

THREE-DIMENSIONAL HUMAN TEXTURE ESTIMATION LEARNING FROM MULTI-VIEW IMAGES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

By
Said Fahri Altındış
December 2023

Three-Dimensional Human Texture Estimation Learning from Multi-View Images

By Said Fahri Altındaş

December 2023

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Ayşegül Dündar Boral(Advisor)

Uğur Güdükbay

Mehmet Erkut Erdem

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

THREE-DIMENSIONAL HUMAN TEXTURE ESTIMATION LEARNING FROM MULTI-VIEW IMAGES

Said Fahri Altındış

M.S. in Computer Engineering

Advisor: Ayşegül DüNDAR Boral

December 2023

In the fields of graphics and vision, accurately estimating 3D human texture from a single image is a critical task. This process involves developing a mapping function that transforms input images of humans in various poses into parametric (UV) space, while also effectively inferring the appearance of unseen parts. To enhance the quality of 3D human texture estimation, our study introduces a framework that utilizes deformable convolution for adaptive input sampling. This convolution is uniquely characterized by offsets learned through a sophisticated deep neural network. Additionally, we introduce an innovative cycle consistency loss, which markedly enhances view generalization. Our framework is further refined by incorporating an uncertainty-based, pixel-level image reconstruction loss, aimed at augmenting color accuracy. Through comprehensive comparisons with leading-edge methods, our approach demonstrates notable qualitative and quantitative advancements in the field.

Keywords: Texture Estimation, Deformable Convolution, Uncertainty Estimation.

ÖZET

ÇOKLU GÖRÜNTÜDEN ÜÇ BOYUTLU İNSAN DOKU TAHMINİNİ GELİŞTİRMEK

Said Fahri Altındış
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Ayşegül Dünder Boral
December 2023

Bilgisayar grafiği ve bilgisayarlı görü alanlarında, tek bir görüntüden üç boyutlu insan dokusunun doğru bir şekilde tahmin edilmesi kritik bir görevdir. Bu süreç, insanların çeşitli pozlardaki giriş görüntülerini parametrik (UV) uzaya dönüştüren bir haritalama fonksiyonu geliştirmeyi ve görünmeyen kısımların görünümünü etkin bir şekilde çıkarmayı içerir. Üç boyutlu insan doku tahmininin kalitesini artırmak için, çalışmamız, deformasyonlu evrilim kullanan bir yöntem sunmaktadır. Bu evrilim, gelişmiş bir derin sinir ağı aracılığıyla öğrenilen ofsetlerle özgün bir şekilde karakterize edilmiştir. Ek olarak çalışmamızda, görünümü genelleştirmeyi belirgin şekilde iyileştiren yenilikçi bir döngü tutarlılığı kaybı tanıtlıyoruz. Metodumuz, renk doğruluğunu artırmayı amaçlayan belirsizliğe dayalı, piksel seviyesinde görüntü yeniden yapılandırma kaybı ekleyerek daha da iyileştirilmiştir. Alanın öncü yöntemleriyle kapsamlı karşılaştırmalar yaparak, yaklaşımımızın alanda dikkate değer niteliksel ve niceliksel ilerlemeler gösterdiğini ortaya koymaktadır.

Anahtar sözcükler: Doku Tahmini, Deformasyonlu Evrilim, Belirsizlik Tahmini.

Acknowledgement

I would like to express my deepest gratitude to my advisor, Ayşegül Dünder Boral, for her invaluable guidance and continuous support throughout my studies. Her insights and expertise have been fundamental to my academic journey, and for this, I am profoundly thankful.

I am also sincerely appreciative of Prof. Uğur Güdükbay and Prof. Erkut Erdem for their willingness to serve on my thesis jury. Their expertise and thoughtful critiques have significantly contributed to the refinement of my research.

A special note of thanks goes to my parents, Mazhar Altındış and Hatice Altındış and my sisters, Şule Gülcan Karaköse and Beyzanur Altındış. Their unwavering guidance and unconditional support, even in the most challenging times, have been the bedrock of my strength and perseverance. Their belief in me has been a constant source of encouragement.

