Semantics-Aligned Representation Learning for Person Re-identification

Xin Jin[1*] **Cuiling Lan**[2†] **Wenjun Zeng**[2] **Guoqiang Wei**[1] **Zhibo Chen**[1†]
University of Science and Technology of China[1] Microsoft Research Asia[2]
{jinxustc,wgq7441}@mail.ustc.edu.cn {culan, wezeng}@microsoft.com chenzhibo@ustc.edu.cn

Abstract

Person re-identification (reID) aims to match person images to retrieve the ones with the same identity. This is a challenging task, as the images to be matched are generally semantically misaligned due to the diversity of human poses and capture viewpoints, incompleteness of the visible bodies (due to occlusion), *etc*. In this paper, we propose a framework that drives the reID network to learn semantics-aligned feature representation through delicate supervision designs. Specifically, we build a Semantics Aligning Network (SAN) which consists of a base network as encoder (SA-Enc) for reID, and a decoder (SA-Dec) for *reconstructing/regressing the densely semantics aligned full texture image*. We jointly train the SAN under the supervisions of person re-identification and aligned texture generation. Moreover, at the decoder, besides the reconstruction loss, we add Triplet ReID constraints over the feature maps as the perceptual losses. The decoder is discarded in the inference and thus our scheme is computationally efficient. Ablation studies demonstrate the effectiveness of our design. We achieve the state-of-the-art performances on the benchmark datasets CUHK03, Market1501, MSMT17, and the partial person reID dataset Partial REID.

1 Introduction

Person re-identification (reID) aims to identify/match persons in different places, times, or camera views. There are large variations in terms of the human poses, capturing view points, incompleteness of the bodies (due to occlusion). These result in *semantics misalignment* across 2D images which makes reID challenging (Shen et al. 2015; Varior et al. 2016; Subramaniam, Chatterjee, and Mittal 2016; Su et al. 2017; Zheng et al. 2017; Zhang et al. 2017; Yao et al. 2017; Li et al. 2017; Zhao et al. 2017; Wei et al. 2017; Zheng, Zheng, and Yang 2018; Ge et al. 2018; Suh et al. 2018; Qian et al. 2018; Zhang et al. 2019).

Semantics misalignment can be interpreted from two aspects. (1) Spatial semantics misalignment: the same spatial position across images may correspond to different semantics of human body or even different objects. As the example in Figure 1 (a) shows, the spatial position A which corresponds to person leg in the first image corresponds to person abdomen in the second image. (2) Inconsistency of visible body regions/semantics: since a person is captured through a 2D projection, only a portion of the 3D surface of a person is visible/projected in an image. The visible body regions/semantics across images are not consistent. As shown in Figure 1(b), front side of a person is visible in one image and invisible in another one.

Figure 1: Challenges in person reID: (a) Spatial misalignment; (b) Inconsistency of the visible body regions/semantics.

Alignment: Deep learning methods can deal with such diversities and misalignment to some extent but it is not enough. In recent years, many approaches explicitly exploit human pose/landmark information to achieve coarse alignment and they have demonstrated their superiority for person reID (Su et al. 2017; Zheng et al. 2017; Yao et al. 2017; Li et al. 2017; Zhao et al. 2017; Wei et al. 2017; Suh et al. 2018). During the inference, these part detection sub-networks are usually required which increases the computational complexity. Besides, the body-part alignment is coarse and there is still spatial misalignment within the parts (Zhang et al. 2019). To achieve fine-granularity spatial alignment, based on estimated dense semantics (Güler, Neverova, and Kokkinos 2018), Zhang *et al*. warp the input person image to a canonical UV coordinate system to have densely semantics aligned images as inputs for reID (Zhang et al. 2019). However, the invisible body regions result in many holes in the warped images and thus the inconsistency of

*This work was done when Xin Jin was an intern at MSRA.
†Corresponding Author.

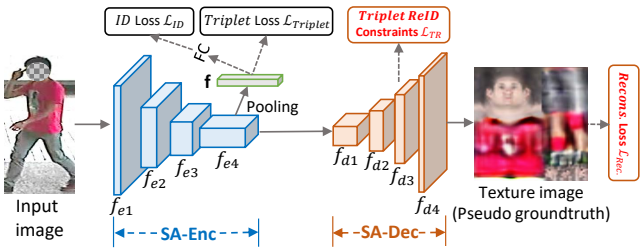

Figure 2: Illustration of the proposed Semantics Aligning Network (SAN), which consists of a base network as encoder (SA-Enc) and a decoder sub-network (SA-Dec). The reID feature vector **f** is obtained by average pooling the feature map f_{e4} of the SA-Enc, followed by the reID losses of $\mathcal{L}_{ID}$ and $\mathcal{L}_{Triplet}$. To encourage the encoder learning semantically aligned features, the SA-Dec is followed which regresses the densely semantically *aligned full* texture image with the pseudo groundtruth supervision $\mathcal{L}_{Rec.}$. The pseudo groundtruth generation is described in Sec. 3.1 without shown here. At the decoder, Triplet ReID constraints $\mathcal{L}_{TR}$ are added as the high level perceptual metric. We use ResNet-50 with four residual blocks as our SA-Enc. In inference, the SA-Dec is discarded.

visible body regions across images. How to better solve the dense semantics misalignment is still an open problem.

