# Distilling CLIP with Dual Guidance for Learning Discriminative Human Body Shape Representation

Feng Liu, Minchul Kim, Zhiyuan Ren, Xiaoming Liu
Department of Computer Science and Engineering
Michigan State University, East Lansing MI 48824
{liufeng6,kimminc2,renzhiy1,liuxm}@msu.edu

## Abstract

*Person Re-Identification (ReID) holds critical importance in computer vision with pivotal applications in public safety and crime prevention. Traditional ReID methods, reliant on appearance attributes such as clothing and color, encounter limitations in long-term scenarios and dynamic environments. To address these challenges, we propose CLIP3DReID, an innovative approach that enhances person ReID by integrating linguistic descriptions with visual perception, leveraging pretrained CLIP model for knowledge distillation. Our method first employs CLIP to automatically label body shapes with linguistic descriptors. We then apply optimal transport theory to align the student model's local visual features with shape-aware tokens derived from CLIP's linguistic output. Additionally, we align the student model's global visual features with those from the CLIP image encoder and the 3D SMPL identity space, fostering enhanced domain robustness. CLIP3DReID notably excels in discerning discriminative body shape features, achieving state-of-the-art results in person ReID. Our approach represents a significant advancement in ReID, offering robust solutions to existing challenges and setting new directions for future research. Code is available.*

## 1. Introduction

Person Re-Identification (ReID) [1, 42, 46, 51, 76] is a fundamental task in computer vision, focusing on recognizing and matching individuals across diverse locations and temporal instances. This technology is crucial for applications ranging from public safety, crime prevention [19, 71], forensic identification, to security monitoring [4, 77], helping track targets across non-overlapping camera views.

Existing works in person ReID have predominantly focused on appearance feature learning [17, 18, 22, 33, 34, 80], adept at decoding superficial characteristics such as clothing and colors. However, these approaches not only show inherent limitations in practical, long-term scenarios where the goal is to recognize individuals over prolonged periods, amidst variations in clothing and human activities, but they also struggle with limited generalization ability in diverse and dynamic environments. In contrast, the human visual system has a remarkable capacity to identify individuals despite appearance changes, such as different clothing, hairstyles, or accessories (*e.g.,* bags, hats, and scarves). This capability is attributed to a holistic understanding of *appearance-irrelevant feature*. *In this work, we aim to emulate this aspect of human perception, focusing on learning the crucial appearance-irrelevant feature to overcome the current limitations in generalization and adaptability.*

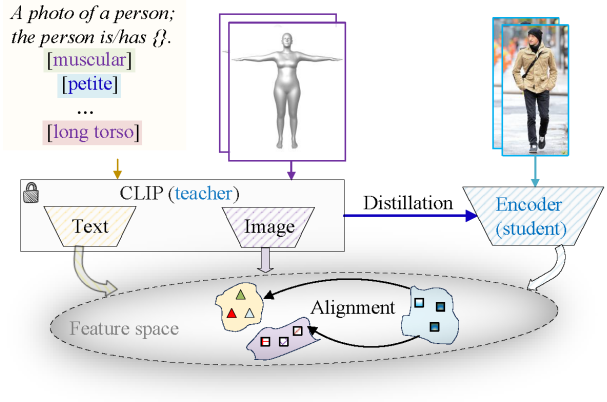


Figure 1. Key intuitions behind **CLIP3DReID**: distilling CLIP with dual guidance to learn discriminative human body shape representations for person ReID.

Early attempts have led to a variety of approaches aimed at tackling the challenge of appearance changes [20, 25, 26, 29, 39, 44, 60, 68, 73]. These methods primarily focus on extracting features that are not dependent on clothing appearance, utilizing custom architectures [26, 27] tailored training processes [39], novel loss functions [20] and advanced data augmentation [78]. Complementing these developments, an innovative series of research introduces 3D priors into person ReID. It involves either converting 2D images into a

3D space [79] or complementing by an auxiliary 3D body reconstruction branch [7], specifically aiming to extract 3D shape-related features to enhance identification accuracy in cloth-changing scenarios. Very recently, 3DInvarReID [45] introduces an innovative method to disentangle identity from pose and clothing in 3D models, while concurrently reconstructing accurate 3D naked body shapes, thereby enhancing person ReID accuracy in scenarios involving diverse activities and clothing changes. However, the lack of training datasets with real images that have 3D ground-truth body shapes presents a limitation. As a result, these approaches might not faithfully capture the intricacies of discriminative body shapes, potentially leading to limited 3D model fitting and overall identification accuracy.

To overcome these challenges, we develop methods to learn inherent discriminative body shape features for person ReID, aiming to relax the reliance on ground-truth 3D body data. Our proposed method is rooted in an intuition:

> *Linguistic descriptions of body shapes (e.g., 'a muscular, broad-shouldered person'), provide distinct and identifying characteristics in a highly concise and complementary manner.*

Thus, we posit that an effective method should involve a regularization strategy where the learned visual representations are closely aligned with these linguistic descriptions, especially focusing on discriminative body features. This approach would bridge the gap between visual perception and linguistic description, harnessing the best of both to enhance the *accuracy* and *efficiency* of person ReID systems.

Building upon this argument, we introduce a novel person ReID approach, **CLIP3DReID**, which leverages the advancements in vision-language models (VLM) like CLIP [56] to train our models. This approach aims to employ knowledge distillation from a large, **pre-trained** and **frozen** vision-language teacher model to optimize a more compact, yet effective visual encoder (student model). The essence of **CLIP3DReID** lies in aligning visual features extracted by our student encoder with linguistic descriptions, while also accurately reflecting discriminative 3D body shape features. Such a distillation process ensures that the model not only captures essential discriminative cues for person matching but also exhibits robust generalization capabilities.

As shown in Fig. 1, to effectively represent the nuanced aspects of body shape, we utilize human-annotated linguistic descriptors, *e.g.* 'muscular', 'petite', 'long torso', forming a set of paired phrases with opposite meanings. These descriptors are transformed into a continuous space using the CLIP text encoder. Concurrently, images are processed by the CLIP image encoder, enabling automatic labeling with body shape descriptors via text-vs-image feature similarities. Once the labels and a corresponding set of text features are obtained, we employ optimal transport (OT) theory to ensure that the *local* visual features of our student model are aligned with these shape-aware tokens, which represent domain-invariant features, via distribution alignment. Further, we utilize MSE loss to effectively distill rich semantic information from the CLIP image encoder into our student model, enriching its global visual feature representation.

Furthermore, we incorporate the 3D SMPL model as an additional regularization for the global feature of our student model. This is purposely implemented to enforce the reconstruction of naked body shapes, deliberately excluding any texture information. Importantly, our method requires no ground-truth shape labels in training. Using synthetic mesh images rendered from 3D body shapes without clothing or texture (see Fig. 1), our method ensures the learned visual features are domain-invariant and physically meaningful, aligning with the core goal of 3D body shape representation.

