] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. IARPA Janus Benchmark-B face dataset. In *CVPRW*, 2017. 2, 5

[70] Peng Xu and Xiatian Zhu. DeepChange: A large long-term person re-identification benchmark with clothes change. 2021. 6

[71] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *PAMI*, 2019. 1, 2, 3, 6

[72] Zhilin Yang. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019. 1

[73] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 2

[74] Dingqiang Ye, Chao Fan, Jingzhe Ma, Xiaoming Liu, and Shiqi Yu. Biggait: Learning gait representation you want by large vision models. In *CVPR*, 2024. 5

[75] Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. *arXiv preprint arXiv:2410.05258*, 2024. 5

[76] Xi Yin, Ying Tai, Yuge Huang, and Xiaoming Liu. Fan: Feature adaptation network for surveillance face recognition and normalization. In *ACCV*, 2020. 1

[77] Hong-Xing Yu, Wei-Shi Zheng, Ancong Wu, Xiaowei Guo, Shaogang Gong, and Jian-Huang Lai. Unsupervised person re-identification by soft multilabel learning. 2019. 2

[78] Shijie Yu, Shihua Li, Dapeng Chen, Rui Zhao, Junjie Yan, and Yu Qiao. COCAS: A large-scale clothes changing person dataset for re-identification. In *CVPR*, 2020. 2

[79] Junkun Yuan, Xinyu Zhang, Hao Zhou, Jian Wang, Zhongwei Qiu, Zhiyin Shao, Shaofeng Zhang, Sifan Long, Kun Kuang, Kun Yao, et al. Hap: Structure-aware masked image modeling for human-centric perception. In *NeurIPS*, 2023. 6, 7, 10

[80] Yunpeng Zhai, Shijian Lu, Qixiang Ye, Xuebo Shan, Jie Chen, Rongrong Ji, and Yonghong Tian. Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification. In *CVPR*, 2020. 2

[81] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *Signal Processing Letters*, 2016. 2

[82] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 1, 2, 7

[83] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 6

[84] Tianyue Zheng and Weihong Deng. Cross-Pose LFW: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5:7, 2018. 2, 6, 7

[85] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-Age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017. 1, 2, 6, 7

[86] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. WebFace260M: A benchmark unveiling the power of million-scale deep face recognition. In *CVPR*, 2021. 1, 2, 6, 7, 3

[87] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification. In *ICME*, 2018. 7, 4