Our work: We intend to fully address the semantics misalignment problems in both aspects. We achieve this by proposing a simple yet powerful Semantics Aligning Network (SAN). Figure 2 shows the overall framework of the SAN, which introduces an aligned texture generation subtask, with densely semantics aligned texture image (see examples in Figure 3) as supervision. Specifically, SAN consists of a base network as encoder (SA-Enc), and a decoder sub-network (SA-Dec). The SA-Enc can be any baseline network used for person reID (*e.g.* ResNet-50 (He et al. 2016)), which outputs a feature map f_{e4} of size $h \times w \times c$. The reID feature vector $\mathbf{f} \in \mathbb{R}^c$ is then obtained by average pooling the feature map f_{e4}, followed by the reID losses. To encourage the SA-Enc to learn semantically aligned features, the SA-Dec is introduced and used to regress/generate the densely semantically aligned full texture image (also referred to as texture image for short) with pseudo groundtruth supervision. We exploit a synthesized dataset for learning pseudo groundtruth texture image generation. This framework enjoys the benefit of dense semantics alignment but without increasing the complexity of inference since the decoder SA-Dec is discarded in inference.

Our main contributions are summarized as follows.

- We propose a simple yet powerful framework for solving the misalignment challenge in person reID without increasing computational cost in inference.
- A semantics alignment constraint is delicately introduced by empowering the encoded feature map with *aligned full* texture generation capability.
- At the SA-Dec, besides the reconstruction loss, we propose Triplet ReID constraints over the feature maps as the perceptual metric.
- There is no groundtruth aligned texture image for the person reID datasets. We address this by generating pseudo groundtruth texture images by leveraging synthesized data with person image and aligned texture image pairs (see Figure 3).

Our method achieves the state-of-the-art performance on the benchmark datasets CUHK03 (Li et al. 2014), Market-1501 (Zheng, Shen, and others 2015), MSMT17 (Wei, Zhang, and others 2018), Partial REID (Zheng et al. 2015).

2 Related Work

Person reID based on deep neural networks has made great progress in recent years. Due to the variations in poses, viewpoints, incompleteness of the visible bodies (due to occlusion), *etc.*, across the images, semantics misalignment is still one of the key challenges.

Alignment with Pose/Part Cues for ReID: To address the spatial semantics misalignment, most of the previous approaches make use of external cues such as pose/part (Li et al. 2017; Yao et al. 2017; Zhao et al. 2017; Kalayeh et al. 2018; Zheng et al. 2017; Su et al. 2017; Suh et al. 2018). Human landmark (pose) information can help align body regions across images. Zhao *et al.* (Zhao et al. 2017) propose a human body region guided Splindle Net, where a body region proposal sub-network (trained with the human pose dataset) is used to extract the body regions, *e.g.*, head-shoulder, arm region. The semantic features from different body regions are separately captured thus the body part features can be aligned across images. Kalayeh *et al.* (Kalayeh et al. 2018) integrate a human semantic parsing branch in their network for generating probability maps associated to different semantic regions of human body, *e.g.*, head, upper-body. Based on the probability maps, the features from different semantic regions of human body are aggregated separately to have part aligned features. Qian *et al.* (Qian et al. 2018) propose to make use of GAN model to synthesize realistic person images of eight canonical poses for matching. However, these approaches usually require pose/part detection or image generation sub-networks, and extra computational cost in inference. Moreover, the alignment based on pose is coarse without considering the finer grained alignment within a part across images.

Zhang *et al.* (Zhang et al. 2019) exploit the dense semantics from DensePose (Alp Güler, Neverova, and Kokkinos 2018) rather than the coarse pose for reID. Their network consists of two streams in training: a main stream takes the original image as input while the other stream learns features from the warped images for regularizing the feature learning of the main stream. However, the invisible body regions result in many holes in the warped images and inconsistency of visible body regions across images, which could hurt the learning efficiency. Moreover, there is a lack of more direct constraints to enforce the alignment. The design of efficient frameworks for dense semantics alignment is still under-explored. In this paper, we propose an elegant framework which adds direct constraints to encourage dense semantics alignment in feature learning.

Figure 3: Examples of texture images (first row) and the corresponding synthesized person images with different poses, viewpoints, and backgrounds (second row). A texture image represents the full texture of the 3D human surface in a surface-based canonical coordinate system (UV space). Each position (u,v) corresponds to a unique semantic identity. For person images of different persons/poses/viewpoints (in the second row), their corresponding texture images are densely semantically aligned.

Semantics Aligned Human Texture: A human body could be represented by a 3D mesh (*e.g*. Skinned Multi-Person Linear Model, SMPL (Loper et al. 2015)) and a texture image (Varol et al. 2017; Hormann, Lévy, and Sheffer 2007) as illustrated in Figure 4. Each position on the 3D body surface has a semantic identity (identified by a 2D coordinate (u,v) in the canonical UV space) and a texture representation (*e.g*. RGB pixel value) (Güler, Neverova, and Kokkinos 2018; Güler et al. 2017). A texture image on the UV coordinate system (*i.e*., surface-based coordinate system) represents the *aligned full* texture of the 3D surface of the person. Note that the texture images across different persons are densely semantically aligned (see Figure 3). In (Güler, Neverova, and Kokkinos 2018), a dataset with labeled dense semantics (*i.e*. DensePose) is established and a CNN-based system is designed to estimate DensePose from person images. Neverova *et al*. (Neverova, Alp Guler, and Kokkinos 2018) and Wang *et al*. (Wang et al. 2019) leverage the aligned texture image to synthesize person image of another pose or view. Yao *et al*. (Yao et al. 2019) propose to regress the 3D human body ((x,y,z) coordinates in 3D space) in the semantics aligned UV space, with the RGB person image as the input to the CNN.

Different from all these works, we leverage the densely semantically aligned full texture image to address the misalignment problem in person reID. We use them as direct supervisions to drive the reID network to learn semantics aligned features.

3 The Semantics Aligning Network (SAN)

To address the cross image misalignment challenge caused by human pose, capturing viewpoint variations, and the incompleteness of the body surface (due to the occlusion when projecting 3D person to 2D person image), we propose a Semantics Aligning Network (SAN) for robust person reID, in which densely semantically aligned full texture images are taken as supervision to drive the learning of semantics aligned features.