In summary, the contributions of this work include:

◇ We propose a novel person ReID method, **CLIP3DReID**, to learn discriminative body shape representation by distilling CLIP with dual guidance.

◇ We devise an innovative application of optimal transport (OT) theory, aligning the local visual features of our student model with domain-invariant, shape-aware tokens of CLIP.

◇ We present a novel approach that utilizes 3D body reconstruction for regularization, thereby enhancing domain invariance and model adaptability by using synthetic data.

◇ Extensive experiments demonstrate the superior performance of **CLIP3DReID** in person ReID.

## 2. Related Work

**Person Re-Identification.** Person ReID matches individuals across images captured by distributed cameras. Traditional approaches mainly focus on short-term scenarios, where a person's clothing remains unchanged [17, 18, 33, 34, 40, 64, 72, 74], a presumption frequently violated in real life. Recognizing this limitation, there is a growing shift towards cloth-changing ReID [20, 25, 26, 29, 39, 60, 60, 62, 67, 68, 68, 73], addressing more realistic scenarios where clothing changes are prevalent. To learn robust features, some works [20, 39, 78] explore disentangled representation learning to separate appearance and structural cues from RGB images, treating structural information as clothing-irrelevant features. Another line of research explores multi-modality information (*e.g.*, skeletons [53], silhouettes [24, 29], radio signals [15], contour sketches [15], or text [9, 16, 37]) to model body shape and extract appearance-irrelevant features. However, multi-modality methods often require additional information during inference. A new direction in this field involves 3D shape-based methods [7, 45, 79], such as 3DInvarReID [45], which learns clothing and pose invariant 3D shape feature, enhancing accuracy in diverse scenarios. However, the scarcity of real images with 3D ground-truth body

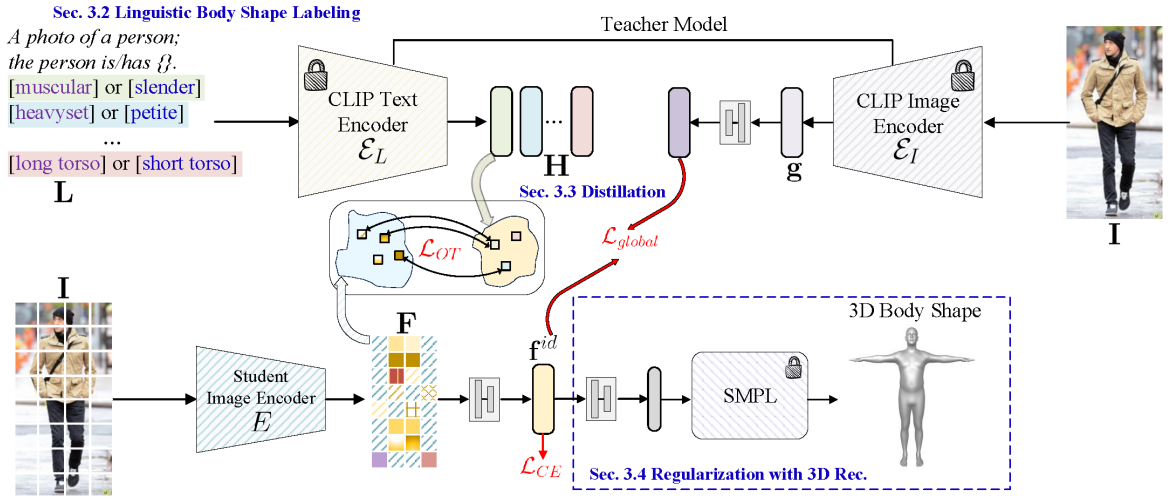


Figure 2. Overview of the proposed **CLIP3DReID**. The key contribution of our work is three effective designs: CLIP-based linguistic body shape labeling, dual distillation from CLIP, and regularization with 3D reconstruction. Incorporating these three designs into the person ReID framework enables us to learn discriminative body shape features.

shapes is a significant challenge, limiting the performance of 3D-based approaches. In contrast, our approach circumvents these limitations by harnessing linguistic cues to enhance the learning of discriminative body shape features, thereby advancing the accuracy and applicability of person ReID.

**Text-based Person Retrieval.** Text-based Person Retrieval [35] retrieves images of individuals from a large-scale gallery using textual descriptions. With the advancement of VLMs, this task has gained increasing attention [2, 3, 5, 31, 32, 36, 41, 43, 54, 59, 63, 69] and is closely related to both person ReID and text-image retrieval. However, text-based retrieval typically does not require learning discriminative features, as the textual descriptions include specific details about appearance, such as clothing and accessories. In contrast, our **CLIP3DReID** focuses on linguistic descriptions that highlight body shape cues, with the goal of extracting discriminative shape features for identification.

**CLIP-based Knowledge Distillation.** Knowledge distillation [13], where a more sophisticated 'teacher' model guides a simpler 'student' model, has seen innovative applications with the pre-trained CLIP model [56], and ChatGPT [57]. CLIP's versatility is demonstrated across a spectrum of tasks, including classification [6, 30, 81, 82], semantic segmentation [65, 83] and detection [14]. The successful distillation of the CLIP model for domain-specific tasks, such as CLIPPING [52] in video-language retrieval and RISE [28] for domain generalization, showcases its ability to transfer domain knowledge into compact networks. ZeroSeg [8] utilizes CLIP for zero-shot semantic segmentation, showcasing its capacity for semantic and visual knowledge transfer. These varied implementations highlight CLIP's capacity to endow student models with a deep understanding of both semantic content and visual nuances. In this paper, we leverage the CLIP model for the specific challenge of person ReID. We delve into the nuances of distilling CLIP's comprehensive domain knowledge into a compact visual encoder network, *e.g.*, ResNet-50, enhancing person ReID by leveraging CLIP's deep semantic insights and linguistic representation.

**Linguistic Body Shape Representation.** Our concept of employing linguistic descriptors for person ReID draws inspiration from several works [10, 23, 50, 55, 61], and early work on attribute-based reID [58]. Notably, BodyTalk [61] and SHAPY [10] demonstrate the creation of perceptually and metrically accurate 3D body models by correlating linguistic attributes with body shape parameters. These investigations emphasize the potential of linguistic attributes as discriminative features for 3D body reconstruction. Building on the concept that language can evoke vivid visual representations of body shapes, our research takes a novel step. We integrate linguistic descriptors with the CLIP model to learn linguistic body shape representations. To the best of our knowledge, this represents the *first* initiative to utilize the CLIP model in this unique capacity, innovatively bridging the gap between linguistic attributes and body shape representation within the context of person ReID.

