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

Supplementary Material

In this supplementary material, we provide:

◇ Additional results include a deeper exploration of the CLIP teacher image encoder's role in our model's performance, further analysis of computational efficiency, and labeled linguistic body shape descriptors.

◇ Additional implementation details.

◇ Detailed information on the person ReID datasets evaluated.

1. Additional Results

1.1. Understanding the CLIP Teacher Image Encoder's Role

Our exploration extends to assessing the potential of directly employing the pre-trained and frozen CLIP image encoder (ViT-L/14) for person Re-Identification (ReID). This inquiry is rooted in understanding the encoder's effectiveness as a teacher in varied ReID contexts. We conduct evaluations across diverse datasets: Celeb-reID [2], its variant Celeb-reID (blur), LTCC [4], and DeepChange [6]. A unique model is developed by freezing the backbone of the CLIP image encoder and appending a fine-tunable fully connected (FC) layer, termed as "Teacher image encoder `Linear`".

Surprisingly, as can be observed in Tab. 1, the pre-trained encoder demonstrates notable effectiveness on the Celeb-reID dataset, with significant performance gains post fine-tuning (Rank1 improvement from 66.0 → 74.8). This impressive outcome led us to hypothesize two contributing factors: 1) *Data Familiarity*: Celeb-reID's compilation of street-collected celebrity images from the web likely intersects with the CLIP training dataset. This overlap, possibly containing named celebrity images, might have provided pre-aligned training advantages. 2) *Facial Feature Recognition*: The dataset's high-quality facial regions could have been a focal point for CLIP's learning, suggesting a dependency on clear facial features.

To test these hypotheses, we evaluate with the Celeb-reID (blur) dataset, where facial regions were deliberately blurred. The pre-trained CLIP model's performance significantly drops (Rank1 dropping from 66.0 to 42.9), and the fine-tuned "Teacher image encoder `Linear`" model only marginally outperforms our student model (53.0 vs. 52.8 Rank1 accuracy). However, on ***real-world captured datasets*** like LTCC and DeepChange, which lack intersections with the CLIP training data, both the pre-trained and fine-tuned CLIP teacher encoders exhibit underwhelming performance (Tab. 1). This indicates the CLIP image encoder's limitations in handling low-quality images and environments dissimilar to its training data. Fig. 1 clearly demonstrates the differences in image quality among these four datasets.

These findings reinforce the necessity of dual guidance for learning discriminative feature representation. It becomes clear that distilling information from the CLIP model is only part of the solution; supplementing it with custom-tailored representations is crucial to address the diverse, real-world variations in human shapes and appearances.

1.2. Computational Efficiency

We conduct a comprehensive evaluation of computational efficiency among the considered models, including the established baselines CAL [1] and 3DInvarReID [3] which utilize the ResNet-50 backbone. To quantify the computational demand, we report the number of parameters and measure inference efficiency using two industry-standard metrics: MACs and FLOPs. FLOPs (Floating Point Operations per Second) are an essential metric for evaluating computational complexity, quantifying the speed of arithmetic operations, including additions and multiplications involving floating-point numbers. MACs (Multiply-Accumulates) are specifically concerned with multiply-accumulate operations, which are ubiquitous in deep learning computations.

In assessing the various CLIP teacher image encoder architectures—ViT-B/16, ViT-B/32, and ViT-L/14—we contrast them with our student encoders ResNet-50. When examining the results on the Celeb-reID (blur) dataset, as detailed in Tab. 2, we observe the following: 1) ResNet-50 emerges as our student encoder of choice due to its optimal balance of reduced parameters alongside lower MACs and FLOPs. This efficiency does not come at the cost of performance; our student encoder attains new state-of-the-art results, outpacing the CAL [1] and 3DInvarReID [3] baselines. 2) The "Teacher image encoder `Linear`" (ViT-L/14), while achieving marginally higher performance on Celeb-reID (blur), demands significantly more computational resources, at 162.12 GFLOPs, compared to the 8.11 GFLOPs required by our model.

Our analysis suggests that the integration of linguistic body shape representations within the student ResNet-50 architecture leads to a favorable trade-off between computational efficiency and the precision of person ReID. This synergy paves the way for deployment in scenarios where computational resources are at a premium, without sacrificing the quality of ReID tasks.

Model	Celeb-reID		Celeb-reID (blur)		LTCC		DeepChange	
	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1
CAL (CVPR22) [1]	13.7	59.2	7.7	48.2	18.0	40.1	19.0	54.0
3DInvarReID (ICCV23) [3]	15.2	61.2	9.6	51.2	18.9	40.9	19.6	55.1
Teacher image encoder	24.1	66.0	8.5	42.9	5.3	14.0	2.9	20.1
Teacher image encoder `Linear`	**37.3**	**74.8**	**14.2**	**53.0**	12.7	31.6	10.5	40.5
Student encoder **CLIP3DReID**	19.2	63.1	11.6	52.8	**21.7**	**42.1**	**20.8**	**56.7**

Table 1. Analysis of the performance of the pre-trained and fine-tuned CLIP teacher image encoder, 'Teacher image encoder `Linear`,' across different datasets, including Celeb-reID [2], its blurred variant Celeb-reID (blur), LTCC [4], and DeepChange [6].