The proposed framework is shown in Figure 2. It consists of a base network as encoder (SA-Enc) for reID, and a decoder sub-network (SA-Dec) (see Sec. 3.2) for generating densely semantically aligned full texture image with supervision. This encourages the reID network to learn semantics aligned feature representation. Since there is no groundtruth texture image of 3D human surface for the reID datasets, we use our synthesized data based on (Varol et al. 2017) to train SAN (with reID supervisions removed) which is then used to generate pseudo groundtruth texture images for the reID datasets (see Sec. 3.1).

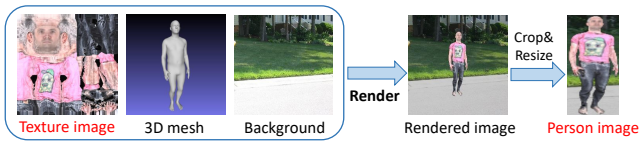

Figure 4: Illustration of the generation of synthesized person image to form a (*person image*, *texture image*) pair. Given a texture image, a 3D mesh, a background image, and rendering parameters, we can obtain a 2D person image through the rendering.

The reID feature vector $\mathbf{f}$ is obtained by average pooling the last layer feature map f_{e4} of the SA-Enc, followed by the reID losses. The SA-Dec is added after the last layer of the SA-Enc to regress densely semantically aligned texture image, with the (pseudo) groundtruth texture supervision. In the SA-Dec, Triplet ReID constraints are further incorporated at different layers/blocks as the high level perceptual metric to encourage identity preserving reconstruction. During inference, the SA-Dec is discarded.

3.1 Densely Semantically Aligned Texture Image

Background: The person texture image in the surface-based coordinate system (UV space) is widely used in the graphics field (Hormann, Lévy, and Sheffer 2007). Texture images for different persons/viewpoints/poses are densely semantically aligned, as illustrated in Figure 3. Each position (u,v) corresponds to a unique semantic identity on the texture image, *e.g*., the pixel on the right bottom of the texture image corresponds to some semantics of a hand. Besides, a texture image contains all the texture of the full 3D surface of a person. In contrast, only a part of the surface texture is visible/projected on a 2D person image.

Motivation: We intend to leverage such aligned texture images to drive the reID network to learn semantics aligned features. For different input person images, the corresponding texture images are well semantics aligned. First, for the same spatial positions on different texture images, the semantics are the same. Second, for person images with different visible semantics/regions, their texture images are semantics consistent/aligned since each one contains the full texture/information of the 3D person surface.

Pseudo groundtruth Texture Images Generation: For the images in the reID datasets, however, there are no groundtruth aligned full texture images. We propose to train the SAN using our synthesized data to enable the generation of a pseudo groundtruth texture image for each image in the reID datasets. We can leverage a CNN-based net-

work to generate pseudo groundtruth texture images. In this work, we reuse the proposed SAN (with the reID supervisions removed) as the network (see Figure 2), which we refer to as SAN-PG (Semantics Aligning Network for Pseudo Groundtruth Generation) for differentiation. Given an input person image, the SAN-PG outputs predicted texture image as the pseudo groundtruth.

To train the SAN-PG, we synthesize a Paired-Image-Texture dataset (PIT dataset), based on SURREAL dataset (Varol et al. 2017), for the purpose of providing the image pairs, *i.e*., the *person image* and *its texture image*. The texture image stores the RGB texture of the *full* person 3D surface. As illustrated in Figure 4, given a texture image, a 3D mesh/shape, and a background image, a 2D projection of a 3D person can be obtained by rendering (Varol et al. 2017). We can control the pose and body form of the person, and projection viewpoint, through changing the parameters of 3D mesh/shape model (*i.e*. SMPL (Loper et al. 2015)) and the rendering parameters. Note that we do not include identity information in the PIT dataset.

To generate the PIT dataset with paired person images and texture images, in particular, we use 929 (451 for female and 478 for male) raster-scanned texture maps provided by the SURREAL dataset (Varol et al. 2017) to generate the *person image* and *texture image* pairs. These texture images are aligned with the SMPL default two-dimensional UV coordinate space (UV space). The same uv coordinate value corresponds to the same semantics. We generate 9,290 different meshes of diverse poses/shapes/viewpoints, by using SMPL body model (Loper et al. 2015) parameters inferred by HMR (Kanazawa et al. 2018) from the person images of the COCO dataset (Lin et al. 2014). For each texture map, we assign 10 different meshes and render these 3D meshes with the texture image by Neural Render (Kato, Ushiku, and Harada 2018). Then we obtain in total 9,290 different synthesized (*person image*, *texture image*) pairs. To simulate real-world scenes, the background images for rendering are randomly sampled from COCO dataset (Lin et al. 2014). Each synthetic person image is centered on a person with resolution 256×128. The resolution of the texture images is 256×256.

Discussion: The texture images which we use for supervisions have three major advantages. 1) They are spatially aligned in terms of the dense semantics of a person surface and thus can guide the reID network to learn semantics aligned representation. 2) A texture image containing the *full* 3D surface of a person can guide the reID network to learn more comprehensive representation of a person. 3) They represent the textures of the human body surface and thus naturally eliminate the interference of diverse background scenes.

There are also some limitations of the current pseudo groundtruth texture image generation process. 1) There is a domain gap between synthetic 2D images (in the PIT dataset) and real-world captured images where the synthetic person is not very realistic. 2) The number of texture images provided by SURREAL (Varol et al. 2017) is not large (*i.e*. 929 in total) which may constraint the diversity of the data in our synthesized dataset. 3) On SURREAL, all faces in the texture image are replaced by an average face of either man or woman (Lin et al. 2014). We leave it as future work to address these limitations. Even with such limitations, our scheme achieves significant performance improvement over the baseline on person reID.