## 3. Methodology

### 3.1. Overview

Fig. 2 provides an overview of CLIP3DReID. For each mini-batch of $B$ training samples, denoted as $\{(\mathbf{I}_i, y_i, \mathbf{L}_i)\}_{i=1}^{B}$, the input consists of human images $\mathbf{I}_i$, the identity label of the image $y_i$, and a set of linguistic descriptors of body

shape $\mathbf{L}_i$, which is detailed in Sec.3.2. We denote the **pre-trained** and **frozen** CLIP teacher text and image encoders as $\mathcal{E}_L$ and $\mathcal{E}_I$, respectively. The focus of our optimization is the student visual encoder, represented as $E$. This encoder is ResNet-50 [21], a standard architecture in person ReID. Such a choice underlines our commitment to balancing between model compactness and the established benchmarks of performance in the field.

The CLIP teacher image encoder $\mathcal{E}_I$ processes the input image $\mathbf{I}$ and generates a feature vector $\mathbf{g} \in \mathbb{R}^d$. In the language component, the CLIP teacher text encoder $\mathcal{E}_L$, working with a set of $M$ linguistic body shape descriptors $\mathbf{L} = \{l_m\}_{m=1}^M$, produces text feature sets $\mathbf{H} = \{\mathbf{h}_m\}_{m=1}^M \in \mathbb{R}^{M\times d}$. The student image encoder $E$ also takes $\mathbf{I}$ as input and outputs local image feature maps $\mathbf{F} = \{\mathbf{f}_n\}_{n=1}^N \in \mathbb{R}^{N\times d'}$, where $N$ is the number of patches. The operations are formally outlined as:

$$\mathbf{g} = \mathcal{E}_I(\mathbf{I}), \quad \mathbf{H} = \mathcal{E}_L(\mathbf{L}), \quad \mathbf{F} = E(\mathbf{I}). \tag{1}$$

To aggregate the local patch image embeddings $\mathbf{F}$ into a single global feature $\mathbf{f}^{id} \in \mathbb{R}^{d'}$, we employ a multilayer perceptron (MLP) with a single hidden layer. During person ReID inference, the similarity between two images is determined using the cosine similarity of their respective features $\mathbf{f}^{id}$. It is worth noting that, the inference process of our ReID system solely relies on the student image encoder $E$, without the need for any additional modules.

### 3.2. Design 1: Labeling Linguistic Body Description

In this section, we elaborate on the process of obtaining $\mathbf{L}$, the linguistic descriptors, for each training image. We draw inspiration from studies highlighting the effectiveness of describing human body shapes linguistically [10, 23, 50, 55, 61], particularly the BodyTalk system [61]. BodyTalk employs 30 linguistic attributes to represent 3D body meshes derived from the SMPL model's shape space, used to train a linear "attribute to shape" regressor.

However, manually annotating these labels for images is a significant challenge. To address this, we leverage the pre-trained CLIP model to automatically label images with appropriate body shape descriptors. For compatibility with CLIP, we ***re-design*** our set of linguistic body shape descriptors, which consists of $M = 16$ pairs of phrases with opposite meanings, as detailed in Tab. 1. Each descriptor is carefully selected for its effectiveness in creating discriminative body shape representations. *The strength of these text descriptors lies in their unwavering consistency, regardless of varying distances, camera views, or variations in clothing and accessories. Such properties are key to unlocking the potential for generalizable representation of body features.*

**Multiple Prompts Determination.** For each human image $\mathbf{I}$, we utilize a series of prompts $\mathbf{L} = \{l_m\}_{m=1}^M$ to generate linguistic descriptor labels. These prompts are transformed into a set of text features $\mathbf{H} = \{\mathbf{h}_m\}_{m=1}^M$. To provide a structured approach, we employ a standardized context prompt, $l =$"*A photo of a person; the person is/has* $\cdot$.", for each pair of contrasting phrases, $P_m^1$ and $P_m^2$. This results in two distinct prompts, $l^{p_m^1}$ and $l^{p_m^2}$, which are then processed by the CLIP text encoder to yield outputs $\mathbf{h}^{p_m^1}$ and $\mathbf{h}^{p_m^2}$, respectively. Given the image feature $\mathbf{g}$ of $\mathbf{I}$, we compute the Cosine Similarities for each phrase pair: $score_m^1 = CS(\mathbf{g}, \mathbf{h}^{p_m^1})$ and $score_m^2 = CS(\mathbf{g}, \mathbf{h}^{p_m^2})$. The final text feature $\mathbf{h}_m$ is determined based on these scores:

$$\mathbf{h}_m = \begin{cases} \mathbf{h}^{p_m^1}, & \text{if } score_m^1 > score_m^2, \\ \mathbf{h}^{p_m^2}, & \text{otherwise.} \end{cases} \tag{2}$$

This method effectively leverages the linguistic descriptors in conjunction with the CLIP model, ensuring that the most relevant features are captured for each image.

| | Phrase 1 | | Phrase 2 |
|---|---|---|---|
| 1 | Muscular | $\leftrightarrow$ | Slender |
| 2 | Broad-Shouldered | $\leftrightarrow$ | Narrow-Shouldered |
| 3 | Heavyset | $\leftrightarrow$ | Petite |
| 4 | Tall | $\leftrightarrow$ | Short |
| 5 | Long Legs | $\leftrightarrow$ | Short Legs |
| 6 | Long Torso | $\leftrightarrow$ | Short Torso |
| 7 | Curvy | $\leftrightarrow$ | Angular |
| 8 | Full-Figured | $\leftrightarrow$ | Skinny |
| 9 | Stocky | $\leftrightarrow$ | Willowy |
| 10 | Pear-Shaped | $\leftrightarrow$ | Apple-Shaped |
| 11 | Athletic | $\leftrightarrow$ | Non-Athletic |
| 12 | Fit | $\leftrightarrow$ | Unfit |
| 13 | Large-Breasted | $\leftrightarrow$ | Small-Breasted |
| 14 | Long-Armed | $\leftrightarrow$ | Short-Armed |
| 15 | Long-Necked | $\leftrightarrow$ | Short-Necked |
| 16 | High-Waisted | $\leftrightarrow$ | Low-Waisted |

Table 1. Paired phrases describing opposite body shape features.

### 3.3. Design 2: Dual Distillation from CLIP

We aim to incorporate rich prior knowledge from both CLIP teacher text and image encoders through distillation.

