SEAS: ShapE-Aligned Supervision for Person Re-Identification

Haidong Zhu Pranav Budhwant Zhaoheng Zheng Ram Nevatia
University of Southern California
{haodongz, budhwant, zhaoheng.zheng, nevatia}@usc.edu

Abstract

*We introduce SEAS, using **S**hap**E**-**A**ligned **S**upervision, to enhance appearance-based person re-identification. When recognizing an individual's identity, existing methods primarily rely on appearance, which can be influenced by the background environment due to a lack of body shape awareness. Although some methods attempt to incorporate other modalities, such as gait or body shape, they encode the additional modality separately, resulting in extra computational costs and lacking an inherent connection with appearance. In this paper, we explore the use of implicit 3-D body shape representations as pixel-level guidance to augment the extraction of identity features with body shape knowledge, in addition to appearance. Using body shape as supervision, rather than as input, provides shape-aware enhancements without any increase in computational cost and delivers coherent integration with pixel-wise appearance features. Moreover, for video-based person re-identification, we align pixel-level features across frames with shape awareness to ensure temporal consistency. Our results demonstrate that incorporating body shape as pixel-level supervision reduces rank-1 errors by 1.4% for frame-based and by 2.5% for video-based re-identification tasks, respectively, and can also be generalized to other existing appearance-based person re-identification methods.*

1. Introduction

Person re-identification [35, 67], which aims to identify an individual from a collection of pedestrian images or videos captured by non-overlapping cameras, is a crucial task for biometric understanding. Existing methods [20, 33, 65, 89] primarily focus on a person's appearance, which can be affected by environmental variations as the appearance is often intertwined with the background. We explore the use of 3-D body shape as supervision to enhance the human-centric appearance and demonstrate significant improvement on public datasets compared to other state-of-the-art methods, as illustrated in Figure 1.

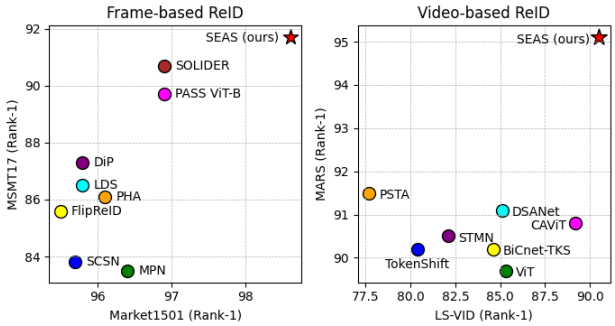

Figure 1. Rank-1 accuracy of using SEAS on ResNet-50 backbone on frame-based (MSMT17 [68] and Market1501 [81]). and PSTA as encoders compared with other state-of-the-art methods on and video-based (LS-VID [37] and MARS [82]) person re-identification datasets.

When introducing a second modality, existing re-identification methods [5, 40, 52, 78] often employ a separate branch alongside appearance for person identification. While these methods can enhance the features for re-identification, they encode each modality independently, thereby diminishing the integration of the two input modalities and weakening their connection. Additionally, the extra encoder required introduces new parameters to the network, leading to increased model size and computational cost.

In this paper, we introduce the use of human body **S**hap**E**-**A**ligned **S**upervision, abbreviated as SEAS, to enhance appearance-based re-identification methods. Instead of using a secondary modality as model input, we utilize it to guide the generation of identity features with a trainable body shape extractor. This extractor takes identity feature maps from the identity encoder and converts them into pixel-level features that represent 3-D body shapes.

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100007]. 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.

Throughout training, we direct the encoder to augment feature extraction with pixel-level shape-related information by supervising the generation of body shapes with an external pretrained model. This also causes the identity feature map to maintain more appearance information across a wider area in the body image, as it is the input for decoding pixel-level body shape features. During inference, the extractor is discarded, allowing the encoder to adeptly extract both shape and appearance-related information without incurring extra time cost. This supervision can be adapted to various encoders, as we demonstrate later in Section 4.3.

As the model input can either be a single frame or a video consisting of multiple consecutive frames, we employ different strategies to leverage the 3-D body shape in augmenting appearance. For frame-based person re-identification, where the input is a single RGB image, we incorporate a pixel-level implicit representation as supplementary guidance to provide body shape details in conjunction with appearance. During training, after the encoder produces the feature map, we attach a series of deconvolutional layers, functioning as the extractor, to upscale this map to a pixel-level representation. We supervise the decoded features from the extractor using the feature map generated by the PIFu [55] encoder with pixel-level shape guidance.

Different from a single frame, video input provides a consistent appearance across multiple frames since they depict the same person. Therefore, we integrate cross-frame appearance consistency with body shape alignment during training based on the per-pixel implicit body representation. We project the pixel-level features of each point on the body in every frame of the same video onto a unified body model and reduce the variance of the features at the same point, enforcing point-level consistency. This alignment generates a shared body model that captures appearance in 3-D space, allowing the identity encoder to extract features for both appearance and body shape with temporal consistency.

As SEAS can be applied to both frame and video-based re-identification, we assess it with both settings. For frame-based re-identification, we evaluate SEAS with ResNet-50 on Market1501 [81] and MSMT17 [68]. For video-based re-identification, we test it with PSTA [65] on MARS [82] and LS-VID [37]. Our results show a 1.4% and 2.5% average reduction in rank-1 errors for frame and video-based re-identification compared to state-of-the-art methods.

In summary, our contributions are as follows: 1) we introduce SEAS, applying shape-aligned supervision to person re-identification, utilizing 3-D body features in addition to appearance without additional cost during inference; 2) we propose a pixel-level feature alignment across frames in video-based re-identification for temporal consistency; and 3) we present the superior performance and adaptability to other encoders of using SEAS for both image and video-based re-identification through extensive experiments.

2. Related Works

Appearance-based Person Re-Identification relies on the visual appearance of individuals in images for identification [25, 31, 46, 70, 83, 91]. Earlier works were focusing on processing the entire image [2, 30, 60, 64, 71, 84], matching the query image with gallery images using the highest response of appearance similarities. Recently, the focus has shifted toward part-based feature extraction strategies for re-identification, aimed at minimizing background biases and concentrating more on human appearance for identity matching. These approaches include part-based recognition [10, 17, 24, 33, 38, 59, 89] and attention-based methods [7, 58, 65, 79]. However, these attention-driven methods tend to generate smaller attention maps, leading to a reliance on a limited portion of the image instead of the appearance of the whole body.