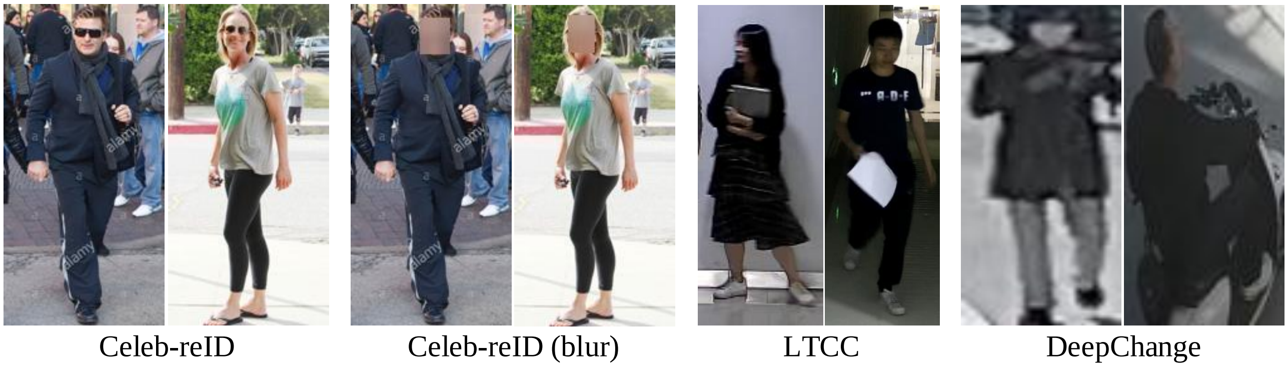

Figure 1. Image examples from Celeb-reID [2], Celeb-reID (blur), LTCC [4], and DeepChange [6] datasets, illustrating the range of image quality.

1.3. Labeled Linguistic Body Shape Descriptors

Fig. 2 provides examples of linguistic body shape descriptors automatically labeled using the pre-trained CLIP model. These illustrations highlight the model's capability in accurately discerning and representing various human body shapes.

2. Additional Implementation Details

In our experimental setup, the CLIP model, in comparison, utilizes a feature space with a higher dimensionality of $d = 768$. To bridge the feature space dimensions, we project the d-dimensional text features into the d'-dimensional space: $\mathbf{H}' = \Psi(\mathbf{H})$. Here, Ψ is a single-layer fully connected (FC) network without an activation function. A similar approach is taken with $\Phi(\cdot)$, which maps the CLIP image features $\mathbf{g} \in \mathbb{R}^d$ to a transformed feature space $\mathbf{f}^{id} \in \mathbb{R}^{d'}$ through a single-layer FC network, also lacking an activation function. Additionally, to align $\mathbf{f}_{id}$ with the 10-dimensional identity space of the SMPL model, we employ a transformation function ϕ, once again realized as a single-layer FC network without an activation function. This consistent approach in the network architecture facilitates the alignment of multi-dimensional features across different modalities, crucial for the integrity and coherence of our representation space.

3. Dataset Information

We provide additional information about the evaluated datasets in the paper.

CCDA [3]. CCDA is a new dataset that encompasses a diverse range of human activities and clothing changes for long-term person ReID evaluation. The dataset specifically includes data from prominent athletes in sports like soccer, tennis, and basketball, as well as well-known figures in the arts, such as fashion models and singers. The CCDA dataset comprises a total of $1,555$ images featuring 100 subjects.

Celeb-reID/Celeb-reID-light [2] . Celeb-reID consists of $34,186$ images of $1,052$ celebrities that are crawled street snap-shots from the web. Data is divided into two parts: 632 subjects with $20,208$ images for training and 420 subjects with $13,978$ images for testing. Among the test images, $2,972$ is used as a query, and $11,006$ is used as a gallery. **Celeb-reID-light** is a light version of Celeb-reID. Unlike Celeb-reID, subjects in Celeb-reID-light do not wear the same clothing twice. It consists of $10,842$ images of 590 subjects: $9,021$ images of 490 subjects for training and $1,821$ images of 100 subjects for testing. In testing, 934 images are in the gallery and 887 images serve as queries.

