Re-Identification Supervised Texture Generation

Jian Wang[1]*, Yunshan Zhong[2]*, Yachun Li[3]*, Chi Zhang[4], and Yichen Wei[4]

[1]State Key Lab. of Computer Science, ISCAS & University of Chinese Academy of Sciences
[2]Peking University [3]Zhejiang University [4]Megvii Technology
wangj@ios.ac.cn, Zhongyunshan@pku.edu.cn, liyachun@zju.edu.cn
{zhangchi,weiyichen}@megvii.com

Abstract

The estimation of 3D human body pose and shape from a single image has been extensively studied in recent years. However, the texture generation problem has not been fully discussed. In this paper, we propose an end-to-end learning strategy to generate textures of human bodies under the supervision of person re-identification. We render the synthetic images with textures extracted from the inputs and maximize the similarity between the rendered and input images by using the re-identification network as the perceptual metrics. Experiment results on pedestrian images show that our model can generate the texture from a single image and demonstrate that our textures are of higher quality than those generated by other available methods. Furthermore, we extend the application scope to other categories and explore the possible utilization of our generated textures.

1. Introduction

The automatic generation of realistic 3D human models is crucial for many applications, including virtual reality, animation, video editing and clothes try-on. Among the 3D human reconstruction approaches, generating the 3D human model from a single image receives more and more attention. A lot of methods have been proposed in both traditional [28, 20, 11] and deep learning manner [25, 40, 49]. Even though these methods succeed in estimating the pose and shape of the human body accurately, the generation of texture is omitted, which is the missing part in the reconstructing the realistic 3D human body.

Even though generating human body textures from a single image is of vital importance, there are only two methods which aim at solving it. [39] first extract the partially observed textures from different images with DensePose [22] and obtain the full textures by combining the partial ones. Then [39] uses the full textures as ground truth and trains a generative network to directly infer the corresponding texture from a single image. This process is computationally expensive and requires high-quality image based dense human pose detection method for extracting the partially observed textures, which would be challenging for lots of in-the-wild images. [26] renders synthetic images with generated textures and minimize the distance between the rendered image and the input image. However, [26] uses an ImageNet-pretrained perceptual loss as the distance metric,

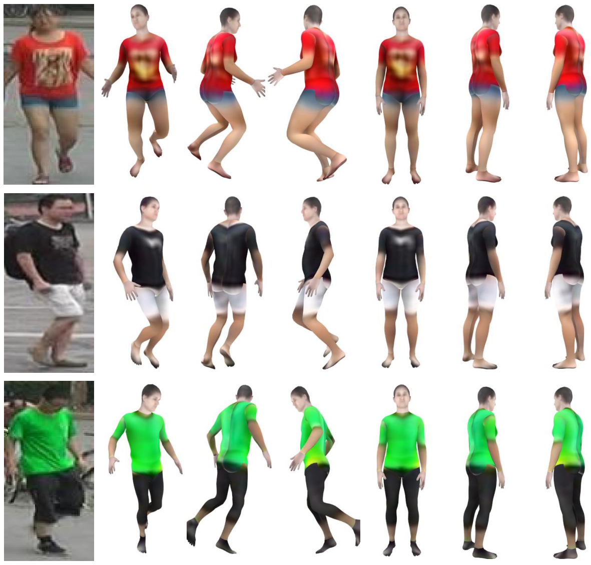

Figure 1: Texture generation results on Market-1501. This figure shows the original images (1st column), rendered 3D models in different view points (2nd-4th columns) and rendered 3D models in standing pose (5th - 7th columns). For better visual results, these 3D models are rendered with Blender [1].

*This work was done during Jian Wang, Yunshan Zhong and Yachun Li were interns at Megvii Technology.

which restricts the quality of their textures.

The shortcomings of existing works indicate that generating textures from a single image is challenging, which is caused by two reasons. Firstly, the occlusion caused by the human body makes it impossible to get the texture information from the occluded parts. Secondly, the diversity of human poses and the background complicates the texture extraction process. For example, the inaccuracy of available 3D pose estimation methods such as [25] makes directly mapping the input images to 3D models difficult.

To overcome these obstacles, we introduce the re-identification to supervise our texture generation model. Re-identification, the person identifying and retrieving method, is explicitly trained to minimize the distance between images from different viewpoints with the same identity. As the person identity is mainly characterized by textures, the re-identification network can serve as the distance metric for the textures partially observed from different viewpoints. This solves the first problem. Moreover, the re-identification network can extract the body features while eliminating the influence of pose and background variations [62], which solves the second problem. From the reasons above, it can perform well as the supervision to guide the training process of texture generation networks.

Based on the supervision of re-identification, we propose a novel method to generate body textures from a single image. Example results are shown in Fig. 1. In order to train our model in an end-to-end way, we render images with the SMPL body model and use the distance between features extracted by re-identification network as the training loss (denoted as *re-identification loss*). Our method shows the strong capability to efficiently generate body textures.

In order to demonstrate the importance of re-identification network, we compare the re-identification loss with other loss functions which are commonly used in image generation tasks. Our experiments indicate that the performance of the model surpasses others in generating body textures.

Aside from generating human body textures, we expand our method to other categories. Our method can generate better bird textures comparing with the approach presented in [26]. In addition, the diversity of generated textures is higher than available 3D-scanned textures. The experiment on the action recognition task has demonstrated that it is beneficial to pretrain the network on dataset synthesized with highly diversified textures.