External Modalities for Re-Identification serve as additional guidance for identifying a person by offering extra distinguishable information beyond appearance. Two of the most common modalities besides appearance are gait [3, 16, 39, 85, 86] and 3-D body shape [5, 40, 45, 66, 78]. Gait can be captured from long distances but requires videos as input with specific walking patterns. For 3-D body modeling, existing methods [19, 40, 88] employ shape priors from SMPL [43]. However, these methods don't accurately model the body shape and encode the body shape separately, not integrating it comprehensively with appearance. As there has been development of other body shape reconstruction methods with implicit representation, integrating these reconstructions can further provide additional guidance to appearance in the input images.

Body Shape Reconstruction has seen significant improvement over recent years. Earlier efforts are primarily focused on explicit methods [43, 49] which include strong priors concerning body shape. However, recent research trends are shifting towards more advanced implicit representations, such as implicit functions [14, 48, 51, 55, 56, 61] and Neural Radiance Fields (NeRF) [29, 50, 69, 80, 87]. Some approaches are now attempting to bridge between SMPL-based methods [43] and refine the final reconstruction results [73, 74, 87]. While most of these methods exhibit promising properties for fine-grained body shape reconstruction and rendering, their potential utility for downstream vision tasks remains largely unexplored.

3. Method

As an enhancement to existing appearance-based person re-identification models, SEAS decodes body shape with an extractor that follows the identity feature encoder generating convolutional features for re-identification, as illustrated in Figure 2. During inference, only the identity encoder is retained, ensuring there is no additional computational cost.

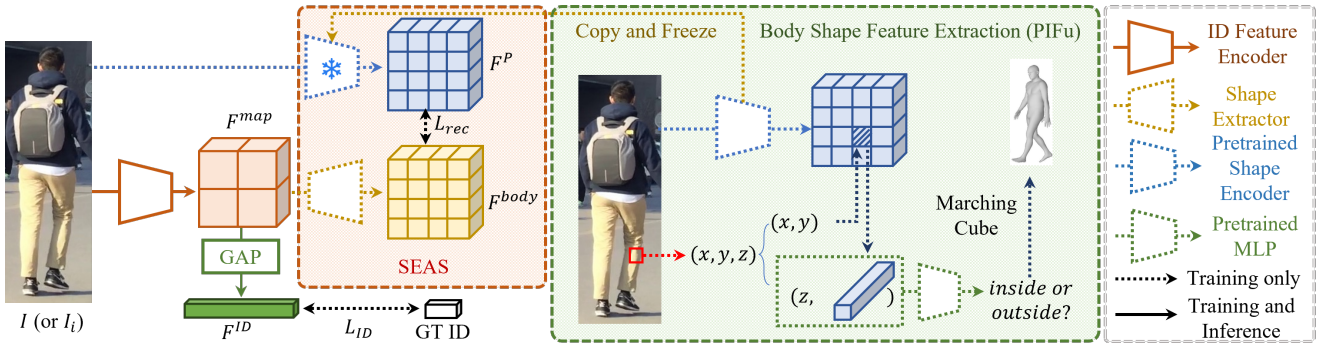

Figure 2. Person re-identification uses implicit body shape extraction as supervision, with (left) the identity feature encoder along with the SEAS pipeline, and (right) a brief overview of the PIFu [55] pipeline. Trapezoids in the figures are trainable models and GAP stands for Global Average Pooling. Dotted lines are used only during training and are excluded during inference.

In the following subsections, we start by providing a brief overview of the person re-identification task and the feature encoders used to generate identity features in Section 3.1. We then detail SEAS in Section 3.2 and provide an in-depth discussion on the training and inference phases with further analysis in Section 3.3.

3.1. Identity Feature Encoder

When taking a video $\boldsymbol{V} = \{\boldsymbol{I}_i\}$ or a single frame image $\boldsymbol{I}$ as input, person re-identification aims to encode the corresponding identity feature $\boldsymbol{F}^{ID}$ from the input modality and use it to search in the gallery for the best match. The label of the feature with the highest response in the gallery is considered as the prediction. As we have two different types of input, videos and single frames, we have two model variations for the feature encoding.

To process single-frame inputs, we first resize the input image to 256×128 and employ a ResNet-50 [21] initialized with parameters trained on ImageNet [54] to encode the image $\boldsymbol{I}$ into its corresponding feature map $\boldsymbol{F}^{map}$. This map is the output from the last convolutional layer in the encoder, which retains spatial information. We then apply a global average pooling (GAP) layer to pool the $\boldsymbol{F}^{map}$, transforming the $H \times W \times C$ feature maps into a $1 \times 1 \times C$ feature vector $\boldsymbol{F}^{ID}$ used for re-identification following

$$\boldsymbol{F}^{ID} = \text{GAP}(\boldsymbol{F}^{map}). \tag{1}$$

Similar to the extraction of the feature map from a single image, when using a video as input with sequential images $\{\boldsymbol{I}_i\}$, we utilize a shared ResNet-50 [21] to encode each frame into its corresponding feature maps $\{\boldsymbol{F}_i^{map}\}$. We follow PSTA [65] and construct the pyramid spatial and temporal attention to aggregate features from different frames, resulting in the final feature map $\boldsymbol{F}^{map}$. As the encoder is not the contribution of this work, more details are available in [65]. We then follow Equation 1 to transform the feature map into a feature vector for identification.

3.2. SEAS: Shape-Aligned Supervision

To use the 3-D shape information as guidance to input frames for re-identification, we introduce SEAS, shape-aligned supervision, to enhance the identity encoder with body shape alongside appearance. We apply SEAS for frame-based and video-based tasks separately as single frames primarily emphasize extracting features from an individual image, while frames within a video are intrinsically interlinked and can be represented by a shared 3-D body model. Given these considerations, we employ pixel-level implicit representation [55, 56] to supervise single frame input. For videos, we introduce per-pixel feature calibration atop implicit features to emphasize temporal consistency.

SEAS for Frame-based Re-Identification. We integrate pixel-level representations extracted from pretrained PIFu [55] as supervision for framewise body shape generated from SEAS. PIFu [55], as depicted on the right of Figure 2, characterizes body shape using a feature map that implicitly records whether points in 3-D space are inside or outside the object. Given an input image $\boldsymbol{I}$, PIFu extracts per-pixel level features $\boldsymbol{F}^P$ with an encoder. For each pixel, PIFu combines a depth value, denoted as z, with its associated feature $\boldsymbol{F}^P(x, y)$ sampled from the feature map, and uses their concatenation as input to an MLP network to decode the signed distance value at that specific depth. The signed distance [11, 48] indicates the distance to the nearest surface, with the sign determining whether the point is inside or outside the object. By aggregating these dense signed distance values in 3-D space, PIFu can reconstruct the object's surface by locating the zero-value surface.