3.2 SAN and Optimization

As illustrated in Figure 2, the SAN consists of an encoder SA-Enc for person reID, and a decoder SA-Dec which enforces constraints over the encoder by requiring the encoded features to be able to predict/regress the semantically aligned full texture images.

SA-Enc: We can use any baseline network used in person reID (*e.g*. ResNet-50 (Sun et al. 2018; Zhang et al. 2017; Zhang et al. 2019)) as the SA-Enc. In this work, we similarly use ResNet-50 and it consists of four residual blocks. The output feature map of the fourth block $f_{e4} \in \mathbb{R}^{h \times w \times c}$ is spatially average pooled to get the feature vector ($\mathbf{f} \in \mathbb{R}^{c}$), which is the reID feature for matching.

For the purpose of reID, on the feature vector $\mathbf{f}$, we add the widely-used identification loss (ID Loss) $\mathcal{L}_{ID}$, *i.e*., the cross entropy loss for identification classification, and the ranking loss of triplet loss with batch hard mining (Hermans, Beyer, and Leibe 2017) ($Triplet$ Loss) $\mathcal{L}_{Triplet}$ as the loss functions in training.

SA-Dec: To encourage the encoder features to learn semantics aligned features, we add a decoder SA-Dec after the fourth block (f_{e4}) of the encoder to regress the densely semantically aligned texture images, supervised by the (pseudo) groundtruth texture images. A reconstruction loss $\mathcal{L}_{Rec.}$ is introduced to minimize $L1$ differences between the generated texture image and its corresponding (pseudo) groundtruth texture image.

Triplet ReID constraints at SA-Dec: Besides the capability of reconstructing the texture images optimized/measured by the $L1$ distance, we also expect the features in the decoder inherit the capability of distinguishing different identities. Wang *et al*. (Wang et al. 2019) use reID network as the perceptual supervision to generate person image, which judges whether the generated person image and the real image have the same identity. Different from (Wang et al. 2019), in considering that the features at each layer of the decoder are spatially semantically aligned across images, we measure the feature distance for each spatial position rather than on the final globally pooled feature. We introduce $Triplet\ ReID$ constraints to minimize the $L2$ differences between the features of the same identity and maximize those of different identities. Specially, for a sample a in a batch, we can randomly select a positive sample p (with the same identity) and a negative sample n. The Triplet ReID constraint/loss over the output feature map of the l^{th} block of the SA-Dec is defined as

$$\mathcal{L}_{TR}^{l} = max(\frac{1}{h_l \times w_l}||f_{dl}(x_l^a) - f_{dl}(x_l^p)||_2^2 - \frac{1}{h_l \times w_l}||f_{dl}(x_l^a - f_{dl}(x_l^n)||_2^2 + m, 0), \tag{1}$$

where $h_l \times w_l$ is the resolution of feature map with c_l channels, $f_{dl}(x_l^a) \in \mathbb{R}^{h_l \times w_l \times c_l}$ denotes the feature map of sample a. $||f_{dl}(x_l^a) - f_{dl}(x_l^p)||_2^2 = \sum_{i=1}^{h_l} \sum_{j=1}^{w_l} ||f_{dl}(x_l^a)(i, j, :$

$) - f_{dl}(x_l^p)(i,j,:)||_2^2$ with $f_{dl}(x_l^a)(i,j,:)$ denotes the feature vector of c_l channels at spatial position (i,j). The margin parameter m is set to 0.3 experimentally

Training Scheme: There are two steps for training our proposed SAN framework for reID:

Step-1, we train a network for the purpose of generating pseudo groundtruth texture images for any given input person image. For simplicity, we reuse a simplified SAN (*i.e.*, SAN-PG) which consists of the SA-Enc and SA-Dec, but with only the reconstruction loss $\mathcal{L}_{Rec.}$. We train the SAN-PG with our synthesized PIT dataset. The SAN-PG model is then used to generate pseudo groundtruth texture image for reID datasets (such as CUHK03 (Li et al. 2014)).

Step-2, we train the SAN for both reID and aligned texture generation. The pre-trained weights of the SAN-PG are used to initialize the SAN. One alternative is to use only the reID dataset for training SAN, where the pseudo groundtruth texture images are used for supervision and all the losses are added. The other strategy is to iteratively use the reID dataset and the synthesized PIT dataset during training. We find the later solution gives superior results because the groundtruth texture images for the synthesized PIT dataset have higher quality than that of reID dataset. The overall loss $\mathcal{L}$ consists of the ID Loss $\mathcal{L}_{ID}$, the $Triplet$ Loss $\mathcal{L}_{Triplet}$, the reconstruction loss $\mathcal{L}_{Rec.}$, and the $Triplet\ ReID$ constraint $\mathcal{L}_{TR}$, *i.e.*, $\mathcal{L}$ = $\lambda_1\mathcal{L}_{ID}$ + $\lambda_2\mathcal{L}_{Triplet}$ + $\lambda_3\mathcal{L}_{Rec.}$ + $\lambda_4\mathcal{L}_{TR}$. For a batch of reID data, we experimentally set λ_1 to λ_4 as 0.5, 1.5, 1, 1. For a batch of synthesized data, λ_1 to λ_4 are set to 0, 0, 1, 0 where the reID losses and Triplet ReID constraints (losses) are not used.