In summary, our contribution can be distributed into three aspects. Firstly, we introduce a new method to generate textures from a single image by incorporating the re-identification loss. Secondly, we provide an in-depth analysis to prove the effectiveness of re-identification loss in the texture generation task. Finally, we extend our method to broader object categories and explore the potential ability of our method as a data augmentation strategy.

2. Related Work

Texture generation. The texture generation is an essential task for reconstructing realistic 3D models because the texture represents crucial information for describing and identifying object instances. Most of the recent works focus on combining texture fragments from different views. [9, 12, 27] blend multiple images into textures with various weighted average strategies. However, these methods are sensitive to noises introduced by camera poses or 3D geometry and end up with blurring and ghosting. Some other methods [51, 42, 6, 17, 29] project images to appropriate vertices and faces. These approaches alleviated the blurring and ghosting problems while they are vulnerable to texture bleeding. Warping-based methods [64, 16, 3, 10] incorporate warping refinement technologies to correct for local misalignments. Specifically, [4] apply back-project technology for 3D human body texture generation while [5] applies a semantic prior and graph optimization strategy to obtain finer details. These methods can build high-quality 3D textures, while images from different views or RGB-D sensors are required.

Aside from the multi-view based texture generation, another challenging problem is generating textures from a single image. [39] proposes a new pose transfer method which incorporates the human body texture generation module and [26] proposed a method for recovering the textures of a specific category. Their approach is either computationally expensive or suffers from the low quality of generated textures, which is indicated in Sec. 1.

Model-based 3D human pose and shape estimation. The model-based method estimates the human body pose and shape from a single image by fitting parameters of a specific body model. Earlier works like [21, 8] optimize the pose and shape parameters of SCAPE [7] under the supervision of human body key-points and silhouettes. However, SCAPE is not consistent with existing animation software, which limits its application scope.

Most of the recent works build their approaches on a simple yet powerful body model: SMPL (Skinned Multi-Person Linear Model) [35]. SMPL renders the body mesh by calculating a linear function of pose and shape parameters, which enables the optimization of SMPL model by learning from massive data. [11] designed a loss function to penalize the difference between projected 3D body joints and detected 2D joints. This loss function also prevents the inter-penetration between limbs and trunk. [28] infers the 3D shape and pose directly from 91 landmark predictions in UP-3D dataset, which accelerates the SMPL by one and two orders of magnitudes. [4] and [8] proposed a method to obtain a visual hull by transforming the silhouette cones corresponding to dynamic human silhouettes, which enables the accurate estimation of body shapes and textures.

More and more recent work applies deep learning methods to fit SMPL parameters. [47, 15, 41] predict SMPL parameters directly by a deep neuron network and get supervision from differentiable rendering of silhouettes. [48] proposed a self-supervised method using 2D human keypoints, 3D mesh motions, and human-background segmentation. [25] regresses the SMPL parameters iteratively and incorporates the prior knowledge to guarantee the reality of human shape and pose. [40] process the image to 12 semantic segmentation parts and predict the SMPL parameters from them. [49] optimizes the volumetric loss to gain higher accuracy in body shape than previous methods.

Person re-identification. Person re-identification aims at spotting a person of interest in different cameras [62]. From the independence of person re-identification task in 2006 [18], it has become increasingly popular due to its wide application. After the incorporation of CNN-based method in [57] and [31], the performance of person re-identification network is promoted drastically. For example, since the release of Market-1501 dataset [61] in 2015, the top-1 accuracy of state-of-the-art method has increased from 44.42% in [61] to 96.6% in [53].

The core idea of deep learning in person re-identification is to extract features of the person from the image [62]. Moreover, the features of different body parts provide more fine-grained information than global features, thus combining local representations from parts of images has become one of the most prevalent strategies in recent works. For example, [14, 32, 46, 53] split the image horizontally and learn local features in each part. [30, 60, 33] apply region proposal methods to extract different human parts. Attention mechanism of [34, 33] shows priority in learning soft pixel-wise parts of the body. Based on recent advances, our approach utilizes the part-based person re-identification method to represent spatial features, which plays a key role in restoring detailed textures.

3. Method

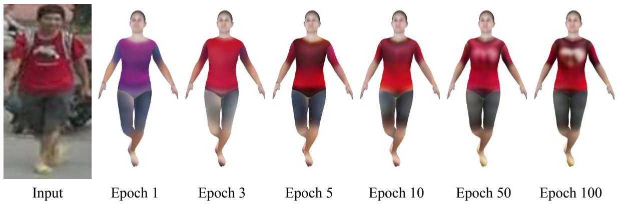

Figure 2: **Visualization of training process.** We render the image with textures generated in different training epochs.

The key idea of our model is to maximize the perceptual similarity between the input images and the images rendered with generated textures. We suppose that the person rendered with better textures shares higher similarity with the person in input images. This is shown in Fig. 2 that similarity between the rendered person and input image is increasing along with the training process.

The overall training procedure is depicted as follow. Firstly, we predict SMPL [36] parameters with HMR [25] and calculate the body mesh from predicted parameters (Sec. 3.1). Subsequently, the texture generator, U-Net [43], is used to generate the texture from a single image. The differentiable renderer, Opendr [37], is further applied to render the human body image with the generated texture (Sec. 3.2). After that, the input and rendered images are sent to a pretrained person re-identification network with part-based convolutional baseline [46] and the distance between extracted features are minimized. In addition, to make the face of the generated texture more realistic, we also minimize the difference in the face parts between generated textures and the 3D scanned textures (Sec. 3.3). The overall architecture of our method, which is trained in an end-to-end way, is shown in Fig. 3.