**Aligning Local Features to CLIP's Body Shape Descriptions.** Upon acquiring the image-specific body shape descriptors $\mathbf{L}$, along with their corresponding text features $\mathbf{H}$, we aim to capitalize on the domain-invariant properties inherent in these body shape descriptions and the features produced by CLIP teacher text encoder. To this end, we employ optimal transport (OT) [48], a method for measuring distances between two distributions [6, 30]. This approach is designed to steer the learning process of our student encoder by aligning the distributions of normalized local patch visual features with text features. Specifically, we align $\mathbf{H} = \{\mathbf{h}_m\}_{m=1}^M \in \mathbb{R}^{M\times d}$ and $\mathbf{F} = \{\mathbf{f}_n\}_{n=1}^N \in \mathbb{R}^{N\times d'}$. The underlying idea is to bring the student's learned representation closer to the teacher's domain-invariant representation, which is derived from the image-specific body shape

descriptors. To compare the distances between $\mathbf{H}$ and $\mathbf{F}$, we introduce a mapping network, $\Psi(\cdot)$, an MLP with a single hidden layer, to project the $d$-dim text features into the $d'$-dim space: $\mathbf{H}' = \Psi(\mathbf{H})$, where $\mathbf{H}' = \{\mathbf{h}'_m\}_{m=1}^{M} \in \mathbb{R}^{M \times d'}$. Formally, we define two discrete distributions as follows:

$$\mathbf{u} = \sum_{n=1}^{N} u_n \delta_{\mathbf{f}_n}, \quad \text{and} \quad \mathbf{v} = \sum_{m=1}^{M} v_m \delta_{\mathbf{h}'_m}. \tag{3}$$

Here, weights $\mathbf{u} = \{u_n\}_{n=1}^{N} \in \Delta_N$ and $\mathbf{v} = \{v_m\}_{m=1}^{M} \in \Delta_M$, with $\Delta$ representing the $N$- and $M$-dim probability simplices. This implies that $\sum_{n=1}^{N} u_n = 1$ and $\sum_{m=1}^{M} v_m = 1$. The terms $\delta_{\mathbf{f}_n}$ and $\delta_{\mathbf{h}'_m}$ denote Dirac delta functions located at the support points $\mathbf{f}$ and $\mathbf{h}'$ within their respective embedding spaces. The discrete OT distance for one training sample is defined as follows:

$$\langle \mathbf{T}, \mathbf{C} \rangle = \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbf{T}_{n,m} \mathbf{C}_{n,m}. \tag{4}$$

OT aims to transport $\mathbf{u}$ to $\mathbf{v}$ at the minimum cost, *i.e.*

$$\begin{aligned} d_{OT}(\mathbf{u}, \mathbf{v} | \mathbf{C}) &:= \min_{\mathbf{T} \in \prod(\mathbf{u}, \mathbf{v})} \langle \mathbf{T}, \mathbf{C} \rangle \\ s.t \quad \mathbf{T}\mathbf{1}_M &= \mathbf{u}, \quad \mathbf{T}\mathbf{1}_N = \mathbf{v}, \quad \mathbf{T} \in \mathbb{R}_{+}^{N \times M}, \end{aligned} \tag{5}$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product and $\mathbf{T}$ is the transport plan to be optimized. The cost matrix $\mathbf{C} \in \mathbb{R}^{n \times m}$ denotes the transportation cost between $\mathbf{f}_n$ and $\mathbf{h}'_m$, *e.g.* $\mathbf{C}_{n,m} = 1 - CS(\mathbf{f}_n, \mathbf{h}'_m)$. The set $\prod(\mathbf{u}, \mathbf{v})$ contains all joint probabilities of $\mathbf{u}$ and $\mathbf{v}$. Due to the high computational cost in solving the optimization problem of Eqn. 5, we adopt the Sinkhorn distance [11], following [6, 30], which apply an entropic constraint for more efficient optimization:

$$d_{OT,\lambda}(\mathbf{u}, \mathbf{v} | \mathbf{C}) := \min_{\mathbf{T} \in \prod(\mathbf{u}, \mathbf{v})} \langle \mathbf{T}, \mathbf{C} \rangle - \lambda h(\mathbf{T}), \tag{6}$$

where $\lambda > 0$ is the regularization weight and $h(\mathbf{T}) = \sum_{n,m} \mathbf{T}_{n,m} \log \mathbf{T}_{n,m}$ is the entropy of the transport plan $\mathbf{T}$. The optimized $\mathbf{T}^*$ can be obtained in a few iterations:

$$\mathbf{T}^* = \text{diag}(\mathbf{u}^{(t)}) \exp(-\mathbf{C}/\lambda) \text{diag}(\mathbf{v}^{(t)}), \tag{7}$$

where $t$ is the iteration step. $\mathbf{u}^{(t)}$ and $\mathbf{v}^{(t)}$ are updated according to the following rules: $\mathbf{u}^{(t)} = \mathbf{u}/\exp(-\mathbf{C}/\lambda \mathbf{v}^{(t-1)})$ and $\mathbf{v}^{(t)} = \mathbf{v}/\exp(-\mathbf{C}/\lambda \mathbf{u}^{(t-1)})$.

As illustrated in Fig. 3, we present visualization examples of transport plans. We can observe that the optimal transport plan aligns specific body regions with their corresponding linguistic descriptions, which demonstrates the aligment of local visual features in our student model with the linguistic descriptions.

**Aligning Global Features to CLIP Teacher Visual Component.** In our pursuit to distill semantic visual information from the CLIP teacher image encoder, we focus on preserving the global manifold structure within the latent space. This involves assessing the relationships, specifically the distances or similarities, between the global features of the student encoder $\mathbf{f}^{id}$ and the features $\mathbf{g}$ generated by the CLIP teacher image encoder, as outlined in Eqn. 1. To facilitate this alignment, we introduce a mapping network, $\Phi(\cdot)$, comprising a multilayer perceptron (MLP) with a single hidden layer. This network projects the teacher's features into a lower-dimensional space that aligns with our student encoder's feature dimensions. Subsequently, we apply an MSE loss function to guide the training of the student encoder. This approach ensures that the student's global features are coherently aligned with those of the teacher. The alignment process is mathematically represented as follows:

$$||\Phi(\mathbf{g}) - \mathbf{f}^{id}||_2^2. \tag{8}$$

This strategic alignment provides a robust framework for the student encoder, leveraging the rich, pre-trained knowledge embedded in the CLIP teacher's visual component.

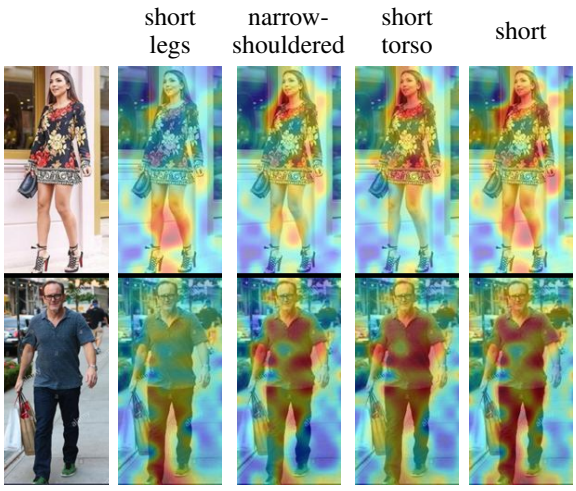


Figure 3. Visualization of transport plans: This figure overlays heatmaps, derived from transport plans $\mathbf{T}$ related to body descriptions, onto raw images. We use a blue-to-red colormap to represent the plans, indicating low to high values.