4 Experiment

4.1 Datasets and Evaluation Metrics

We conduct experiments on six benchmark person reID datasets, including CUHK03 (Li et al. 2014), Market1501 (Zheng, Shen, and others 2015), DukeMTMC-reID (Zheng, Zheng, and Yang 2017), the large-scale MSMT17 (Wei, Zhang, and others 2018), and two challenging partial person reID datasets of Partial REID (Zheng et al. 2015) and Partial-iLIDS (He et al. 2018)

We follow the common practices and use the cumulative matching characteristics (CMC) at Rank-k, k = 1, 5, 10, and mean average precision (mAP) to evaluate the performance.

4.2 Implementation Details

We use ResNet-50 (He et al. 2016) (which are widely used in some re-ID systems (Sun et al. 2018; Zhang et al. 2019)) to build our SA-Enc. We also take it as our baseline (Baseline) with both ID loss and triplet loss. Similar to (Sun et al. 2018; Zhang et al. 2019), the last spatial down-sample operation in the last *Conv* layer is removed. We build a light weight decoder SA-Dec by simply stacking 4 residual up-sampling blocks with about 1/3 parameters of the SA-Enc. This facilitates our model training using only a single GPU.

4.3 Ablation Study

We perform comprehensive ablation studies to demonstrate the effectiveness of the designs in the SAN framework, on the datasets of CUHK03 (labeled bounding box setting) and Market-1501 (single query setting).

Table 1: Comparisons (%) of our SAN and baseline.

Model	CUHK03(L)		Market1501	
	Rank-1	mAP	Rank-1	mAP
Baseline (ResNet-50)	73.7	69.8	94.1	83.2
SAN-basic	77.9	73.7	95.1	85.8
SAN w/ $\mathcal{L}_{TR}$	78.9	74.9	95.4	86.9
SAN w/ syn. data	78.8	75.8	95.7	86.8
SAN	**80.1**	**76.4**	**96.1**	**88.0**

Effectiveness of Dense Semantics Alignment. In Table 1, ***SAN-basic*** denotes our basic semantics aligning model which is trained with the supervision of the pseudo groundtruth texture images with loss of $\mathcal{L}_{Rec.}$, the reID losses $\mathcal{L}_{ID}$ and $\mathcal{L}_{Triplet}$. ***SAN w/*$\mathcal{L}_{TR}$** denotes that the Triplet ReID constraints at the SA-Dec is added on top of the *SAN-basic*. ***SAN w/syn. data*** denotes that the (*person image*, *texture image*) pairs of our PIT dataset is also used in training the SAN on top of the *SAN-basic* network. ***SAN*** denotes our final scheme with both the Triplet ReID constraints and the groundtrth texture image supervision from the PIT on top of the *SAN-basic* network.

We have the following observations/conclusions. **1)** Thanks to the drive to learn semantics aligned features, our *SAN-basic* significantly outperforms the baseline scheme by about 4% in both Rank-1 and mAP accuracy on CUHK03. **2)** The introduction of high-level Triplet ReID constraints ($\mathcal{L}_{TR}$) as the perceptual loss can regularize the feature learning and it brings about additional 1.0% and 1.2% improvements in Rank-1 and mAP accuracy on CUHK03. Note that we add them after each block of the first three blocks in the SA-Dec. **3)** The use of the synthesized PIT dataset (syn. data) with the input image and groundtruth texture image pairs for training the SAN remedies the imperfection of the generated pesudo groundtruth texture images (with errors/noise/blurring). It improves the performance over *SAN-basic* by 0.9% and 2.1% in Rank-1 and mAP accuracy. **4)** Our final scheme *SAN* significantly outperforms the baseline, *i.e.*, by **6.4%** and **6.6%** in Rank-1 and mAP accuracy on CUHK03, but with the same inference complexity. On Market1501, even though the baseline performance is already very high, our *SAN* achieves 2.0% and 4.8% improvement in Rank-1 and mAP.

Different Reconstruction Guidance. We study the effect of using different reconstruction guidance and show results in Table 2. We design another two schemes for comparisons. For the same input image, the three schemes use the same encoder-decoder networks (the same network as SAN-basic) but to reconstruct (a) the input person image, (b) pose aligned person image, and (c) proposed texture image (see Figure 5). To have pose aligned person image as supervision, during synthesizing the PIT dataset, for each projected person image, we also synthesized a person image of a given fixed pose (frontal pose here). Thus, the pose aligned person images are also semantically aligned. In this case, only

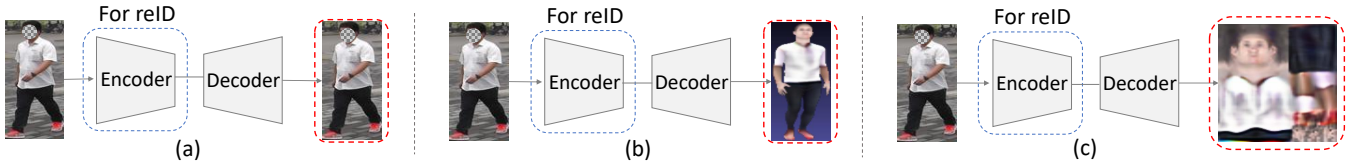

Figure 5: The same encoder-decoder networks but with different reconstruction objectives of reconstructing the (a) input image, (b) pose aligned person image, and (c) texture image, respectively.

Table 2: Performance (%) comparisons of the same encoder-decoder networks but with different reconstruction objectives of reconstructing the input image, pose aligned person image, and texture image respectively.

Model	CUHK03(L)		Market1501	
	Rank-1	mAP	Rank-1	mAP
Baseline (ResNet-50)	73.7	69.8	94.1	83.2
Enc-Dec rec. input	74.4	70.8	94.3	84.0
Enc-Dec rec. pose	75.8	72.0	94.4	84.5
Enc-Dec rec. PN-GAN pose	76.1	72.6	94.3	84.7
Enc-Dec rec. texture (SAN-basic)	**77.9**	**73.7**	**95.1**	**85.8**

partial texture (frontal body regions) of the full 3D surface texture is retained with information loss. In addition, corresponding to (b), we also use the pose aligned person images generated by PN-GAN (Qian et al. 2018) as the reconstruction guidance and get *Enc-Dec rec. PN-GAN pose*.