This journey would not have been possible without these individuals, and I am immensely grateful for their contributions to my academic and personal growth.

Contents

1	Introduction	1
2	Related Work	4
2.1	Self-Supervised Learning with Differentiable Renderers	5
2.2	Improving Texture Predictions for Invisible Regions	6
3	Method	7
3.1	Architecture	9
3.1.1	The Refinement Module: Incorporating Deformable Con- volutions	11
3.1.2	Integration of Mask-Fusion Method	12
3.1.3	Supervision of the Framework through Rendering and Backpropagation	13
3.2	Loss Functions	14
3.2.1	The Overall Base Loss	15
3.2.2	Uncertainty-based Reconstruction Loss	18

3.2.3	Cycle Consistency Loss	20
3.2.4	Total Loss	21

4 Experiments 23

4.1	Dataset Selection and Utilization	23
4.1.1	Market-1501 Dataset	23
4.1.2	DeepFashion	24
4.2	The Overall Architecture	25
4.2.1	Deformable Convolution Layers	25
4.2.2	Attention-Based Network Structure	25
4.2.3	Deformable Convolution Specifics	26
4.2.4	Additional Architectural Features	26
4.2.5	The Confidence Model	27
4.2.6	Training Specifications	28
4.3	Evaluation Metrics	28
4.3.1	Evaluation Across Different Views	29
4.4	Comparison with the State-of-the-art	31
4.4.1	Performance Analysis of Competing Methods	32
4.4.2	User Study	33
4.4.3	Comparison on DeepFashion	33

4.4.4	Comparison on Pose-guided Human Image Generation . . .	34
4.5	Ablation Study	35
4.5.1	Refinement Module (BL+Refine)	35
4.5.2	Uncertainty-Based Reconstruction Loss (BL+URL)	36
4.5.3	Cycle Consistency Loss (BL+Cycle)	36
5	Analysis and Discussions	41
5.1	Comparative Analysis with State-of-the-Art Methods	41
5.2	Ablation Study Insights	42
5.3	Limitations and Future Directions	42
5.3.1	Limitations	42
5.3.2	Future Directions	43
5.4	Discussions	44
6	Conclusion	45
	Bibliography	46

List of Figures

1.1	Texture estimation is not aligned	3
3.1	The overall framework	8
3.2	Cycle Consistency Loss	10
4.1	Input/same view and novel view ground-truth images are provided in the first two columns.	24
4.2	Qualitative results of our method and state-of-the-art competing methods on the Market-1501 dataset.	37
4.3	Qualitative results of our method and state-of-the-art competing methods on the DeepFashion dataset.	38
4.4	Qualitative results of our method and Coordinate-based texture inpainting [1] on the Market-1501 dataset.	38
4.5	Qualitative results of Ablation Study.	39
4.6	Visualization of offsets in input images.	40

List of Tables

4.1	Comparisons of models trained with multi-view consistency and without multi-view consistency	31
4.2	Evaluation results of our method on Market-1501 dataset and state-of-the-art competing methods	31
4.3	Comparison of Parameter Sizes Across Different Methods	32
4.4	Evaluation results of our method on DeepFashion dataset and state-of-the-art competing methods	34
4.5	Ablation Study on Market-1501 dataset.	36

Chapter 1

Introduction

The task of estimating three-dimensional (3D) human texture from a single image holds crucial significance across diverse domains, spanning virtual reality (VR), augmented reality (AR), gaming, robotics, and digital cloth try-on applications. This challenge becomes particularly intricate due to the dual requirements of predicting textures for non-visible human body parts and accommodating the vast spectrum of human poses and appearances.