Recently, the generative adversarial network (GAN) [19] has been widely used in image generation tasks [24, 38] as it can generate images that look superficially authentic to human observers. However, combining the loss with a GAN-style discriminator will not work in our method. As there is an obvious style gap between the rendered images and real ones, the discriminator in GAN can always distinguish them easily, which causes the gradient of generator diminishes.

3.1. Body Mesh Reconstruction

In our method, we render our textures on the SMPL body model due to its outstanding realism and high computational efficiency. SMPL parameterizes human body mesh with shape parameters $\beta \in \mathbb{R}^{10}$, pose parameters $\theta \in \mathbb{R}^{72}$ and translation parameters $\gamma \in \mathbb{R}^3$. The shape parameters control how individuals vary in height, weight and body proportions, while the pose parameters model the 3D rotation of both the human body and the $K = 23$ joints in axis-angle representation. The translation parameters are optional as it controls the position of the human body mesh in the orthogonal coordinate system. SMPL uses a differentiable function $M(\theta, \beta, \gamma) \in \mathbb{R}^{3\times N}$ to give the triangulated body mesh with $N = 6890$ vertices.

Although the re-identification network can reduce the influence caused by variations in body pose and translation, these variations, especially the position and orientation of the human body, can still interfere with the training process, which is shown in Sec. 4.4. Thus, we still need to align the rendered person with the input image by estimating body shape, pose, and translation of the input image. To tackle

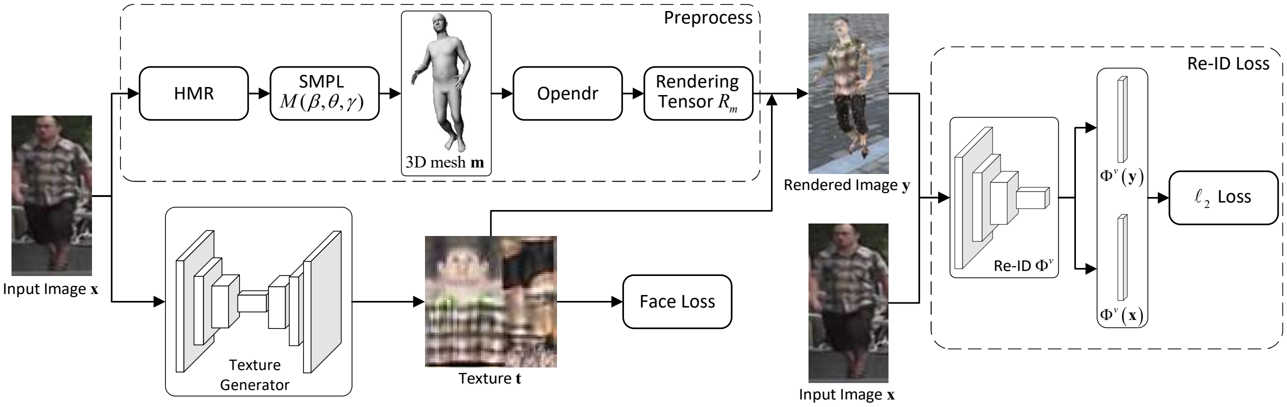

Figure 3: **Overview of the proposed framework.** Firstly, the image is sent to HMR and SMPL parameters are predicted. The 3D body mesh is calculated with HMR [25] and SMPL [36] and the rendering tensor is generated with Opendr. This step can be finished in the preprocessing procedure. Afterward, the texture is generated directly from the U-Net and the face loss of the generated texture is calculated. Finally, the rendered image $\mathbf{y}$ is generated as the product of the texture and rendering tensor. The background of $\mathbf{y}$ is randomly cropped from PRW dataset [63] and CUHK-SYSU dataset [56]. The feature of the rendered image $\mathbf{y}$ and the input image $\mathbf{x}$ (denoted as $\Phi^v(\mathbf{x})$ and $\Phi^v(\mathbf{y})$ respectively) is extracted with a pretrained person re-identification model and the ℓ_2 loss between $\Phi^v(\mathbf{x})$ and $\Phi^v(\mathbf{y})$ is calculated.

this issue, we adopt HMR, the state-of-the-art method for the 3D body pose and shape estimation. HMR produces the shape, pose and translation parameters for SMPL with an iterative 3D regression module. Thus, the estimated 3D mesh $\mathbf{m}$ from the image $\mathbf{x}$ can be expressed as follows: $\mathbf{m} = M(\beta, \theta, \gamma) = M(hmr(\mathbf{x}))$.

3.2. Texture Rendering

In this step, we generate the texture with U-Net and apply Opendr [37], a differentiable renderer, to map the generated texture to the 3D mesh. With the UV correspondence provided by SMPL, the rendering function $R(\mathbf{m}, \mathbf{t})$ of Opendr directly assigns pixels to surface on the 3D mesh polygon and fills in the gaps with linear interpolation.