From Table 2, we have the following observations/conclusions. 1) The addition of a reconstruction subtask helps improve the reID performance which encourages the encoded feature to preserve more original information. *Enc-Dec rec. input* improves the performance of the baseline by 0.7% and 1.0% in Rank-1 and mAP accuracy. However, the input images (and their reconstructions) are not semantically aligned across images. 2) *Enc-Dec rec. pose* and *Enc-Dec rec. PN-GAN pose* both enforce the supervision to be *pose aligned person images*. This has a superior performance to *Enc-Dec rec. input*, demonstrating the effectiveness of **alignment**. But they are sub-optimal which may lose information. For example, for an input back-facing person image, such fixed (frontal) pose supervision may mistakenly guide the features to drop the back-facing body information. 3) In contrast, our full aligned texture image as supervision can provide comprehensive and densely semantics aligned information, which results in the best performance.

Why not *Directly* use Generated Texture Image for ReID? How about the performance when the generated texture images are used as the input for reID? Results show that our scheme significantly outperforms them. The inferior performance is caused by the low quality of the generated texture image (with the texture smoothed/blurred).

How does the Quality of Textures affect reID Performance? We use different backbone networks, *e.g.*, ResNet-101, DenseNet-121, *etc.*, to train the pseudo texture generators, and then the generated pseudo textures are used to train our SAN-basic network for reID. We find that using deeper and more complex generators can improve the texture quality, which in turn further boosts the reID performance.

Figure 6: Two sets of examples of the pairs. Each pair corresponds to the original input image and the generated texture image.

4.4 Comparison with State-of-the-Arts

Table 3 shows the performance comparisons of our proposed SAN with the state-of-the-art methods. Our scheme SAN achieves the best performance on CUHK03, Market1501, and MSMT17. It consistently outperforms the approach *DSA-reID* (Zhang et al. 2019) which also considers the dense alignment. On the DukeMTMC-reID dataset, *MGN* (Wang et al. 2018b) achieves better performance, however, it ensembles the local features of multiple granularities and the global features.

4.5 Visualization of Generated Texture Image

For the different images with varied poses, viewpoints, or scales, we find the generated texture images from our SAN are well semantically aligned (see Figure 6).

4.6 Partial Person ReID

Partial person reID is more challenging as the misalignment problem is more severe, where two partial person images are generally not spatially semantics aligned and usually have less overlapped semantics. We also demonstrate the effectiveness of our scheme on the challenging partial person reID datasets of Partial REID (Zheng et al. 2015) and Partial-iLIDS (He et al. 2018).

Benefiting from the *aligned full* texture generation capability, our SAN exhibits outstanding performance. Figure 7 shows our regressed texture images from the SA-Dec are semantically aligned across images even though the input images have severe misalignment.

Table 4 shows the experimental results. Note that we train SAN on the Market1501 dataset (Zheng, Shen, and others 2015) and test on the partial datasets. We directly take the trained model for Market1501 for testing, *i.e.*, Baseline (ResNet-50), SAN. In this case, the network seldom

Table 3: Performance (%) comparisons with the state-of-the-art methods. Bold numbers denote the best performance, while the numbers with underlines denote the second best.

	Method	CUHK03 Labeled		CUHK03 Detected		Market1501		DukeMTMC-reID		MSMT17	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Pose /Part /Mask -related	IDE (ECCV) (Sun et al. 2018)	43.8	38.9	-	-	85.3	68.5	73.2	52.8	-	-
	MGN (ACMMM) (Wang et al. 2018b)	68.0	67.4	66.8	66.0	95.7	86.9	**88.7**	**78.4**	-	-
	AACN (CVPR) (Xu et al. 2018)	-	-	-	-	85.9	66.9	76.8	59.3	-	-
	MGCAM (CVPR) (Song et al. 2018)	50.1	50.2	46.7	46.9	83.8	74.3	-	-	-	-
	MaskReID (ArXiv) (Qi et al. 2018)	-	-	-	-	90.0	70.3	78.9	61.9	-	-
	SPReID (CVPR) (Kalayeh et al. 2018)	-	-	-	-	92.5	81.3	84.4	71.0	-	-
	Pose Transfer (CVPR) (Liu et al. 2018)	33.8	30.5	30.1	28.2	87.7	68.9	68.6	48.1	-	-
	PSE (CVPR) (Sarfraz et al. 2018)	-	-	30.2	27.3	87.7	69.0	79.8	62.0	-	-
	PN-GAN (ECCV) (Qian et al. 2018)	-	-	-	-	89.4	72.6	73.6	53.2	-	-
	Part-Aligned (ECCV) (Suh et al. 2018)	-	-	-	-	91.7	79.6	84.4	69.3	-	-
	PCB+RPP (ECCV) (Sun et al. 2018)	63.7	57.5	-	-	93.8	81.6	83.3	69.2	-	-
Attention -based	DuATM (CVPR) (Si et al. 2018)	-	-	-	-	91.4	76.6	81.8	64.6	-	-
	Mancs (ECCV) (Wang et al. 2018a)	69.0	63.9	65.5	60.5	93.1	82.3	84.9	71.8	-	-
	FD-GAN (NIPS) (Ge et al. 2018)	-	-	-	-	90.5	77.7	80.0	64.5	-	-
	HPM (AAAI) (Fu et al. 2019)	63.9	57.5	-	-	94.2	82.7	86.6	74.3	-	-
Semantics	DSA-reID (CVPR) (Zhang et al. 2019)	78.9	75.2	78.2	73.1	95.7	87.6	86.2	74.3	-	-
Others	GoogLeNet (CVPR) (Wei, Zhang, and others 2018)	-	-	-	-	-	-	-	-	47.6	23.0
	PDC (CVPR) (Wei, Zhang, and others 2018)	-	-	-	-	-	-	-	-	58.0	29.7
	GLAD (CVPR) (Wei, Zhang, and others 2018)	-	-	-	-	-	-	-	-	61.4	34.0
Ours	Baseline (ResNet-50)	73.7	69.8	69.7	66.1	94.1	83.2	85.9	71.8	73.8	47.2
	SAN	**80.1**	**76.4**	**79.4**	**74.6**	**96.1**	**88.0**	87.9	75.5	**79.2**	**55.7**