### 3.4. Design 3: 3D Reconstruction Regularization

In our approach, we introduce a unique regularization technique using the 3D Skinned Multi-Person Linear (SMPL) model [47] for our student model's global feature learning. This method is distinguished by **two novel traits**: First, it focuses on domain-invariant feature learning by reconstructing naked body shapes, deliberately omitting textural information to ensure the learned features are universally applicable across various domains. Second, it operates independently of ground-truth shape labels, utilizing synthetic mesh images rendered from the 3D body model without clothing or

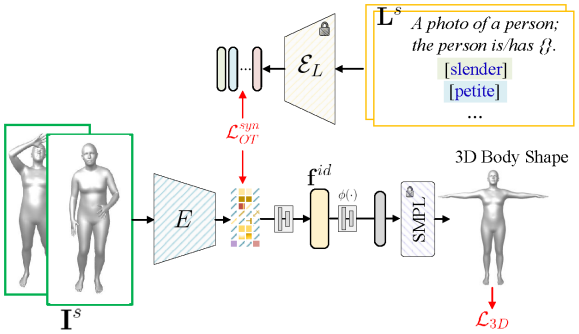


Figure 4. Illustration of regularization using 3D reconstruction with synthetic mesh images, each annotated with their corresponding linguistic body shape descriptors.

texture. This approach allows the student model to focus on the fundamental aspects of body shape, thereby enhancing its generalization across different datasets.

Specifically, our process begins with the generation of synthetic bodies using the identity component of the SMPL. We synthesize 500 females and 500 males in a neutral pose by randomly sampling the first 10 principal shape directions. Subsequently, we utilize an off-the-shelf method [49] to predict body pose parameters for 100 random human images from the training subset of the Celeb-reID dataset [25], thus creating a pool of 3D body poses. For each of the $1,000$ synthetic subjects, we randomly select 10 poses from this pool to render body mesh images $\mathbf{I}^s$ with varied body poses, resulting in a total of $10,000$ synthetic body mesh images. Leveraging the shape-to-attribute framework in [49], we generate true linguistic body shape descriptions for each synthetic image, eliminating the need for employing the CLIP to determine the descriptions $\mathbf{L}^s$. Consequently, each mini-batch of $B$ synthetic training samples is represented as a tuple of training sets $\{(\mathbf{I}_i^s, \mathbf{L}_i^s, \mathbf{P}_i)\}_{i=1}^B$, where $\mathbf{P}$ is the ground-truth canonical identity shape.

For the final 3D regularization loss, it is formulated as:

$$\mathcal{L}_{3D-Regu} = \mathcal{L}_{OT}^{syn} + \mathcal{L}_{3D}, \tag{9}$$

where $\mathcal{L}_{OT}^{syn} = \frac{1}{B}\sum_{i=1}^{B} d_{OT}$ (Eqn. 5). The $\mathcal{L}_{3D}$ is defined as the squared Euclidean distance between the predicted body shape, derived from the identity features $\mathbf{f}^{id}$, and the ground truth, $\mathcal{L}_{3D} = \sum_{i=1}^{B} ||SMPL(\phi(\mathbf{f}_i^{id})) - \mathbf{P}_i||_2^2$, with $\phi(\cdot)$ denoting a single-hidden-layer MLP that maps identity features to the SMPL body shape latent space (Fig. 4).

### 3.5. Overall Training Objective

The overall training loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{OT} + \beta\mathcal{L}_{global} + \gamma\mathcal{L}_{3D-Regu}, \tag{10}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss on $\mathbf{f}^{id}$, $\mathcal{L}_{OT} = \frac{1}{B}\sum_{i=1}^{B} d_{OT}$, $\mathcal{L}_{global} = \frac{1}{B}\sum_{i=1}^{B} ||\Phi(\mathbf{g}_i) - \mathbf{f}_i^{id}||_2^2$. The parameters $\alpha$, $\beta$, and $\gamma$ are weights assigned to balance the loss terms. *For practical implementation, in our training, we integrate both synthetic and real images within each training batch.*

| Method | Celeb-reID | | Celeb-reID-light | |
|---|---|---|---|---|
| | mAP | Rank1 | mAP | Rank1 |
| ReIDCaps (TCSVT20) [26] | 15.8 | 63.0 | 19.0 | 33.5 |
| RCSAnet (ICCV21) [27] | 11.9 | 55.6 | 16.7 | 29.5 |
| CASE-Net (WACV21) [39] | 18.2 | **66.4** | 20.4 | 35.1 |
| CAL (CVPR22) [20] | 13.7 | 59.2 | 18.5 | 33.6 |
| 3DInvarReID (ICCV23) [45] | 15.2 | 61.2 | 21.8 | 37.0 |
| **CLIP3DReID** | **19.2** | 63.1 | **26.3** | **39.4** |

Table 2. Comparison on Celeb-reID and Celeb-reID-light datasets.

| Method | Celeb-reID (blur) | | Celeb-reID-light (blur) | |
|---|---|---|---|---|
| | mAP | Rank1 | mAP | Rank1 |
| CAL [20] | 7.7 | 48.2 | 13.4 | 22.5 |
| 3DInvarReID [45] | 9.6 | 51.2 | 17.2 | 29.6 |
| **CLIP3DReID** | **11.6** | **52.8** | **21.3** | **32.1** |

Table 3. Comparison on face-blurred versions of Celeb-reID and Celeb-reID-light datasets.

**Implementation Details.** We employ the ViT-L/14 trained by CLIP [56] as our teacher model and ResNet-50 [21] as our student model. We implement in Pytorch, use Adam optimizer, and set $t$=1, $\lambda$=1, $\alpha$=0.3, $\beta$=0.5, $\gamma$=0.3.

## 4. Experiment

### 4.1. Person ReID

In the evaluation, we utilize the standard retrieval metrics: the Cumulative Matching Characteristics (CMC) and mean Average Precision (mAP).

#### 4.1.1 Results on Cloth-changing Person ReID datasets

**Dataset and Baseline.** We evaluate on six popular cloth-changing ReID datasets: Celeb-reID/Celeb-reID-light [25, 26], PRCC [68], LTCC [60] and the recent CCVID [20, 75], DeepChange [67] and CCDA [45] dataset. We compare with eight state-of-the-art (SoTA) cloth-changing ReID works: ReIDCaps [26], 3DSL [7], RCSAnet [27], FSAM [24], CAL [20], AIM [70] and 3DInvarReID [45].

**Results on Celeb-reID and Celeb-reID-light.** Tab. 2 shows results on the Celeb-reID and Celeb-reID-light datasets. Our **CLIP3DReID** significantly surpasses all SoTA baselines. For instance, on Celeb-reID-light, our model elevates the mAP from 21.8, achieved by the best baseline, to 26.3. These results clearly indicate that the linguistic body shape features in our method are highly discriminative. This enhancement in performance underlines the effectiveness of our proposed approach in person ReID.