With the fixed 3D human mesh $\mathbf{m}$, the rendering function $R(\mathbf{m}, \mathbf{t})$ can be viewed as a linear transformation that maps from the space of texture $\mathbf{t}$ to the space of rendered image $\mathbf{y}$:

$$R_{\mathbf{m}} = \mathbf{R}_{h_{\mathbf{t}} \times w_{\mathbf{t}} \times c \times h_{\mathbf{y}} \times w_{\mathbf{y}} \times c} \tag{1}$$

where $h_{\mathbf{t}}$ and $w_{\mathbf{t}}$ stand for the height and width of texture image, $h_{\mathbf{y}}$ and $w_{\mathbf{y}}$ stand for the height and width of rendered image and c stands for the size of image channels.

The rendering process can be simplified as the multiplication of tensors:

$$\mathbf{y} = R_{\mathbf{m}}(\mathbf{t}) = \mathbf{t} \otimes R_{\mathbf{m}} \tag{2}$$

The rendering tensor $R_{\mathbf{m}}$ will not change as long as the human 3D mesh is fixed. This provides a trick for accelerating the training procedure: we can predict all 3D meshes and calculate all rendering matrices $R_{\mathbf{m}}$ of training data in the preprocess step. In this way, we can avoid the time-consuming process of calculating rendering tensor $R_{\mathbf{m}}$ in the training phase.

3.3. Loss Functions

Re-identification loss. The re-identification loss is the layer-wise feature distance between rendered image and input image. Given a pair of input and rendered image $\{\mathbf{x}, \mathbf{y}\}$, we use the pretrained re-identification network as a feature extractor for both $\mathbf{x}, \mathbf{y}$. We penalize the ℓ_2 distance of the respective intermediate feature activations Φ^v at $n = 4$ different network layers ($v = 1, ..., n$) after the Resnet block.

$$\mathcal{L}_{reid}(\mathbf{x}, \mathbf{y}) = \sum_{v=1}^{n} \|\Phi^v(\mathbf{x}) - \Phi^v(\mathbf{y})\|_2 \tag{3}$$

This loss penalizes differences in low- mid- and high-level feature statistics, captured by respective network filters.

The setting of re-identification loss is similar to the perceptual loss which is widely used for image generation while the perceptual loss uses a network pretrained on ImageNet. However, our method outperforms the model trained on perceptual loss, which will be shown in Sec. 4.4. This is because the re-identification network has been explicitly trained to minimize the distance of the images of the same person and maximize that of the different persons. As the person identity is mostly characterized by the body texture, the re-identification network performs better for guiding the texture generation process.

In our proposed approach, we use the person re-identification model with PCB [46] because of its simplicity and efficiency in extracting features from different body parts. Other re-identification models can reach similar results while they perform badly when generating the details of the human body, which will be shown in Sec. 4.4.

Face loss. In order to improve the realism of generated texture, we design the face loss as the ℓ_1 loss of face and hand parts between the generated texture $\mathbf{t}$ and 3D scanned texture $\mathbf{t_s}$ from SURREAL [50]. Given the mask $\mathcal{M}$ of head and hand parts, the face loss is defined in the following way:

$$\mathcal{L}_{face}(\mathbf{t}, \mathbf{t_s}) = \|\mathcal{M} \odot (\mathbf{t} - \mathbf{t_s})\|_1 \tag{4}$$

The face loss makes the face part of the generated texture similar to the corresponding part in the scanned texture. The reason why we use the face loss in the training procedure rather than simply covering the generated face parts with scanned ones is that the former approach can eliminate the color contrast between head and torso. From Fig. 4 we can see the model trained without face loss produces results of low quality. If we substitute the face part with textures in SURREAL, there will be an obvious color contrast on the neck.

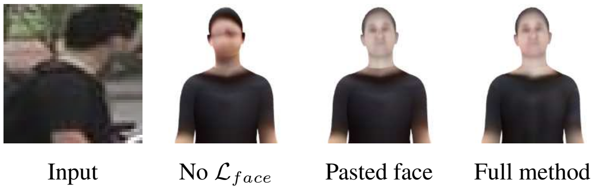

Figure 4: **Results with and without face loss.** For "Pasted face", we copy the face part of textures from SURREAL and paste it to the texture of "No $\mathcal{L}_{face}$".

In all, our overall loss function is:

$$\mathcal{L} = \lambda_{reid}\mathcal{L}_{reid} + \lambda_{face}\mathcal{L}_{face} \tag{5}$$

where the λ_{reid} and λ_{face} are the weight of re-identification loss and face loss respectively.

4. Experiments

4.1. Datasets and Metrics

Datasets. We perform our experiments on commonly-used re-identification dataset Market-1501 [61], containing 32,668 images of 1,501 identities captured from six disjoint surveillance cameras. All images are resized to 64 × 128 pixels. In our experiment, we select 100 person identities for testing and the remaining 1401 for training. We also removed all the images with unknown human labels. This results in 30,470 training and 1,747 testing images.

Metrics. Evaluating the quality of image generation method is a tricky task. In our experiments, we adopt a redundancy of metrics and a user study to evaluate the quality of generated textures. Following [38], we use the Structural Similarity (SSIM) [55], Inception Score (IS) [44] and the masked version of them: mask-SSIM and mask-IS [38].

The mask-SSIM is incorporated in order to reduce the influence of background in our evaluation. A pose mask is added to both the generated and the target image before computing SSIM. In this way, we only focus on measuring the generation quality of a person's appearance.

Though the SSIM performs well in estimating similarity both in body structure and textures, the Inception Score is not useful in our problem. This is because it only rewards the inter-class divergence and penalizes the inner-class divergence, which means that it is not relevant to the with-in class object generation [39]. We also have empirically observed its instability with respect to the perceived quality and structural similarity. Thus, we do not expect to draw strong conclusions from it.