We employ the extracted pixel-level body shape representation $\boldsymbol{F}^P$ to guide the extraction of body shapes using the encoded identity feature map $\boldsymbol{F}^{map}$. Since $\boldsymbol{F}^{map}$ emanates from the last convolutional layer of the identity feature encoder, it retains the highest level of semantic representation, along with spatial information before global av-

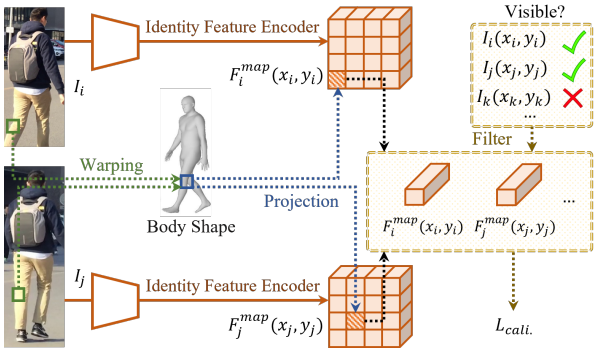

Figure 3. Feature calibration across different frames. We use a point near the left knee as an example. We warp the 2-D points using a shared 3-D body shape and project them onto the feature maps F^{map} to extract point-level features with interpolation, followed by a calibration loss $\mathcal{L}_{cali.}$ to reduce the variance of features for corresponding points across different frames that are visible and mapped to the same location on the shared body shape.

erage pooling. Furthermore, as we employ the feature post-pooling directly for identification in frame-based input per Equation 1, leveraging $\boldsymbol{F}^{map}$ to decode body shape information ensures maximal preservation of body features in the final identification representation $\boldsymbol{F}^{ID}$. Given the pooling operations in the identity and PIFu feature encoders, we upscale $\boldsymbol{F}^{map}$ to align with the size of $\boldsymbol{F}^{P}$ as follows:

$$\boldsymbol{F}^{body} = \text{UpConv} \cdot \ldots \cdot \text{UpConv}(\boldsymbol{F}^{map}). \tag{2}$$

Here, $\boldsymbol{F}^{body}$ represents the decoded per-pixel feature map for the body shape. The `UpConv` sequence consists of 2-D transposed convolutions, BatchNorm, and ReLU, followed by 2-D convolutions, BatchNorm, and ReLU, in that order. We exclude BatchNorm and ReLU from the final `UpConv` step because $\boldsymbol{F}^{P}$ may include non-positive values. The width and height of the feature map are doubled with each `UpConv` layer, with the total number of layers determined by the size difference between $\boldsymbol{F}^{map}$ and $\boldsymbol{F}^{body}$.

With the extracted body shape feature $\boldsymbol{F}^{body}$ and the corresponding PIFu-encoded feature $\boldsymbol{F}^{P}$, we apply $\mathcal{L}_{rec}$ to train the identity encoder and the extractor in SEAS using the smooth L1 loss following

$$\mathcal{L}_{rec} = \frac{1}{HW} \sum \text{SmoothL1}(\boldsymbol{F}^{body}, \boldsymbol{F}^{P}) \tag{3}$$

to compute the average of per-pixel differences across feature maps of size $H \times W$. The smooth L1 loss generates reasonable gradients near the zero point of the differences and prevents the model from overly penalizing values with large deviations from zero.

SEAS for Video-based Re-Identification. Video input includes appearance consistency across frames for the same individual, while an encoder without such correspondence awareness might generate disparate features due to varying poses or image resolutions. Inspired by [50], which suggests that the per-pixel appearance of the same body part can be decoded from the same point-level features even with different poses, maintaining pixel-level feature consistency enforces appearance consistency across frames. Therefore, in addition to SEAS for single frames, we integrate feature calibration for points across multiple frames that correspond to the same body part, as depicted in Figure 3, which ensures temporal consistency at the pixel level.

To align the body shape across different frames, we adopt three steps for feature calibration: 1) extract the per-pixel level features, 2) determine the correspondence between points across various frames and project the per-pixel features onto a shared body shape, and 3) align the features from different frames that represent the same point. We utilize the feature maps from the identity encoder $\{\boldsymbol{F}_i^{map}\}$ as our input since they also maintain the highest level of semantic representations and can avoid the potential conflict between the alignment and $\{\boldsymbol{F}_i^{P}\}$ during supervision.

With the extracted framewise per-pixel features, we establish dense correspondence using SMPL [43], which provides a predefined body shape with 6,890 vertices on the body surface, ensuring each point has a specific order and representation. For different frames in the video $\{\boldsymbol{I}_i\}$, we first warp the images using a shared body shape model to build pixel-level point correspondence between the frames and the shared body shape. We then sample k points on the body shape and locate their corresponding points in each frame, gathering the corresponding pixel-level features using bilinear interpolation from the nearest four points on the feature map. As points may be occluded, we use the normal to determine visibility and accordingly filter the features.

After collecting the features of these k points from the frames in which they are visible, we calculate the variance of the features for each point and aggregate them following

$$\mathcal{L}_{cali.} = \frac{1}{kn} \sum_{k} \sum_{n} \text{Variance}(\boldsymbol{F}_i^{map}(x_i, y_i)) \tag{4}$$

where n represents the number of frames in which the sampled points are visible. By minimizing the variance for the sampled k points, we can integrate the temporal consistency across different frames within a single video.

3.3. Training and Inference

During training, we employ three different losses: the re-identification loss $\mathcal{L}_{ID}$, the reconstruction loss $\mathcal{L}_{rec}$, which supervises the 3-D shape features produced by the SEAS extractor, and the feature calibration loss $\mathcal{L}_{cali.}$ for video-based re-identification, aligning features of the same body part across various frames. We provide more details about $\mathcal{L}_{ID}$ in Section 4.1. The overall loss is defined as

$$\mathcal{L} = \mathcal{L}_{ID} + \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{cali.} \tag{5}$$

where λ_1 and λ_2 are hyperparameters to balance the different loss terms. λ_2 is set to 0 when SEAS is applied exclusively to frame-based re-identification tasks.

During inference, we detach SEAS from the identity encoder, keeping only the encoder for identity feature extraction. We process its output feature F^{map} and convert it into F^{ID} for person identification using global average pooling. Since the body shape features can be noisy, we replace each gallery feature with the average of features from non-overlapping cameras and find the highest response of cosine similarity. This ensures the number of gallery examples does not change for a fair comparison.