Table 4: Partial person reID performance on the datasets of Partial REID and Partial-iLIDS (partial images are used as the probe set and holistic images are used as the gallery set). "*" means that the network is fine-tuned with holistic and partial person images from Market1501.

Model	Partial REID			Partial-iLIDS		
	Rank-1	Rank-5	Rank-10	Rank-1	Rank-5	Rank-10
AMC+SWM	36.0	-	-	49.6	-	-
DSR (single-scale)*	39.3	-	-	51.1	-	-
DSR (multi-scale)*	43.0	-	-	**54.6**	-	-
Baseline (ResNet-50)	37.8	65.0	74.5	42.0	65.5	73.2
SAN	39.7	67.5	80.5	46.9	71.2	78.2
Baseline (ResNet-50)*	38.9	67.7	78.2	46.1	69.6	76.1
SAN*	**44.7**	**72.4**	**86.0**	53.7	**77.4**	**81.9**

sees partial person data. Similar to (He et al. 2018), we also fine-tune with the holistic and partial person images cropped from Market1501 (marked by *). SAN* outperforms *Baseline**, AMC+SWM (Zheng et al. 2015) and is comparable with the state-of-the-art partial reID method DSR (He et al. 2018). SAN* outperforms Baseline (ResNet-50)* by 5.8%, 4.7%, 7.8% on Rank-1, Rank-5, and Rank-10 respectively on the Partial REID dataset, and by 7.6%, 7.8%, 5.8% on Rank-1, Rank-5, and Rank-10 respectively on the other Partial-iLIDS dataset. Even without fine-tune, our SAN also significantly outperforms the baseline.

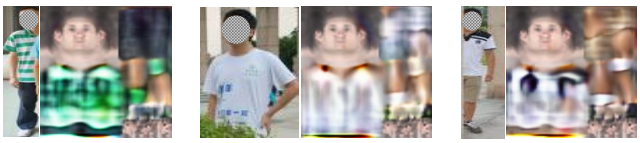

Figure 7: Three example pairs of (input image, regressed texture images by our SAN) from the Partial REID dataset.

5 Conclusion

In this paper, we proposed a simple yet powerful Semantics Aligning Network (SAN) for learning semantics-aligned feature representations for efficient person reID, under the joint supervisions of person reID and semantics aligned texture generation. At the decoder, we add Triplet ReID constraints over the feature maps as the perceptual loss to regularize the learning. We have synthesized a Paired-Image-Texture dataset (PIT) to train a SAN-PG model, with the purpose to generate pseudo groundtruth texture images for the reID datasets, and to train the SAN. Our SAN achieves the state-of-the-art performances on the datasets CUHK03, Market1501, MSMT17, and the Partial REID, without increasing computational cost in inference.

6 Acknowledgments

This work was supported in part by NSFC under Grant 61571413, 61632001.

References

[Alp Güler, Neverova, and Kokkinos 2018] Alp Güler, R.; Neverova, N.; and Kokkinos, I. 2018. Densepose: Dense human pose estimation in the wild. In *CVPR*, 7297–7306.

[Fu et al. 2019] Fu, Y.; Wei, Y.; Zhou, Y.; et al. 2019. Horizontal pyramid matching for person re-identification. In *AAAI*, volume 33, 8295–8302.

[Ge et al. 2018] Ge, Y.; Li, Z.; Zhao, H.; et al. 2018. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In *NeurIPS*.

[Güler et al. 2017] Güler, R. A.; Trigeorgis, G.; Antonakos, E.; Snape, P.; Zafeiriou, S.; and Kokkinos, I. 2017. Densereg: Fully convolutional dense shape regression in-the-wild. In *CVPR*.

[Güler, Neverova, and Kokkinos 2018] Güler, R. A.; Neverova, N.; and Kokkinos, I. 2018. DensePose: Dense human pose estimation in the wild. *CVPR*.

[He et al. 2016] He, K.; Zhang, X.; Ren, S.; et al. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

[He et al. 2018] He, L.; Liang, J.; Li, H.; and Sun, Z. 2018. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *CVPR*, 7073–7082.

[Hermans, Beyer, and Leibe 2017] Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

[Hormann, Lévy, and Sheffer 2007] Hormann, K.; Lévy, B.; and Sheffer, A. 2007. Mesh parameterization: Theory and practice.

[Kalayeh et al. 2018] Kalayeh, M. M.; Basaran, E.; Gökmen, M.; et al. 2018. Human semantic parsing for person re-identification. In *CVPR*.

[Kanazawa et al. 2018] Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *CVPR*.

[Kato, Ushiku, and Harada 2018] Kato, H.; Ushiku, Y.; and Harada, T. 2018. Neural 3d mesh renderer. In *CVPR*, 3907–3916.

[Li et al. 2014] Li, W.; Zhao, R.; Tian, L.; et al. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 152–159.