4.2. Implementation Details

We train texture generator with the Adam optimizer (learning rate: 2×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay: 1×10^{-5}). In all experiments, the batch size is set to 16 and the training proceeds for 120 epochs, totally 64k iterations. Every batch contains four groups while each group constitutes of four images from the same person identity. The balancing weights λ between different losses (described in Sec. 3.3) are set empirically to $\lambda_{reid} = 5 \times 10^3$, $\lambda_{face} = 1.0$.

4.3. Comparison with Available Methods

In this section, we demonstrate that the re-identification plays an indispensable role in our method. We compare the re-identification loss in our approach with two commonly-used loss functions for the texture generation task: the pixel-wise ℓ_1 loss [39] and perceptual loss [26]. We show our qualitative results in Fig. 5 and quantitative results in Table. 1. From the qualitative result, we can conclude that the model trained with re-identification loss performs better than models with other loss functions. The reason why the pixel-wise ℓ_1 and perceptual loss performs bad is analyzed in the following paragraphs.

The pixel-wise ℓ_1 loss is defined as the ℓ_1 loss between the rendered and input image. The model trained with pixel-wise ℓ_1 reconstruction loss performs bad especially in generating details, such as hands and shins, of the image. This is caused by the inaccurate estimation of human body poses and shapes provided by HMR [25] module.

As described in [13], the perceptual loss is defined as the ℓ_2 distance between features of two images extracted from 5 intermediate layers of a VGG19 network which is pre-

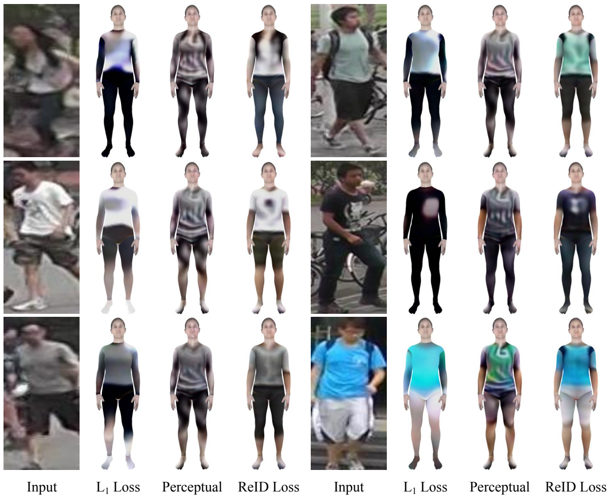

Figure 5: **Qualitative results.** The result of different loss functions is shown above. Each column shows (in order from left): input images (from test set), pixel-wise ℓ_1 loss, perceptual loss, the re-identification loss

Model	pixel-wise ℓ_1	Perceptual	ReID Loss
SSIM	0.162	0.149	**0.164**
mask-SSIM	**0.374**	0.356	0.372

Table 1: **Quantitative results.** The SSIM of our method is higher than others while our mask-SSIM score is only 0.02 less than results of pixel-wise ℓ_1 loss.

trained on ImageNet. The model trained with perceptual loss ended up with an even worse result. The perceptual quality of the torso part is poor. This is because the network trained on Imagenet tends to extract general features of objects rather than concentrating on body textures.

The SSIM and mask-SSIM score of ℓ_1 reconstruction loss are among the highest scores in all of the experiment results, this is because the ℓ_1 loss optimize the generated texture in a pixel-to-pixel way, which is equivalent to directly optimizing the SSIM score in the training process. While the re-ID loss does not directly optimize the SSIM score, the SSIM score is still high using re-ID loss. This verifies that re-ID loss is indeed effective.

The IS scores of the three models are 3.96, 4.04 and 3.96 respectively while the mask-IS scores are 2.90, 2.59 and 2.52. We do not show these results in Table 1 for the same reason in Sec. 4.1.

4.4. Ablation Study

In this section, we carry out the following experiments to explore the influence of different model settings. The qualitative result is shown in Fig. 6 while the quantitative result is shown in Table. 2.

ℓ_1 loss of deep features. In the re-identification network with PCB, the image is passed into a resnet-50 network and a pooling layer, producing the feature g of 6×256 dimensions. The deep feature loss is defined as the ℓ_1 distance between deep features g of re-identification network.

The result of deep feature loss is shown in the column labeled "Deep Feature" of Fig. 6. Compared with the proposed re-identification loss, this result is good while ignoring some details, e.g. the patterns on clothes. This is because the deep features can hardly represent the details of human texture as the deep features g can hold less information than features from different layers. The qualitative result is also confirmed by the SSIM and mask-SSIM scores.

With vs. without body pose alignment. In our method, we estimate the body pose and shape parameters of the images with HMR and render the SMPL body model with these parameters. This can be viewed as the alignment of pose and position of human body. Even though the re-identification is believed to own the ability of filtering out the influence of body posture and position, we still suppose that the body pose alignment contributes to the texture generation process because the influence of body pose and position is inevitable [59].

In this experiment, we substitute the SMPL parameters with the randomly chosen parameter from the walking sequences of the CMU MoCap database [2]. The result is shown in the column labeled "No-pose" of Fig. 6. The experiment shows that the model without pose alignment can generate textures of acceptable quality. However, the defects in arm parts indicates the significance of body alignment. Moreover, in the first example of Fig. 6 where human only occupies half part of the image, the model without pose alignment cannot recovery the human body size and location automatically and it regards the background as a part of the body.