Discussion. As the input to the body shape extractor is the output from the identity feature encoder, the only data used to decode body shapes originates from this encoder. Since pixel-level body shapes are predicted using the information from pixel-level features, this process facilitates a pixel-level integration of appearance and body shape information. This ensures that the feature map $\boldsymbol{F}^{map}$, which is used for re-identification after global average pooling, not only is explicitly shape-aware but also maximally retains the pixel-level appearance information for the body area to decode pixel-level body shapes, instead of focusing on a small patch of it for re-identification. Additionally, when processing video inputs, employing a shared body shape projector aligns different points from various input perspectives into a unified space. This alignment provides explicit temporal consistency across frames, allowing the encoder to extract temporally coherent information inherently. We verify this with multiple model variations as well as attention visualization, which we present later in Section 4.3.

Compared with existing methods [5, 40] that explicitly decode body shapes from images and use them for identification, SEAS differs in two main aspects: 1) SEAS does not explicitly require 3-D reconstruction as guidance for identification, while it preserves it as implicit supervision for pixel-level coherence with appearance. 2) SEAS directly connects the appearance and shape across different frames within a video. Existing methods [5, 40] reconstruct the frame-wise body shapes and do not aggregate the temporal information across different frames, while SEAS can explicitly provide such guidance with variance computation on the same vertex of the SMPL body to offer more appearance consistency without extra computational cost during inference, as only the encoder is needed.

4. Experiments

4.1. Experimental Details

Datasets. For our experiments, we utilize Market1501 [81] with MSMT17 [68] for frame-based person re-identification and MARS [82] in conjunction with LS-VID [37] for video-based person re-identification.

Market1501 [81] and MSMT17 [68] serve as two datasets for the person re-identification public benchmark with single-frame images as input. Market1501 [81] includes 1,501 identities, with 750 designated for training and the remaining 751 for testing, captured across six cameras. The training set, gallery, and query consist of 12,936, 3,368, and 19,732 cropped images, respectively. MSMT17 [68] comprises 4,101 identities, with 1,041 for training and 3,060 for testing. Images are collected from 15 cameras (12 for outdoor scenes and 3 for indoor scenes) with the numbers of images for training, gallery, and query at 30,248, 82,161, and 11,659, respectively.

MARS [82] and LS-VID [37] are two public datasets for video-based re-identification. MARS [82] contains 1,261 unique identities and a total of 17,503 tracklets, with 625 identities reserved for training and 636 for testing. LS-VID [37] includes 842 identities for training and 2,730 for testing, with training sequences comprising 8,298 video segments, while the gallery and query sets contain 11,310 and 1,980 segments, respectively.

Implementation Details. In our pipeline, we use two shape representations for guidance: PIFu [55] for both videos and frames, and SMPL [43] for videos. Since appearance is largely preserved in images, we use the pretrained PIFu surface model [55] and keep it forzen during training for shape guidance, accompanied by a pretrained DeepLab-v3 [6] to remove the background. For SMPL [43], we follow SPIN [34] to extract the 3-D body shapes for each frame. The number of points used for training, k, is set to 500. We empirically set λ_1 as 1 for both cases and λ_2 as 0.001 for video-based re-identification. During inference, we calculate the pairwise cosine similarity between the query and gallery examples for both videos and images.

For frame-based person re-identification, we resize the input images to 256×128 and use a ResNet-50 [21] to encode the features into $\boldsymbol{F}^{map}$. The shape extractor, which follows the feature map, includes 2 `UpConv` operations to upscale the feature map to its pixel-level implicit representation feature map $\boldsymbol{F}^{P}$. Each `UpConv` operation increases the feature dimension by 2 on both height and width. During training, we incorporate a Triplet loss $\mathcal{L}_{triplet}$ [57] with a 0.3 margin, along with a Cross Entropy loss $\mathcal{L}_{CE}$ as

$$\mathcal{L}_{ID} = \mathcal{L}_{triplet} + \mathcal{L}_{CE}. \tag{6}$$

We train the model for 120 epochs, starting with a learning rate of $3.5e^{-4}$ using the Adam optimizer, and reduce it by $\frac{1}{10}$ at epoch 40 and 70.

For video-based re-identification, we follow PSTA [65] and use the ResNet-50 as our feature encoder, followed by a three-layer pyramid for spatial and temporal attention extraction and aggregation. We include a Triplet loss $\mathcal{L}_{triplet}$ [57] with 0.3 as margin along with a Cross Entropy loss $\mathcal{L}_{CE}$ as Equation 6. We train the model for 500 epochs

with the Adam optimizer, with an initial learning rate set at $3.5e^{-4}$. We reduce the learning rate by 0.3 at epochs 70, 140, 210, 310, and 410. During training, we randomly select 8 frames from each clip. In the inference phase for MARS [82], we select 8 frames per video, starting with the first frame and proceeding at intervals of 8 frames. If a video's length doesn't permit an 8-frame interval, we halve the interval until 8 frames can be selected. For LS-VID, we utilize all available frames in the video and average the features from the last layer of PSTA as its representation.

We conduct all our experiments on a machine equipped with 2 A100 GPUs, and recommend using 1 GPU with 24GB memory for training frame-based datasets and 2 for video-based datasets. The training time for frame-based re-identification ranges from 12 to 30 hours; for video-based re-identification, it varies between one to three days, depending on the datasets used and the I/O speed.

Baseline Methods. For frame-based re-identification, we build SEAS on a ResNet-50 encoder and compare it with multiple state-of-the-art methods, including earlier ReID methods [9, 12], recent transformer-based approaches [8, 13, 22, 89, 90], and others [18, 36, 47, 53, 63, 75–77]. In addition, we compare with some other re-identification methods [5, 40] that use 3-D body shape in addition to appearance. We assess the methods based on rank-1 accuracy and mean average precision (mAP) for comparison.

For video-based re-identification, we compare with several state-of-the-art models, including models with 2-D convolutions [15, 23, 28, 32, 41, 42, 65], 3-D convolutions [1, 19, 26, 27], and ViT-based methods [4, 13, 72], measuring rank-1 accuracy and mAP. Following PSTA [65], we use a ResNet-50 [21] as our per-frame image encoder, followed by the pyramid spatial-temporal aggregation.