[Li et al. 2017] Li, D.; Chen, X.; Zhang, Z.; et al. 2017. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*.

[Lin et al. 2014] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755.

[Liu et al. 2018] Liu, J.; Ni, B.; Zhuang, Y.; et al. 2018. Pose transferrable person re-identification. In *CVPR*.

[Loper et al. 2015] Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. Smpl: A skinned multi-person linear model. *TOG*.

[Neverova, Alp Guler, and Kokkinos 2018] Neverova, N.; Alp Guler, R.; and Kokkinos, I. 2018. Dense pose transfer. In *ECCV*, 123–138.

[Qi et al. 2018] Qi, L.; Huo, J.; Wang, L.; Shi, Y.; and Gao, Y. 2018. Maskreid: A mask based deep ranking neural network for person re-identification. *arXiv preprint arXiv:1804.03864*.

[Qian et al. 2018] Qian, X.; Fu, Y.; Wang, W.; et al. 2018. Pose-normalized image generation for person re-identification. In *ECCV*.

[Sarfraz et al. 2018] Sarfraz, M. S.; Schumann, A.; Eberle, A.; and Stiefelhagen, R. 2018. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*.

[Shen et al. 2015] Shen, Y.; Lin, W.; Yan, J.; et al. 2015. Person re-identification with correspondence structure learning. In *ICCV*.

[Si et al. 2018] Si, J.; Zhang, H.; Li, C.-G.; Kuen, J.; Kong, X.; Kot, A. C.; and Wang, G. 2018. Dual attention matching network for context-aware feature sequence based person re-identification. In *CVPR*.

[Song et al. 2018] Song, C.; Huang, Y.; Ouyang, W.; et al. 2018. Mask-guided contrastive attention model for person re-identification. In *CVPR*.

[Su et al. 2017] Su, C.; Li, J.; Zhang, S.; et al. 2017. Pose-driven deep convolutional model for person re-identification. In *ICCV*.

[Subramaniam, Chatterjee, and Mittal 2016] Subramaniam, A.; Chatterjee, M.; and Mittal, A. 2016. Deep neural networks with inexact matching for person re-identification. In *NeurIPS*, 2667–2675.

[Suh et al. 2018] Suh, Y.; Wang, J.; Tang, S.; et al. 2018. Part-aligned bilinear representations for person re-identification. In *ECCV*.

[Sun et al. 2018] Sun, Y.; Zheng, L.; Yang, Y.; et al. 2018. Beyond part models: Person retrieval with refined part pooling. In *ECCV*.

[Varior et al. 2016] Varior, R. R.; Shuai, B.; Lu, J.; Xu, D.; and Wang, G. 2016. A siamese long short-term memory architecture for human re-identification. In *ECCV*.

[Varol et al. 2017] Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M. J.; Laptev, I.; and Schmid, C. 2017. Learning from synthetic humans. In *CVPR*, 109–117.

[Wang et al. 2018a] Wang, C.; Zhang, Q.; Huang, C.; et al. 2018a. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *ECCV*.

[Wang et al. 2018b] Wang, G.; Yuan, Y.; Chen, X.; et al. 2018b. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 274–282.

[Wang et al. 2019] Wang, J.; Zhong, Y.; Li, Y.; Zhang, C.; and Wei, Y. 2019. Re-identification supervised texture generation. In *CVPR*.

[Wei et al. 2017] Wei, L.; Zhang, S.; Yao, H.; et al. 2017. Glad: global-local-alignment descriptor for pedestrian retrieval. In *ACM MM*.

[Wei, Zhang, and others 2018] Wei, L.; Zhang, S.; et al. 2018. Person transfer GAN to bridge domain gap for person re-identification. In *CVPR*, 79–88.

[Xu et al. 2018] Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; and Ouyang, W. 2018. Attention-aware compositional network for person re-identification. *arXiv preprint arXiv:1805.03344*.

[Yao et al. 2017] Yao, H.; Zhang, S.; Zhang, Y.; Li, J.; and Tian, Q. 2017. Deep representation learning with part loss for person re-identification. *arXiv preprint arXiv:1707.00798*.

[Yao et al. 2019] Yao, P.; Fang, Z.; Wu, F.; Feng, Y.; and Li, J. 2019. Densebody: Directly regressing dense 3d human pose and shape from a single color image. *arXiv preprint arXiv:1903.10153*.

[Zhang et al. 2017] Zhang, X.; Luo, H.; Fan, X.; Xiang, W.; Sun, Y.; Xiao, Q.; Jiang, W.; Zhang, C.; and Sun, J. 2017. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*.

[Zhang et al. 2019] Zhang, Z.; Lan, C.; Zeng, W.; et al. 2019. Densely semantically aligned person re-identification. In *CVPR*.

[Zhao et al. 2017] Zhao, H.; Tian, M.; Sun, S.; et al. 2017. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*.

[Zheng et al. 2015] Zheng, W.-S.; Li, X.; Xiang, T.; Liao, S.; Lai, J.; and Gong, S. 2015. Partial person re-identification. In *ICCV*, 4678–4686.

[Zheng et al. 2017] Zheng, L.; Huang, Y.; Lu, H.; and Yang, Y. 2017. Pose invariant embedding for deep person re-identification. *arXiv preprint arXiv:1701.07732*.

[Zheng, Shen, and others 2015] Zheng, L.; Shen, L.; et al. 2015. Scalable person re-identification: A benchmark. In *ICCV*, 1116–1124.

[Zheng, Zheng, and Yang 2017] Zheng, Z.; Zheng, L.; and Yang, Y. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 3754–3762.

[Zheng, Zheng, and Yang 2018] Zheng, Z.; Zheng, L.; and Yang, Y. 2018. Pedestrian alignment network for large-scale person re-identification. *TCSVT*.