The SSIM and mask-SSIM scores of the model without pose alignment are 0.158 and 0.365 respectively, which are lower than our method. This result indicates that the feature extracted by the re-identification model more or less influenced by the pose and position of the human body, which is consistent with the conclusion in [59].

With vs. without body part features. In our method, we employ PCB model [46] as the feature extraction module of re-identification loss because it can extract features from every part of the human body, which is supposed to be beneficial for reconstructing details of human body. To demonstrate this, we compared the performance of our method with the model without body part features. In order to exclude the influence caused by the different performances of re-identification network, the top-1 accuracy of the two re-id networks are both around 92% on the Market-1501 dataset.

The result is shown in the column labeled “No-PCB” of Fig. 6. The result with and without body part features is mostly similar while they only differs on the arms or clothing details. This aligns well with the SSIM and mask-SSIM results, where the scores of No-PCB is slightly lower than proposed method. The high quality textures generated by the model without PCB can be ascribed to the high accuracy of re-identification model which extracts precise features.

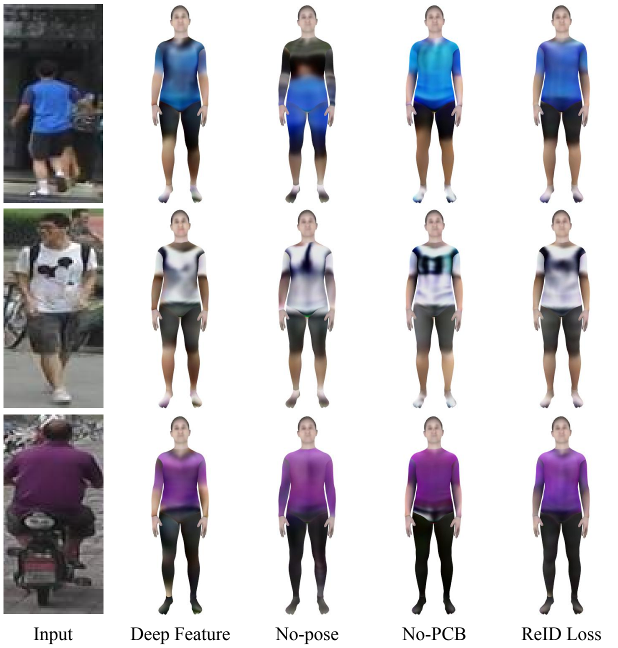

Figure 6: **Qualitative results of ablation study.** The result of different model settings is shown above. Each column shows (in order from left): input images (from test set), the ℓ_1 loss on deep features, model without the pose alignment, re-identification loss without body part features, the re-identification loss proposed in Sec. 3.3

Model	Deep Feature	No-Pose	No-PCB	ReID Loss
SSIM	0.155	0.158	0.159	**0.164**
mask-SSIM	0.354	0.365	0.369	**0.372**

Table 2: **Quantitative results of ablation study.** The SSIM and mask-SSIM score of ReID loss is the higher than other loss functions.

The IS scores of the four methods are 3.77, 4.07, 3.75 and 3.96 respectively while the mask-IS scores are 2.37, 2.63, 2.77 and 2.52. We do not show these results in Table 2 for the reason in Sec. 4.1.

4.5. User Study

A commonly used way to further assess the reality of generated texture is the user study, as human judgment is the ultimate criterion in the generative model. However, unlike previous works [23], our network generates the textures instead of images of human, which makes it impossible for an unprofessional user to tell which texture is better. Moreover, the available rendering software cannot bridge the style gap between rendered and real images, which makes a direct comparison between them impossible. To tackle this issue, we designed the user study aiming at comparing the generated textures with 3D-scanned textures which can be considered as the “real image” in the domain of texture. We generated 55 image pairs and each pair contains one image rendered with the generated texture and another one rendered with the real texture from SURREAL [50]. 30 users have to choose one image with higher quality among two images in 2 seconds. The first 5 image pairs are used for practice thus are ignored when computing scores.

From the result of our user study, users consider the quality of generated textures is higher than scanned textures in 32% image pairs. This shows the relatively high quality of textures generated by our method. By reviewing our user study, we find that the generated textures suffer from blurring while the 3D scanning tends to preserve the details. There are two reasons behind this. Firstly, our textures are generated from blurred images in Market-1501 dataset. Secondly, the differential render, Opendr, only performs well on textures of small size, which limits the resolution of our generated textures.

4.6. Bird Texture Generation

Apart from generating textures of the human body, our framework can also be applied to other object categories. [26] presents a learning framework called CMR for recovering the 3D shape, camera, and texture of an object from a single image. CMR projects the 3D mesh to 2D images and uses the perceptual loss [58] between the rendered image and input image as the loss function. We believe our re-identification supervised method can outperform CMR as our loss function performs better than the perceptual loss, which is demonstrated in Sec. 4.3.

To compare with them, we trained the re-identification network on CUB-200-2011 dataset [52] and use it as the perceptual metrics in the texture generation module. The CUB-200-2011 dataset has 6,000 training and 6,000 test images from 200 birds species. Following [26], we filter out nearly 300 images where the visible number of keypoints are less than or equal to 6. Our experiment result is shown in Fig. 7.