4.2. Results and Analysis

Results for Frame-based Re-ID. We present the numerical results for frame-based person re-identification in Table 1, comparing the application of SEAS on ResNet-50 with other state-of-the-art methods on Market1501 [62] and MSMT17 [68] in terms of rank-1 accuracy and mean average precision (mAP). Using SEAS with ResNet-50 as the image encoder outperforms all other methods on both datasets across these metrics. On Market1501 [62], ResNet-50 with SEAS achieves a rank-1 accuracy of 98.6% and an mAP of 98.9%, surpassing SOLIDER [8], which had the previous highest rank-1 accuracy of 96.9% with external training data. We also observe significant improvements on MSMT17 [68], with rank-1 accuracy increasing from 90.7% to 91.7% and mAP from 77.1% to 92.8%.

Moreover, we also include a comparison with explicit appearance reconstruction as in SAN [30] and other methods [5, 40] using body shapes. SAN includes a decoder-like structure for explicitly reconstructing the appearance of the entire body, even when occlusions are present. We note that using SEAS outperforms SAN on both datasets; the gap is more significant on MSMT17 [68], where occlusions are more common, making the reconstruction of the entire body appearance impractical. Compared with other methods that use features of 3-D body shapes [5, 40] for identification, ResNet-50 with SEAS also shows significant improvement. The use of a body shape extractor with pixel-level body shape guidance establishes a stronger connection between shape and observed appearance, providing a more complete understanding for person re-identification.

Method	Market1501 [81]		MSMT17 [68]	
	Rank-1	mAP	Rank-1	mAP
ViT-B [13]	94.0	87.6	82.8	63.6
TransReID [22]	95.2	89.5	86.2	69.4
AAFormer [90]	95.4	87.7	83.6	63.2
AGW [75]	95.5	89.5	81.2	59.7
FlipReID [47]	95.5	89.6	85.6	68.0
CAL [53]	95.5	89.5	84.2	64.0
PFD [63]	95.5	89.7	83.8	64.4
SAN [30]	96.1	88.0	79.2	55.7
LDS [76]	95.8	90.4	86.5	67.2
DiP [36]	95.8	90.3	87.3	71.8
MPN [12]	96.4	90.1	83.5	62.7
MSINet [18]	95.3	89.6	81.0	59.6
SCSN [9]	95.7	88.5	83.8	58.5
PHA [77]	96.1	90.2	86.1	68.9
PASS ViT-B [89]	<u>96.9</u>	93.3	89.7	74.3
SOLIDER [8]	<u>96.9</u>	<u>93.9</u>	<u>90.7</u>	<u>77.1</u>
ASSP* [5]	95.0	87.3	-	-
3DInvarReID* [40]	95.1	87.9	80.8	59.1
Baseline (ResNet-50)	94.1	83.2	73.8	47.2
SEAS (ResNet-50)	**98.6**	**98.9**	**91.7**	**92.8**

Table 1. Rank-1 accuracy and mAP on Market1501 [81] and MSMT17 [68] datasets. We bold the numbers that are the best performance and underline the second best ones. Methods ending with (*) include 3-D body shapes in addition to appearance.

Results for Video-based Re-ID. In addition to single-frame-based person re-identification, we present our results for video-based datasets [37, 82] in Table 2. Across both datasets, we note a significant improvement over other baseline methods. For MARS [82], the highest rank-1 and mAP from the current state-of-the-art methods are 91.5% and 87.2%, respectively. In comparison, our method achieves 95.1% for rank-1 and 96.7% for mAP, surpassing the existing baselines and reducing the errors from 8.5% to 4.9%. We also observe a performance boost on LS-VID [37]. Against the best performance for rank-1 and mAP, both from CAViT [72], which stand at 89.2% and 79.2%, using a PSTA with SEAS enhanced with $\mathcal{L}_{cali.}$ exceeds these metrics, reaching a rank-1 of 90.5% and an mAP of 93.4%. This demonstrates the efficacy of SEAS with simple encoders over other methods that leverage vision transformers.

Method	MARS [82]		LS-VID [37]	
	Rank-1	mAP	Rank-1	mAP
GRL [42]	91.0	84.8	-	-
TokenShift [4]	90.2	86.6	80.4	68.7
ViT [13]	89.7	86.4	85.3	76.4
TCLNet [26]	89.8	85.1	-	-
AP3D [19]	90.1	85.1	-	-
DenseIL [23]	90.8	87.0	-	-
STMN [15]	90.5	84.5	82.1	69.2
BiCnet-TKS [27]	90.2	86.0	84.6	75.1
STRF [1]	90.3	86.1	-	-
RFCnet [28]	90.7	86.3	-	-
CTL [41]	91.4	86.7	-	-
DSANet [32]	91.1	86.6	85.1	75.5
CAViT [72]	90.8	87.2	89.2	79.2
Baseline (PSTA) [65]	91.5	85.8	77.7	67.2
SEAS (PSTA)	**95.1**	**96.6**	**90.5**	**93.4**

Table 2. Rank-1 accuracy and mAP on MARS [82] and LS-VID [37] datasets. In SEAS, we use ResNet-50 for identity feature map extraction and PSTA [65] for temporal fusion.

Method		Rank-1	mAP	Params	FLOPs
(I) Appearance	Baseline (Market1501)	94.1	83.2	23.51M	4.07G
(II) Body shape as input	+ PIFu as 2^{nd} branch	94.1 (+0.0)	84.8 (+1.6)	34.80M	6.28G
	+ PIFu concatenation	94.3 (+0.2)	85.8 (+2.6)	34.89M	4.26G
(III) Body shape as supervision	+ SEAS (SPIN)	97.1 (+3.0)	97.8 (+14.6)	23.51M	4.07G
	+ SEAS (PIFu)	**98.6** (+4.5)	**98.9** (+15.7)	23.51M	4.07G
(IV) SEAS w/ calibration for video frames	Baseline (MARS)	91.5	85.8	35.43M	37.70G
	+ SEAS (w/o $\mathcal{L}_{cali.}$)	94.8 (+3.3)	96.5 (+10.7)	35.43M	37.70G
	+ SEAS (w/ $\mathcal{L}_{cali.}$)	**95.1** (+3.6)	**96.7** (+10.9)	35.43M	37.70G

Table 3. Model variations analysis, with ResNet-50 as the baseline for Market1501 and PSTA [65] for MARS.