Model	Backbone	Input Resolution	Params(M) ↓	MACs(G) ↓	FLOPs(G) ↓	mAP ↑	Rank1 ↑
CAL (CVPR22) [1]	ResNet-50	384 × 192	23.52	9.12	18.30	7.7	48.2
3DInvarReID (ICCV23) [3]	ResNet-50	384 × 192	23.52	9.12	18.30	9.6	51.2
Teacher image encoder	ViT-B/16	224 × 224	151.28	4.41	8.82	1.4	14.1
	ViT-B/32	224 × 224	149.62	17.56	35.15	2.4	22.2
	ViT-L/14	224 × 224	427.62	81.01	162.11	2.5	24.1
Teacher image encoder `Linear`	ViT-B/16	224 × 224	151.57	4.41	8.82	10.4	50.3
	ViT-B/32	224 × 224	149.92	17.56	35.15	9.1	49.8
	ViT-L/14	224 × 224	428.06	81.01	162.12	**14.2**	**53.0**
Student encoder **CLIP3DReID**	ResNet-50	384 × 192	25.87	9.12	18.30	11.6	52.8

Table 2. Comparative evaluation of model efficiencies, detailing parameter counts, MACs, and FLOPs for varying architectures on the Celeb-reID (blur) dataset.

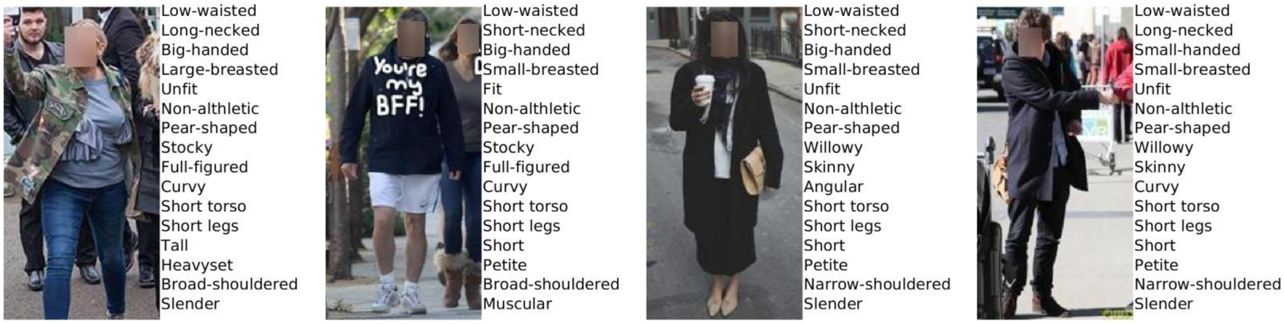

Figure 2. Examples of linguistic body shape descriptors automatically generated by the pre-trained CLIP model, showcasing its ability to capture discriminative body shape representations.

LTCC [4] . LTCC dataset is focused on indoor clothes-changing re-identification (re-ID). It comprises 17,138 images of 152 individuals wearing 478 different outfits, which were captured by 12 cameras over a two-month period. On average, each person has five clothing outfits, with the number of outfit changes ranging from 2 to 14.

PRCC [7]. PRCC dataset comprises 33,698 real-scenario images of 221 individuals captured from three different camera views. When the same person appears under Camera A and Camera B, they wear the same clothes, while the same person appearing under Camera A and Camera C wears different clothes.

CCVID [1] . CCVID dataset is a new clothes-changing *video* person-reID dataset which is made from the raw data of the gait recognition dataset FVG [8]. FVG includes 2856 videos of 226 subjects and each subject has 2∼5 different sets of clothing. The re-purposed CCVID contains $347,833$ bounding boxes of 226 subjects. It is split into two parts: 75 subjects for training, and 151 subjects for testing. In the test set, 834 videos are in the query set and $1,074$ videos are in the gallery.

DeepChange [6] . DeepChange dataset sets a new benchmark for long-term person ReID. It is distinguished by its unprecedented scale, featuring the largest collection to date with 17 cameras, encompassing $1,121$ identities, and a total of $178,407$ bounding boxes. Remarkably, this dataset spans over 12 months, capturing a wide array of clothing changes, thereby offering unparalleled diversity and depth for person ReID research.

Market-1501 [9]. The Market-1501 dataset contains 32,668 annotated images of 1,501 individuals captured by six cameras in an outdoor market. The dataset includes challenging variations in pose, illumination, occlusion, and background. It is a widely used short-term person ReID dataset.

MSMT17 [5]. MSMT17 dataset contains 126,441 images of 4,112 identities captured by 15 cameras. The images were collected from a university campus over a period of four years. This short-term person ReID dataset includes a wide range of variations in terms of lighting conditions, camera viewpoints, pose, and occlusion.

References

[1] 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, 3

[2] 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

[3] 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, 3

[4] 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. 1, 2, 3

[5] 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. 3

[6] Peng Xu and Xiatian Zhu. DeepChange: A large long-term person re-identification benchmark with clothes change. In *ICCV*, 2023. 1, 2, 3

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

[8] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *CVPR*, 2019. 3

[9] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 3