From the experiment results we can conclude that the textures generated with re-identification loss are more accurate than those generated with the perceptual loss. Par-

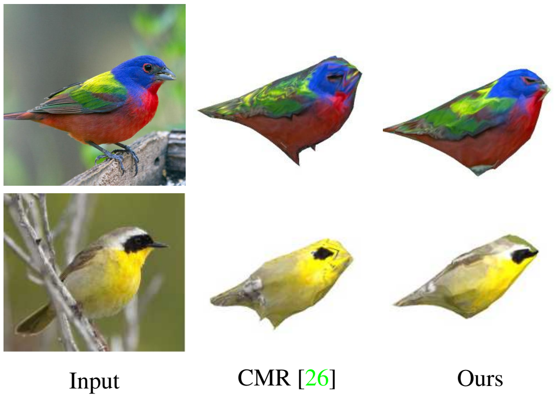

Figure 7: **Sample results.** The textures generated by our method is of higher quality than textures of CMR. This is clearly shown on the heads of birds.

ticularly, our method succeeds in reconstructing details of the input image. For example, the bird heads of our generated textures look more realistic than CMR [26]. This experiment shows that our method holds the potential to be applied in the texture generation task for general categories.

5. Action Recognition

SURREAL [50] provides an effective approach to generate the synthetic dataset and proves that pretraining on synthetic dataset can enhance the performance of human parsing and depth estimation models. However, due to the shortage of available human body textures, the generated dataset suffers from a lack of texture diversity. This limits the generalize ability of models pretrained on the synthetic datasets.

Our method provides an efficient way to generate a large number of textures, which can be used to synthesize datasets with higher diversity and tackle the aforementioned issue. In order to prove this, we carry out the experiment to compare the networks pretrained on action recognition datasets synthesized with different textures. One dataset is the SURREAL dataset generated with 772 scanned textures while another dataset called SURREAL++ is generated with 1.5k textures extracted from Market-1501. We generate SURREAL++ with the method proposed in SURREAL using sequences of 2607 action categories from CMU MoCap dataset [2]. This makes the SURREAL++ embody 67,582 continuous image sequences containing 6.5 million frames, which is of the same size as the SURREAL dataset.

To evaluate the generated dataset, we implement the non-local neural networks [54] which is commonly used for action recognition task and pretrain it on both the SURREAL and SURREAL++ dataset. Then we fine-tune the networks and test them on UCF101 dataset [45] to estimate the networks' performance. The UCF101 dataset contains 13320 videos from 101 action categories. It uses three train/test splits and each split contains around 9.5k training videos and 3.7k test videos. We report our method by the average of 3-fold cross-validation. In addition, we use the non-local network trained on UCF101 dataset as our baseline model.

Training Data	Top-1 Acc. (%)	Top-5 Acc. (%)
UCF101 (baseline)	82.03	94.43
SURREAL	85.83	96.98
SURREAL++	**86.89**	**97.04**

Table 3: **Experiment results.** We show the top-1 and top-5 accuracy of models trained on different datasets.

Table. 3 summarizes test results on UCF101. The model pretrained on the SURREAL dataset and fine-tuned on UCF101 is 3.80% higher in top-1 accuracy than the baseline model, while the model pretrained on SURREAL++ dataset is 4.86% higher in top-1 accuracy. This result shows that the dataset with richer texture diversity can elevate the generalize ability of networks and such kind of diversity can be obtained with our method.

6. Conclusion and Future Work

In this paper, we present an end-to-end framework for generating the texture from a single RGB image. This is achieved by incorporating the pretrained re-identification network as the supervision for texture generation. We have shown that re-identification network can work as a good supervisor in the texture generation task due to its ability to extract body features while reducing the influence in pose variations. We have also proved the extensive application potential of re-identification network in the 3D reconstruction of general categories. To provide the possible usage of our generated body textures, we have demonstrated the diversity in our textures can provide the pretrained model with higher performance.

As the quality of the generated human body texture is restricted by low-quality differentiable render, we suppose that a high-quality renderer will enhance the performance of our method dramatically. We also note that as our framework renders a synthetic image in a similar pose as that of the input image, the quality of texture in occluded parts is not guaranteed. However, from the training images, we can find another image $\mathbf{x}'$ with the same identity as the input image$\mathbf{x}$, but in a different pose. Then, we can align the pose of rendered image $\mathbf{y}$ with $\mathbf{x}'$ and therefore supervise the texture generation process under another viewpoint. This extension will be explored in the future.

Acknowledgement. This research was supported by National Key R&D Program of China (No. 2017YFA0700800).

References

[1] Blender - a 3d modelling and rendering package. https://www.blender.org/. 1

[2] Carnegie-mellon mocap database. http://mocap.cs.cmu.edu/. 6, 8

[3] Ehsan Aganj, Pascal Monasse, and Renaud Keriven. Multi-view texturing of imprecise mesh. In *Asian Conference on Computer Vision*, pages 468–476, 2009. 2

[4] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[5] Thiemo Alldieck, Marcus A. Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *International Conference on 3D Vision*, pages 98–109, 2018. 2

[6] Cédric Allène, Jean-Philippe Pons, and Renaud Keriven. Seamless image-based texture atlases using multi-band blending. In *International Conference on Pattern Recognition*, pages 1–4, 2008. 2

[7] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: shape completion and animation of people. *ACM Trans. Graph.*, 24(3):408–416, 2005. 2

[8] Alexandru O. Balan, Leonid Sigal, Michael J. Black, James E. Davis, and Horst W. Haussecker. Detailed human shape and pose from images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 2