4.3. Ablation Studies and Model Variations

Module Analysis. We present model components analysis in Table 3, comparing performance and computational costs with other variations on Market1501 [62] and MARS [82]. Based on the input and different ways of aggregating the body shape feature, we split the table into four categories and have the following observations:

Enhancing Appearance with Body Shape. Compared to the baseline method that relies solely on appearance in (I), incorporating body shape in (II) and (III) generally demonstrates better performance. We experiment with three variations: 1) encoding the PIFu feature with a secondary branch encoder (ResNet-18-like) and concatenating it with the appearance feature, 2) channel-wise concatenation of PIFu features with appearance as input, and 3) employing body shape as supervision following SEAS. All variations result in improved performance, confirming that the body shape feature enhances the appearance-based method.

Using Shape as Input vs. Supervision. Compared to using body shape as input to identify the person in (II), using body shape as supervision in (III) shows better performance for both metrics. While employing body shape as input incorporates general shape information, utilizing it as supervision establishes a stronger connection between pixel-level aligned body shape and appearance, as it guides the encoder to extract shape-related information based on the appearance, facilitating a more coherent integration.

Pixel and Image-level Supervision. We include experiments comparing the use of SMPL features generated by SPIN [34] and PIFu features as supervision in (III). SMPL features are image-level, and we use the reconstructed body shape β for supervision. We observe that using either of them as supervision enhances performance, while employing PIFu for pixel-level supervision shows the best results. Compared with image-level supervision, pixel-level supervision provides a more specific guidance between pixel-level features, making the prediction more accurate.

Cross-frame Consistency in Video. For video-based re-identification, as shown in (IV), applying PIFu as single-frame body shape guidance leads to improvements compared to the baseline method [65]. Moreover, pixel-level calibration across frames further enhances performance.

Time Consumption[1]. Since SEAS is employed only during training, its introduction in (III) and (IV) does not increase the number of parameters and FLOPs required. Conversely, explicitly using the body shape as input in (II) significantly increases these two metrics.

Generalizability. To validate its generalizability, we incorporate SEAS with other frame-based re-identification methods such as BoT [44] and LDS [76], as well as video-

Method	Market1501 [81]		MSMT17 [68]	
	Rank-1	mAP	Rank-1	mAP
BoT [44]	94.5	85.9	74.1	50.2
w/ SEAS	**95.9**	**97.5**	**81.3**	**86.2**
LDS [76]	95.8	90.4	86.5	67.2
w/ SEAS	**96.3**	**97.8**	**86.6**	**90.1**

Table 4. Results for applying SEAS on BoT [44] and LDS [76].

Method	MARS [82]		LS-VID [37]	
	Rank-1	mAP	Rank-1	mAP
STMN [15]	90.5	84.5	82.1	69.2
w/ SEAS	**92.2**	**94.9**	**84.1**	**88.9**
BiCnet-TKS [27]	**90.2**	86.0	84.6	75.1
w/ SEAS	90.1	**87.9**	**86.7**	**90.8**

Table 5. Results for applying SEAS on STMN [15] and BiCnet-TKS [27] for video-based person re-identification.

[1]For the calculation of FLOPs and the number of parameters, we refer to https://github.com/facebookresearch/fvcore/blob/main/docs/flop_count.md.

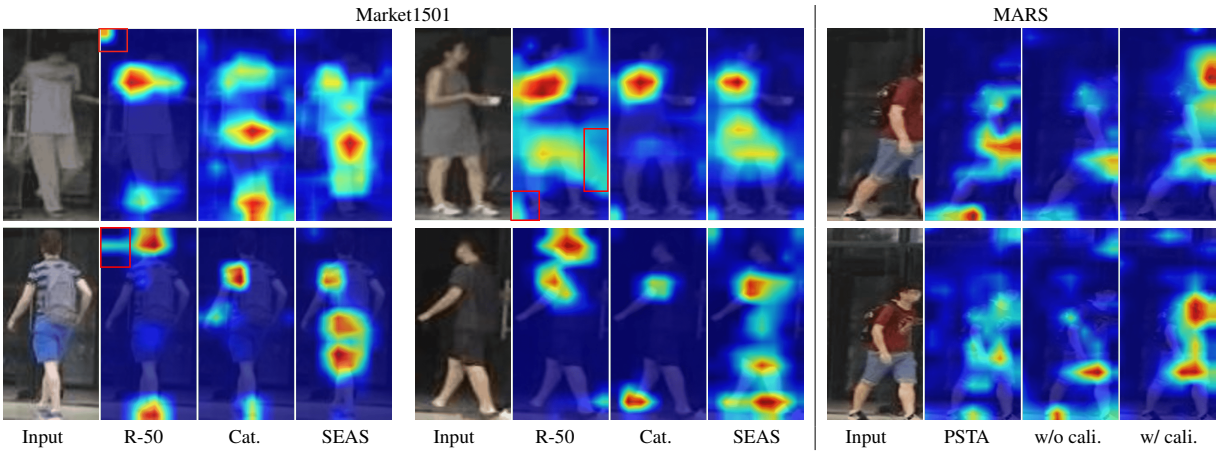

Figure 4. GradCam visualizations on Market1501 and MARS. For Market1501, we compare the baseline ResNet-50 (R-50) with using PIFu as input through concatenation (Cat.) and using SEAS. For MARS, we compare baseline PSTA with SEAS without and with calibration.

based re-identification models including STMN [15] and BiCnet-TKS [27]. We present results comparing models with and without SEAS in Tables 4 and 5, and observe improvements on both datasets for these settings and metrics. We also evaluate SEAS with CTL [70] and improve rank-1 accuracy from 98.0 to 98.6 by using the centroid of the gallery features for matching, indicating that incorporating SEAS generally enhances performance.

Visualization. We present attention maps overlayed with input images in Figure 4, verifying where the model is looking at for prediction. For Market1501, we compare the baseline with using PIFu as input and SEAS. The baseline ResNet-50, lacking body shape guidance, often erroneously directs attention to the background, as highlighted by the areas within red boxes. Introducing PIFu as input helps focus on the body by delineating its boundary, yet attention remains confined to a narrow area. SEAS shows a more even distribution across the body. As pixel-level body shape features are predicted from pixel-level appearances, using aligned body shapes as supervision ensures the appearance features related to the body region go deeper into the network and enrich the feature map with information across all body-present regions, enhancing the identity feature vector. Furthermore, we assess PSTA against SEAS, with and without calibration, on two frames of the same video in MARS. PSTA scatters attention across frames and focuses narrowly on each frame, while SEAS with calibration, generates a more expansive attention map, demonstrating its capability in providing enriched and consistent features.