To investigate the impact of facial anonymization on model performance, we employed a face detection [12] cou-

| Method | LTCC | | PRCC | |
|---|---|---|---|---|
| | mAP | Rank1 | mAP | Rank1 |
| 3DSL (CVPR21) [7] | 14.8 | 31.2 | – | 51.3 |
| FSAM (CVPR21) [24] | 16.2 | 38.5 | – | 54.5 |
| CAL (CVPR22) [20] | 18.0 | 40.1 | 55.8 | 55.2 |
| AIM (CVPR23) [70] | 19.1 | 40.6 | 58.3 | 57.9 |
| 3DInvarReID (ICCV23) [45] | 18.9 | 40.9 | 57.2 | 56.5 |
| **CLIP3DReID** | **21.7** | **42.1** | **59.3** | **60.6** |

Table 4. Comparison on the LTCC and PRCC datasets, with focus on the cloth-changing protocol.

| Method | CCVID | | | | CCDA | |
|---|---|---|---|---|---|---|
| | General | | Cloth-changing | | | |
| | mAP | Rank1 | mAP | Rank1 | mAP | Rank1 |
| CAL [20] | 81.3 | 82.6 | 79.6 | 81.7 | 19.3 | 10.0 |
| 3DInvarReID [45] | 82.6 | 83.9 | 81.3 | 81.7 | 21.7 | 11.1 |
| **CLIP3DReID** | **83.9** | **84.5** | **83.2** | **82.4** | **25.7** | **15.5** |

Table 5. Comparison on CCVID and CCDA datasets.

| Method | mAP | Rank1 |
|---|---|---|
| ReIDCaps (TCSVT20) [26] | 11.3 | 39.5 |
| CAL (CVPR22) [20] | 19.0 | 54.0 |
| 3DInvarReID (ICCV23) [45] | 19.6 | 55.1 |
| **CLIP3DReID** | **20.8** | **56.7** |

Table 6. Comparison on the DeepChange datasets.

pled with a Gaussian blurring on facial regions to create anonymized blur versions of the Celeb-reID and Celeb-reID-light datasets. This process involves retraining the baselines CAL [20] and 3DInvarReID [45], alongside our own models, on these altered datasets. According to Tab. 3, our method consistently surpasses these two baselines. This finding underscores the resilience and efficacy of our approach, even in scenarios where facial details are obscured for anonymity.

**Results on LTCC and PRCC.** The comparative results on the LTCC and PRCC datasets are presented in Tab. 4. Consistently, our method sets a new SoTA performance. For example, on the LTCC dataset, it surpasses the 3DInvarReID [70] by 1.5% in mAP and 2.6% in Rank-1 accuracy. Similarly, on PRCC, our method shows improvements of 1.0% and 2.7% in mAP and Rank-1 accuracy, respectively. We observed limited improvement in the LTCC and PRCC datasets. This difference in performance is largely due to the lower image resolution in these datasets compared to others. The reduced resolution diminishes the semantic correlation effectiveness in the pretrained CLIP model, which could impact the efficiency and accuracy of our linguistic body shape labeling process.

**Results on CCVID, DeepChange and CCDA.** We further evaluate our method on recent cloth-changing datasets, including CCVID, DeepChange, and CCDA. Following the same protocol as [45], the model evaluated on CCDA is trained on the Celeb-reID dataset. As illustrated in Tabs. 5 and 6, our CLIP3DReID model demonstrates superior performance over baselines. Notably, on the most recent CCDA dataset, characterized by its diversity in human activities, our model shows significant improvements, with a 4.0% increase in mAP and a 4.4% enhancement in Rank1 accuracy, surpassing the 3DInvarReID model, which also employs 3D shape feature extraction.

| Method | Market-1501 | | MSMT17 | |
|---|---|---|---|---|
| | mAP | Rank1 | mAP | Rank1 |
| 3DSL (CVPR21) [7] | 87.3 | 95.0 | – | – |
| FSAM (CVPR21) [24] | 85.6 | 94.6 | – | – |
| CAL (CVPR22) [20] | 87.5 | 94.7 | 57.3 | 79.3 |
| 3DInvarReID (ICCV23) [45] | 87.9 | 95.1 | 59.1 | 80.8 |
| **CLIP3DReID** | **88.4** | **95.6** | **61.2** | **81.5** |

Table 7. Comparison on the short-term ReID datasets: Market-1501 and MSMT17 datasets.

| Method | LTCC → PRCC | | PRCC → LTCC | |
|---|---|---|---|---|
| | mAP | Rank1 | mAP | Rank1 |
| CAL [20] | 35.9 | 38.0 | 3.3 | 8.4 |
| 3DInvarReID [45] | 36.1 | 40.1 | 3.3 | 9.1 |
| **CLIP3DReID** | **37.5** | **41.7** | **4.5** | **9.9** |

Table 8. Cross-dataset comparison on the cloth-changing ReID datasets: LTCC and PRCC datasets.

#### 4.1.2 Results on Short-term Person ReID datasets

While our method is primarily tailored for long-term scenarios, we also conduct comparisons on two conventional short-term ReID datasets: Market-1501 [76] and MSMT17 [66], as detailed in Tab. 7. Our evaluations on the MSMT17 dataset indicate an enhancement of 0.7% in Rank-1 accuracy and 2.1% in mAP. These results highlight the beneficial impact of our linguistic shape feature, demonstrating its applicability and value even in short-term ReID scenarios.

#### 4.1.3 Results on the Cross-Dataset Setting

Additionally, we evaluate in a ***cross-dataset setting***, by training on one dataset and testing on another. Specifically, we test on two cloth-changing datasets: LTCC [60] and PRCC [68]. The results in Tab. 8 reveal that **CLIP3DReID** consistently surpasses the baselines, further affirming its effectiveness in diverse dataset scenarios. We refrain from comparing ours with domain-invariant person ReID methods, as they typically utilize information from *both* datasets in training. For instance, the study in [38] employs a technique to transfer style from the target to the source dataset.

### 4.2. 3D Body Shape Reconstruction

Consistent with the methodologies described in [10] and [45], we assess our 3D body shape reconstruction performance using the Human Body in the Wild (HBW) dataset. This dataset comprises 237 in-the-wild images of 10 subjects,

Figure 5. Examples of our 3D body reconstruction. For enhanced shape visualization, we have applied body poses estimated by an established off-the-shelf method [49] for each reconstructed shape.

each paired with ground-truth 3D body scans, providing a robust framework for evaluating identity-specific body shapes. Following [45], we use the Chamfer Distance (CD-$L_2$) as the metric. This involves uniformly sampling $10,000$ points on both the ground-truth and predicted meshes in the canonical space. As shown in Tab. 9, our model achieves comparable performance to two established baselines, despite not utilizing any real 3D body shapes in our training. This achievement underlines two important points: first, it validates the effective integration of our linguistic body shape features with actual 3D body shapes; and second, it confirms the efficacy of our framework, which is designed to utilize solely synthetic data for training, in accurately reconstructing 3D identity body shapes from real images. Fig. 5 presents qualitative results of our reconstruction, demonstrating promising performance, which indicates that the feature $\mathbf{f}^{id}$ we learned is meaningful and effective in the 3D body shape space.