[9] Fausto Bernardini, Ioana M. Martin, and Holly Rushmeier. High-quality texture reconstruction from multiple scans. *IEEE Transactions on Visualization and Computer Graphics*, 7(4):318–332, 2001. 2

[10] Sai Bi, Nima Khademi Kalantari, and Ravi Ramamoorthi. Patch-based optimization for image-based texture mapping. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2017)*, 36(4), 2017. 2

[11] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV 2016*, pages 561–578, 2016. 1, 2

[12] Marco Callieri, Paolo Cignoni, Massimiliano Corsini, and Roberto Scopigno. Masked photo blending: Mapping dense photographic data set on high-resolution sampled 3d models. *Computers & Graphics*, 32(4):464–473, 2008. 2

[13] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *IEEE International Conference on Computer Vision*, pages 1520–1529, 2017. 5

[14] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based CNN with improved triplet loss function. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1335–1344, 2016. 3

[15] Endri Dibra, Himanshu Jain, A. Cengiz Öztireli, Remo Ziegler, and Markus H. Gross. Human shape from silhouettes using generative HKS descriptors and cross-modal neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5504–5514, 2017. 3

[16] Martin Eisemann, Bert De Decker, Marcus Magnor, Philippe Bekaert, Edilson De Aguiar, Naveed Ahmed, Christian Theobalt, and Anita Sellent. Floating textures. In *Computer graphics forum*, volume 27, pages 409–418, 2008. 2

[17] Ran Gal, Yonatan Wexler, Eyal Ofek, Hugues Hoppe, and Daniel Cohen-Or. Seamless montage for texturing models. In *Computer Graphics Forum*, volume 29, pages 479–486, 2010. 2

[18] Niloofar Gheissari, Thomas B. Sebastian, and Richard I. Hartley. Person reidentification using spatiotemporal appearance. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1528–1535, 2006. 3

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[20] Peng Guan, Alexander Weiss, Alexandru O. Balan, and Michael J. Black. Estimating human shape and pose from a single image. In *IEEE International Conference on Computer Vision*, pages 1381–1388, 2009. 1

[21] Peng Guan, Alexander Weiss, Alexandru O. Balan, and Michael J. Black. Estimating human shape and pose from a single image. In *IEEE International Conference on Computer Vision*, pages 1381–1388, 2009. 2

[22] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. *arXiv preprint arXiv:1802.00434*, 2018. 1

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5967–5976, 2017. 7

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[25] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 3, 4, 5

[26] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV 2018*, pages 386–402, 2018. 1, 2, 5, 7, 8

[27] Wadim Kehl, Nassir Navab, and Slobodan Ilic. Coloured signed distance fields for full 3d object reconstruction. In *British Machine Vision Conference*, 2014. 2

[28] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4704–4713, 2017. 1, 2

[29] Victor Lempitsky and Denis Ivanov. Seamless mosaicing of image-based texture maps. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2007. 2

[30] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7398–7407, 2017. 3

[31] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159, 2014. 3

[32] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2194–2200, 2017. 3

[33] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *IEEE International Conference on Computer Vision*, page 2, 2018. 3

[34] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Shuai Yi, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *IEEE International Conference on Computer Vision*, pages 350–359, 2017. 3

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 2

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248, 2015. 3, 4

[37] Matthew M. Loper and Michael J. Black. Opendr: An approximate differentiable renderer. In *ECCV 2014*, pages 154–169, 2014. 3, 4

[38] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing System*, pages 405–415, 2017. 3, 5

[39] Natalia Neverova, Riza Alp Güler, and Iasonas Kokkinos. Dense pose transfer. In *ECCV 2018*, pages 128–143, 2018. 1, 2, 5

[40] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. *arXiv preprint arXiv:1808.05942*, 2018. 1, 3

[41] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. *arXiv preprint arXiv:1805.04092*, 2018. 3

[42] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Transactions on graphics (TOG)*, 22(3):313–318, 2003. 2

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015. 3

[44] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2226–2234, 2016. 5

[45] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 8

[46] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and A strong convolutional baseline). In *ECCV 2018*, pages 501–518, 2018. 3, 4, 6

[47] Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *British Machine Vision Conference*, 2017. 3

[48] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems*, pages 5242–5252, 2017. 3

[49] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *ECCV 2018*, pages 20–38, 2018. 1, 3

[50] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4627–4635, 2017. 5, 7, 8

[51] Michael Waechter, Nils Moehrle, and Michael Goesele. Let there be color! large-scale texturing of 3d reconstructions. In *ECCV 2014*, pages 836–850. Springer, 2014. 2

[52] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 7

[53] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM Multimedia Conference on Multimedia Conference*, pages 274–282, 2018. 3

[54] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 8

[55] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13(4):600–612, 2004. 5

[56] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2, 2016. 4

[57] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Deep metric learning for person re-identification. In *International Conference on Pattern Recognition*, pages 34–39, 2014. 3

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep networks as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 7

[59] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017. 6

[60] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *IEEE International Conference on Computer Vision*, pages 3239–3248, 2017. 3

[61] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 3, 5

[62] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016. 2, 3

[63] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3346–3355, 2017. 4

[64] Qian-Yi Zhou and Vladlen Koltun. Color map optimization for 3d reconstruction with consumer depth cameras. *ACM Transactions on Graphics (TOG)*, 33(4):155, 2014. 2