We also visualize the extracted body shapes using examples from MSMT17 [68] in Figure 5. Each example displays the original image on the left, the SMPL from SPIN [34] in the middle, and the PIFu [55] shape reconstruction using the extractor on the right. The SMPL body shape from SPIN aligns well with the overall shape of the person, showing the capacity of providing dense correspondence on feature maps $\{\boldsymbol{F}_i^{map}\}$ with lower resolution. PIFu reconstructions, while limited by the image quality in re-identification datasets, effectively outline body shapes with its depth and contour, showing it is helpful to guide the identity feature encoder with pixel-level shape representations.

Figure 5. Two visualization examples from MSMT17 for SPIN (second column) and PIFu (third column) reconstructions.

5. Conclusion and Limitation

In this work, we introduce SEAS, using Shape-Aligned Supervision, to enhance the appearance feature for person re-identification. We utilize implicit body shape representations to supervise the training of the appearance-based identity encoder, with a shape extractor to translate the feature map into pixel-level body shapes, providing pixel-level shape guidance. For video-based person re-identification, we also incorporate temporal consistency across the appearance of different frames within the same video by adding pixel-level calibrations. Our method achieves state-of-the-art performance in both frame-based and video-based person re-identification evaluations on public datasets.

Limitations. The effectiveness of our proposed extractor relies on a pretrained shape encoder for supervision, meaning the overall performance improvement correlates with the quality of the extracted body shape features.

References

[1] Abhishek Aich, Meng Zheng, Srikrishna Karanam, Terrence Chen, Amit K Roy-Chowdhury, and Ziyan Wu. Spatio-temporal representation factorization for video-based person re-identification. In *ICCV*, pages 152–162, 2021. 6, 7

[2] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *CVPR*, pages 2109–2118, 2018. 2

[3] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, pages 8126–8133, 2019. 2

[4] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *ECCV*, pages 660–676, 2020. 6, 7

[5] 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*, pages 8146–8155, 2021. 1, 2, 5, 6

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 5

[7] Peixian Chen, Wenfeng Liu, Pingyang Dai, Jianzhuang Liu, Qixiang Ye, Mingliang Xu, Qi'an Chen, and Rongrong Ji. Occlude them all: Occlusion-aware attention network for occluded person re-id. In *ICCV*, pages 11833–11842, 2021. 2

[8] 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*, pages 15050–15061, 2023. 6

[9] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Salience-guided cascaded suppression network for person re-identification. In *CVPR*, pages 3300–3310, 2020. 6

[10] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *CVPR*, pages 7308–7318, 2022. 2

[11] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, pages 303–312, 1996. 3

[12] Changxing Ding, Kan Wang, Pengfei Wang, and Dacheng Tao. Multi-task learning with coarse priors for robust part-aware person re-identification. *TPAMI*, 44(3):1474–1488, 2020. 6

[13] 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. *arXiv preprint arXiv:2010.11929*, 2020. 6, 7

[14] Yueqi Duan, Haidong Zhu, He Wang, Li Yi, Ram Nevatia, and Leonidas J Guibas. Curriculum deepsdf. In *ECCV*, pages 51–67, 2020. 2

[15] Chanho Eom, Geon Lee, Junghyup Lee, and Bumsub Ham. Video-based person re-identification with spatial and temporal memory networks. In *ICCV*, pages 12036–12045, 2021. 6, 7, 8

[16] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *CVPR*, pages 14225–14233, 2020. 2

[17] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *CVPR*, pages 11744–11752, 2020. 2

[18] Jianyang Gu, Kai Wang, Hao Luo, Chen Chen, Wei Jiang, Yuqiang Fang, Shanghang Zhang, Yang You, and Jian Zhao. Msinet: Twins contrastive search of multi-scale interaction for object reid. In *CVPR*, pages 19243–19253, 2023. 6

[19] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *ECCV*, pages 228–243. Springer, 2020. 2, 6, 7

[20] Ke Han, Shaogang Gong, Yan Huang, Liang Wang, and Tieniu Tan. Clothing-change feature augmentation for person re-identification. In *CVPR*, pages 22066–22075, 2023. 1

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 5, 6

[22] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, 2021. 6

[23] Tianyu He, Xin Jin, Xu Shen, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Dense interaction learning for video-based person re-identification. In *ICCV*, pages 1490–1501, 2021. 6, 7

[24] Tianyu He, Xu Shen, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Partial person re-identification with part-part correspondence learning. In *CVPR*, pages 9105–9115, 2021. 2

[25] 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*, pages 10513–10522, 2021. 2

[26] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. In *ECCV*, pages 388–405. Springer, 2020. 6, 7

[27] Ruibing Hou, Hong Chang, Bingpeng Ma, Rui Huang, and Shiguang Shan. Bicnet-tks: Learning efficient spatial-temporal representation for video person re-identification. In *CVPR*, pages 2014–2023, 2021. 6, 7, 8

[28] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Feature completion for occluded person re-identification. *TPAMI*, 44(9):4894–4912, 2021. 6, 7

[29] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *ICCV*, pages 9352–9364, 2023. 2

[30] Xin Jin, Cuiling Lan, Wenjun Zeng, Guoqiang Wei, and Zhibo Chen. Semantics-aligned representation learning for

person re-identification. In *AAAI*, pages 11173–11180, 2020. 2, 6

[31] 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*, pages 14278–14287, 2022. 2

[32] Minjung Kim, MyeongAh Cho, and Sangyoun Lee. Feature disentanglement learning with switching and aggregation for video-based person re-identification. In *WACV*, pages 1603–1612, 2023. 6, 7

[33] Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, and Kwanghoon Sohn. Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In *CVPR*, pages 18621–18632, 2023. 1, 2

[34] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, pages 2252–2261, 2019. 5, 7, 8

[35] Qingming Leng, Mang Ye, and Qi Tian. A survey of open-world person re-identification. *TCSVT*, 30(4):1092–1108, 2019. 1

[36] Dengjie Li, Siyu Chen, Yujie Zhong, Fan Liang, and Lin Ma. Dip: Learning discriminative implicit parts for person re-identification. *arXiv preprint arXiv:2212.13906*, 2022. 6

[37] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *ICCV*, pages 3958–3967, 2019. 1, 2, 5, 6, 7

[38] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *CVPR*, pages 2898–2907, 2021. 2

[39] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *ICCV*, pages 14648–14656, 2021. 2

[40] Feng Liu, Minchul Kim, ZiAng Gu, Anil Jian, and Xiaoming Liu. Learning clothing and pose invariant 3d shape representation for long-term person re-identification. *ICCV*, 2023. 1, 2, 5, 6