### 4.3. Ablation Study

All models in ablation are trained on the Celeb-reID dataset.

**Contribution of Synthetic Mesh Images.** We train a model without using synthetic mesh images and find a significant performance drop w.r.t. our standard model (Tab. 10). This highlights the pivotal role of synthetic mesh images in improving linguistic body descriptions.

**Contribution of Loss Terms.** Tab. 10 systematically evaluates the performance impact of various loss terms in CLIP3DReID, including visual-text feature alignment $\mathcal{L}_{OT}$, global feature alignment $\mathcal{L}_{global}$, and 3D reconstruction regularization ($\mathcal{L}_{3D-Regu}$), on Celeb-reID and Celeb-reID-light datasets. **1)** When the $\mathcal{L}_{OT}$ is omitted, we observe a notable decrease in performance, with the mAP dropping to $17.0\%$ and $24.8\%$ on Celeb-reID. This underscores the importance of $\mathcal{L}_{OT}$ in capturing fine-grained local features that are critical for accurate person ReID. **2)** The absence of the global feature alignment ($\mathcal{L}_{global}$ ) also leads to performance drop, albeit to a lesser extent. This indicates that while $\mathcal{L}_{global}$ contributes to the model's overall performance, the local features captured by $\mathcal{L}_{OT}$ have a more pronounced impact on the accuracy of ReID. **3)** Using 3D reconstruction regularization ($\mathcal{L}_{3D-Regu}$) improves model performance, highlighting its effectiveness in discriminative body feature learning.

| | SHAPY [10] | 3DInvarReID [45] | **CLIP3DReID** |
|---|---|---|---|
| CD-$L_2$ | 0.632 | **0.610** | 0.642 |

Table 9. Comparison of 3D identity shape reconstruction on HBW.

| Method | Celeb-reID | | Celeb-reID-light | |
|---|---|---|---|---|
| | mAP | Rank1 | mAP | Rank1 |
| w/o Synthetic mesh images | 17.9 | 61.8 | 25.3 | 38.7 |
| w/o $\mathcal{L}_{OT}$ | 17.0 | 61.2 | 24.8 | 38.1 |
| w/o $\mathcal{L}_{global}$ | 16.5 | 60.1 | 23.9 | 37.6 |
| w/o $\mathcal{L}_{3D-Regu}$ | 18.5 | 62.5 | 26.0 | 38.9 |
| w/o $\mathcal{L}_{OT}$, $\mathcal{L}_{global}$, $\mathcal{L}_{3D-Regu}$ | 14.2 | 57.5 | 17.4 | 29.1 |
| **CLIP3DReID** (Ours) | **19.2** | **63.1** | **26.3** | **39.4** |

Table 10. Ablation studies on varied loss function configurations.

| Method | CLIP Teacher | Celeb-reID | |
|---|---|---|---|
| | | mAP | Rank1 |
| Model-1 | ViT-B/32 | 15.2 | 58.7 |
| Model-2 | ViT-B/16 | 17.7 | 61.2 |
| **CLIP3DReID** (Ours) | ViT-L/14 | **19.2** | **63.1** |

Table 11. Ablation studies on the architectural variations of the CLIP-Based teacher and our student models.

**Effect of Different Teacher Models.** We also ablate on the architectural variations of the CLIP-based teacher model, as summarized in Tab. 11. Model-1 employs the smaller ViT-B/32 and sees a slight dip in performance. Model-2, with the mid-sized ViT-B/16, improves upon this outcome. CLIP3DReID utilizes the ViT-L/14 teacher model for its broad feature extraction capacity, which surpasses other configurations in performance.

## 5. Conclusion

In this paper, we introduce **CLIP3DReID**, a new approach to person ReID, harnessing the power of knowledge distillation from the CLIP model. Our innovative integration of vision-language models and 3D body shape understanding significantly enhances both accuracy and robustness in ReID systems. The successful alignment of visual features with linguistic descriptions and the novel use of the 3D SMPL model as a non-reliant tool on ground-truth shape labels are key highlights of our approach. The superior performance of **CLIP3DReID** on multiple datasets underscores its potential for diverse real-world applications.

**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] Ejaz Ahmed, Michael Jones, and Tim K Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 1

[2] Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. Rasa: Relation and sensitivity aware representation learning for text-based person search. In *IJCAI*, 2023. 3

[3] Yang Bai, Jingyao Wang, Min Cao, Chen Chen, Ziqiang Cao, Liqiang Nie, and Min Zhang. Text-based person search without parallel image-text data. In *ACMMM*, 2023. 3

[4] Apurva Bedagkar-Gala and Shishir K Shah. A survey of approaches and trends in person re-identification. *Image and vision computing*, 2014. 1

[5] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of CLIP for text-based person search. In *AAAI*, 2024. 3

[6] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. In *ICLR*, 2023. 3, 4, 5

[7] 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, 6, 7

[8] Jun Chen, Deyao Zhu, Guocheng Qian, Bernard Ghanem, Zhicheng Yan, Chenchen Zhu, Fanyi Xiao, Sean Chang Culatana, and Mohamed Elhoseiny. Exploring open-vocabulary semantic segmentation from clip vision encoder distillation only. In *ICCV*, 2023. 3

[9] Zhong Chen, Zhizhong Zhang, Xin Tan, Yanyun Qu, and Yuan Xie. Unveiling the power of CLIP in unsupervised visible-infrared person re-identification. In *ACMMM*, 2023. 2

[10] Vasileios Choutas, Lea Müller, Chun-Hao P Huang, Siyu Tang, Dimitrios Tzionas, and Michael J Black. Accurate 3D body shape regression using metric and semantic attributes. In *CVPR*, 2022. 3, 4, 7, 8

[11] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013. 5

[12] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. In *CVPR*, 2020. 6

[13] Xing Di, Yiyu Zheng, Xiaoming Liu, and Yu Cheng. ProS: Facial omni-representation learning via prototype-based self-distillation. In *WACV*, 2024. 3

[14] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model CLIP. In *AAAI*, 2022. 3

[15] Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, and Dina Katabi. Learning longterm representations for person re-identification using radio signals. In *CVPR*, 2020. 2

[16] Takuro Fujii and Shuhei Tarashima. Bilma: Bidirectional local-matching for text-based person re-identification. In *ICCVW*, 2023. 2

[17] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 1, 2

[18] 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. 1, 2

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

[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

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 6

[22] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, 2021. 1

[23] Matthew Q Hill, Stephan Streuber, Carina A Hahn, Michael J Black, and Alice J O'Toole. Creating body shapes from verbal descriptions by linking similarity spaces. *Psychological science*, 2016. 3, 4

[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, 6, 7

[25] 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. 1, 2, 6

[26] 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. 1, 2, 6, 7

[27] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing status awareness for long-term person re-identification. In *ICCV*, 2021. 1, 6