[41] Jiawei Liu, Zheng-Jun Zha, Wei Wu, Kecheng Zheng, and Qibin Sun. Spatial-temporal correlation and topology learning for person re-identification in videos. In *CVPR*, pages 4370–4379, 2021. 6, 7

[42] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, and Xiaoyun Yang. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *CVPR*, pages 13334–13343, 2021. 6, 7

[43] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 2, 4, 5

[44] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR workshops*, 2019. 7

[45] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 3d clothed human reconstruction in the wild. In *ECCV*, pages 184–200, 2022. 2

[46] Kien Nguyen, Clinton Fookes, Sridha Sridharan, Feng Liu, Xiaoming Liu, Arun Ross, Dana Michalski, Huy Nguyen, Debayan Deb, Mahak Kothari, et al. Ag-reid 2023: Aerial-ground person re-identification challenge results. 2

[47] Xingyang Ni and Esa Rahtu. Flipreid: closing the gap between training and inference in person re-identification. In *EUVIPW*, pages 1–6, 2021. 6

[48] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, pages 165–174, 2019. 2, 3

[49] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 2

[50] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Animatable neural radiance fields for human body modeling. *arXiv preprint arXiv:2105.02872*, 2021. 2, 4

[51] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, pages 9054–9063, 2021. 2

[52] Haocong Rao and Chunyan Miao. Transg: Transformer-based skeleton graph prototype contrastive learning with structure-trajectory prompted reconstruction for person re-identification. In *CVPR*, pages 22118–22128, 2023. 1

[53] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *ICCV*, pages 1025–1034, 2021. 6

[54] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015. 3

[55] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, pages 2304–2314, 2019. 2, 3, 5, 8

[56] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, pages 84–93, 2020. 2, 3

[57] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 5

[58] Minho Shim, Hsuan-I Ho, Jinhyung Kim, and Dongyoon Wee. Read: Reciprocal attention discriminator for image-to-video re-identification. In *ECCV*, pages 335–350, 2020. 2

[59] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person re-identification. In *WACV*, pages 1613–1623, 2023. 2

[60] Maoqing Tian, Shuai Yi, Hongsheng Li, Shihua Li, Xuesen Zhang, Jianping Shi, Junjie Yan, and Xiaogang Wang. Eliminating background-bias for robust person re-identification. In *CVPR*, pages 5794–5803, 2018. 2

[61] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *ECCV*, pages 1–19, 2022. 2

[62] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703, 2014. 6, 7

[63] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *AAAI*, pages 2540–2549, 2022. 6

[64] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. Person re-identification with cascaded pairwise convolutions. In *CVPR*, pages 1470–1478, 2018. 2

[65] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid spatial-temporal aggregation for video-based person re-identification. In *ICCV*, pages 12026–12035, 2021. 1, 2, 3, 5, 6, 7

[66] Yanan Wang, Xuezhi Liang, and Shengcai Liao. Cloning outfits from real-world images to 3d characters for generalizable person re-identification. In *CVPR*, pages 4900–4909, 2022. 2

[67] Zheng Wang, Zhixiang Wang, Yinqiang Zheng, Yang Wu, Wenjun Zeng, and Shin'ichi Satoh. Beyond intra-modality: A survey of heterogeneous person re-identification. In *IJCAI*, 2020. 1

[68] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 1, 2, 5, 6, 7, 8

[69] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, pages 16210–16220, 2022. 2

[70] Mikołaj Wieczorek, Barbara Rychalska, and Jacek Dąbrowski. On the unreasonable effectiveness of centroids in image retrieval. In *ICONIP*, pages 212–223, 2021. 2, 8

[71] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, pages 5380–5389, 2017. 2

[72] Jinlin Wu, Lingxiao He, Wu Liu, Yang Yang, Zhen Lei, Tao Mei, and Stan Z Li. Cavit: Contextual alignment vision transformer for video object re-identification. In *ECCV*, pages 549–566, 2022. 6, 7

[73] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *CVPR*, pages 13286–13296, 2022. 2

[74] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *CVPR*, pages 512–523, 2023. 2

[75] 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*, 44(6):2872–2893, 2021. 6

[76] Xianghao Zang, Ge Li, Wei Gao, and Xiujun Shu. Learning to disentangle scenes for person re-identification. *IVC*, 116: 104330, 2021. 6, 7

[77] Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, and Shiliang Pu. Pha: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *CVPR*, pages 14133–14142, 2023. 6

[78] Quan Zhang, Kaiheng Dang, Jian-Huang Lai, Zhanxiang Feng, and Xiaohua Xie. Modeling 3d layout for group re-identification. In *CVPR*, pages 7512–7520, 2022. 1, 2

[79] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, pages 3186–3195, 2020. 2

[80] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *CVPR*, pages 7743–7753, 2022. 2

[81] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 1, 2, 5, 6, 7

[82] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884, 2016. 1, 2, 5, 6, 7

[83] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *CVPR*, pages 1367–1376, 2017. 2

[84] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J Radke. Re-identification with consistent attentive siamese networks. In *CVPR*, pages 5735–5744, 2019. 2

[85] Haidong Zhu, Wanrong Zheng, Zhaoheng Zheng, and Ram Nevatia. Gaitref: Gait recognition with refined sequential skeletons. *IJCB*, 2023. 2

[86] Haidong Zhu, Zhaoheng Zheng, and Ram Nevatia. Gait recognition using 3-d human body shape inference. In *WACV*, pages 909–918, 2023. 2

[87] Haidong Zhu, Zhaoheng Zheng, Wanrong Zheng, and Ram Nevatia. Cat-nerf: Constancy-aware tx2former for dynamic body modeling. In *CVPRW*, pages 6618–6627, 2023. 2

[88] Haidong Zhu, Wanrong Zheng, Zhaoheng Zheng, and Ram Nevatia. Sharc: Shape and appearance recognition for person identification in-the-wild. In *WACV*, 2024. 2

[89] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. Pass: Part-aware self-supervised pre-training for person re-identification. In *ECCV*, pages 198–214, 2022. 1, 2, 6

[90] Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Jing Liu, Jinqiao Wang, and Ming Tang. Aaformer: Auto-aligned transformer for person re-identification. *TNNLS*, 2023. 6

[91] Yang Zou, Xiaodong Yang, Zhiding Yu, BVK Vijaya Kumar, and Jan Kautz. Joint disentangling and adaptation for cross-domain person re-identification. In *ECCV*, pages 87–104, 2020. 2