[28] Zeyi Huang, Andy Zhou, Zijian Ling, Mu Cai, Haohan Wang, and Yong Jae Lee. A sentence speaks a thousand images: Domain generalization through distilling CLIP with language guidance. In *ICCV*, 2023. 3

[29] 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

[30] Yiming Lei, Zilong Li, Yangyang Li, Junping Zhang, and Hongming Shan. Lico: Explainable models with language-image consistency. In *NeurIPS*, 2023. 3, 4, 5

[31] Huafeng Li, Shedan Yang, Yafei Zhang, Dapeng Tao, and Zhengtao Yu. Progressive feature mining and external knowledge-assisted text-pedestrian image retrieval. *arXiv preprint arXiv:2308.11994*, 2023. 3

[32] Jiachen Li and Xiaojin Gong. Prototypical contrastive learning-based CLIP fine-tuning for object re-identification. *arXiv preprint arXiv:2310.17218*, 2023. 3

[33] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, 2018. 1, 2

[34] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised tracklet person re-identification. *PAMI*, 2019. 1, 2

[35] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, 2017. 3

[36] Shuang Li, Jiaxu Leng, Ji Gan, Mengjingcheng Mo, and Xinbo Gao. Shape-centered representation learning for visible-infrared person re-identification. *arXiv preprint arXiv:2310.17952*, 2023. 3

[37] Siyuan Li, Li Sun, and Qingli Li. CLIP-ReID: exploiting vision-language model for image re-identification without concrete text labels. In *AAAI*, 2023. 2

[38] Yu-Jhe Li, Ci-Siang Lin, Yan-Bo Lin, and Yu-Chiang Frank Wang. Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In *ICCV*, 2019. 7

[39] Yu-Jhe Li, Xinshuo Weng, and Kris M Kitani. Learning shape representations for person re-identification under clothing change. In *WACV*, 2021. 1, 2, 6

[40] 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

[41] Yin Lin, Cong Liu, Yehansen Chen, Jinshui Hu, Bing Yin, Baocai Yin, and Zengfu Wang. Exploring part-informed visual-language learning for person re-identification. *arXiv preprint arXiv:2308.02738*, 2023. 3

[42] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *ECCV*, 2012. 1

[43] Delong Liu and Haiwen Li. Unleashing the imagination of text: A novel framework for text-to-image person retrieval via exploring the power of words. *arXiv preprint arXiv:2307.09059*, 2023. 3

[44] Feng Liu, Ronghang Zhu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Disentangling features in 3d face shapes for joint face reconstruction and recognition. In *CVPR*, 2018. 1

[45] 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. 2, 6, 7, 8

[46] Feng Liu, Ryan Ashbaugh, Nicholas Chimitt, Najmul Hassan, Ali Hassani, Ajay Jaiswal, Minchul Kim, Zhiyuan Mao, Christopher Perry, Zhiyuan Ren, Yiyang Su, Pegah Varghaei, Kai Wang, Xingguang Zhang, Stanley Chan, Arun Ross, Humphrey Shi, Zhangyang Wang, Anil Jain, and Xiaoming Liu. FarSight: A physics-driven whole-body biometric system at large distance and altitude. In *WACV*, 2024. 1

[47] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *TOG*, 2015. 5

[48] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, 1781. 4

[49] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3D hand pose estimation for whole-body 3d human mesh estimation. In *CVPR*, 2022. 6, 8

[50] Blake A Myers, Lucas Jaggernauth, Thomas M Metz, Matthew Q Hill, Veda Nandan Gandi, Carlos D Castillo, and Alice J O'Toole. Recognizing people by body shape using deep networks of images and words. In *IJCB*, 2023. 3, 4

[51] Kien Nguyen, Clinton Fookes, Sridha Sridharan, Feng Liu, Xiaoming Liu, Arun Ross, Dana Michalski, Huy Nguyen, Debayan Deb, Mahak Kothari, Manisha Saini, Dawei Du, Scott McCloskey, Gabriel Bertocco, Fernanda Andal´o, Terrance E. Boult, Anderson Rocha, Haidong Zhu, Zhaoheng Zheng, Ram Nevatia, Zaigham Randhawa, Sinan Sabri, and Gianfranco Doretto. Ag-reid 2023: Aerial-ground person re-identification challenge results. In *IJCB*, 2023. 1

[52] Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *CVPR*, 2023. 3

[53] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *ACCV*, 2020. 2

[54] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *CVPR*, 2024. 3

[55] María Alejandra Quirós-Ramírez, Stephan Streuber, and Michael J Black. Red shape, blue shape: political ideology influences the social perception of body shape. *Humanities and Social Sciences Communications*, 2021. 3, 4

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 6

[57] Zhiyuan Ren, Yiyang Su, and Xiaoming Liu. ChatGPT-powered hierarchical comparisons for image classification. In *NeurIPS*, 2023. 3

[58] Joseph Roth and Xiaoming Liu. On the exploration of joint attribute learning for person re-identification. In *ACCV*, 2014. 3

[59] Fei Shen, Xiangbo Shu, Xiaoyu Du, and Jinhui Tang. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In *ACMMM*, 2023. 3

[60] 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

[61] Stephan Streuber, M Alejandra Quiros-Ramirez, Matthew Q Hill, Carina A Hahn, Silvia Zuffi, Alice O'Toole, and Michael J Black. Body talk: Crowdshaping realistic 3d avatars with words. *TOG*, 2016. 3, 4

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

[63] Guanshuo Wang, Fufu Yu, Junjie Li, Qiong Jia, and Shouhong Ding. Exploiting the textual potential from vision-language pre-training for text-based person search. *arXiv preprint arXiv:2303.04497*, 2023. 3

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

[65] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: CLIP-driven referring image segmentation. In *CVPR*, 2022. 3

[66] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer GAN to bridge domain gap for person re-identification. In *CVPR*, 2018. 7

[67] Peng Xu and Xiatian Zhu. DeepChange: A large long-term person re-identification benchmark with clothes change. In *ICCV*, 2023. 2, 6

[68] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *PAMI*, 2019. 1, 2, 6, 7

[69] Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *ACMMM*, 2023. 3

[70] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *CVPR*, 2023. 6, 7

[71] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *TPAMI*, 2021. 1

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

[73] Shijie Yu, Shihua Li, Dapeng Chen, Rui Zhao, Junjie Yan, and Yu Qiao. COCAS: A large-scale clothes changing person dataset for re-identification. In *CVPR*, 2020. 1, 2

[74] 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

[75] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *CVPR*, 2019. 6

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

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

[78] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, 2019. 1, 2

[79] Zhedong Zheng, Nenggan Zheng, and Yi Yang. Parameter-efficient person re-identification in the 3D space. *TNNLS*, 2022. 2

[80] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, 2019. 1

[81] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 3

[82] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 3

[83] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting CLIP for zero-shot semantic segmentation. In *CVPR*, 2023. 3