Within the research community, the endeavor to generate 3D human textured models from a single image has witnessed a surge in interest. The introduction of differentiable, or neural, renderers [2] has streamlined the training of deep learning models for this task, allowing for end-to-end training by approximating rasterization gradients. Despite this advancement, numerous existing methods rely on labor-intensive and costly data for training, such as 3D scanning [3, 4, 5, 6] or dense human pose estimation [4, 7]. Our research aims to circumvent the need for expensive 3D labels by exclusively leveraging multi-view images [8, 9, 10, 11, 12]. Notable methodologies include Zhao et al.’s [11] use of cross-view consistency, Wang et al.’s [13] integration of a re-identification loss, and Xu and Loy’s [12] implementation of an attention-based architecture.

Despite significant strides in this domain, existing works grapple with limitations that impede the quality of 3D human texture estimation. Primarily, the setup of texture estimation from a single image introduces a spatial misalignment between input and output images, rendering them unsuitable for Convolutional Neural Networks (CNNs) with local receptive fields. For instance, hands may appear anywhere in input images, yet they possess fixed corresponding locations in the parametric uv space (cf. Fig. 1.1). While prior works propose attention-based architectures to address this issue, our research demonstrates a further enhancement through a deformable convolution-based module, termed the *refinement* module. The learnable offsets from the deformable convolution, derived from a deep attention-based architecture, enable the refinement module to adaptively sample input images and generate high-fidelity texture predictions for both visible and invisible pixels.

Moreover, conventional approaches sidestep the use of pixel-level reconstruction loss between rendered and ground-truth images due to its subpar performance in detail generation, attributable to inaccurate human body pose and shape estimations causing misalignments. In response, we advocate for a confidence-based pixel-level reconstruction loss, effectively addressing misalignments and yielding substantial improvements. Additionally, we augment our texture estimation with a novel cycle consistency loss, leveraging the estimation of texture from a single image, rendering it from a new perspective, and re-encoding the texture to enhance model generalizability across diverse views.

The evaluation of our method against the current state-of-the-art employs metrics such as Structural Similarity Index Measure (SSIM) [14], Learned Perceptual Image Patch Similarity (LPIPS) [15], and Cosine Similarity (CosSim) [16], comparing input and rendered images from the same viewpoint. However, recognizing that the essence of 3D models lies in predicting appearance from *novel* viewpoints, our evaluation extends to performance assessments from new perspectives, aiming to enhance results for both the original and additional viewpoints.

Our contributions are as follows:



Figure 1.1: Texture estimation is not aligned, unlike many tasks with input-output alignment (e.g., part segmentation).

1. We present a deformable convolution-based framework to adeptly map spatially diverse input images into fixed parametric uv maps, addressing the inherent challenge of spatial misalignment.
2. We propose a confidence-based pixel-level reconstruction loss to rectify misalignments in the ground truth, enabling training with pixel-level reconstruction loss and ensuring closer color fidelity to the input image.
3. We devise a novel cycle consistency loss that significantly elevates the quality of texture estimation.
4. We conduct an extensive analysis of the models from the perspective of novel viewpoints to assess their effectiveness in rendering unseen areas in 3D human texture estimation.

We perform extensive evaluations with an array of quantitative metrics that show the effectiveness of our scheme compared to the several state-of-the-art approaches.

Chapter 2

Related Work

3D texture estimation from images can enable various VR/AR applications and attracts much interest from the research community [17, 18, 19, 20, 21, 22, 23], especially the 3D human texture estimation [24, 25, 26, 27, 3, 4, 28, 6, 13, 11, 29, 12]. Many methods have been proposed for 3D human reconstruction that take multi-view images and optimize them [30, 31, 32]. Especially videos have been explored with implicit representations for modeling scenes [33, 34] as well as for human reconstruction such as the recently proposed Vid2Avatar [35]. While achieving impressive results, these models are overfitting to a sequence and can not be used for single-image inference.

In this work, we are also interested in 3D human reconstruction but inferring them from single-images [36, 37, 1, 11, 29, 12] since it is more applicable to real-world use cases. Even though many works aim to infer texture from single-view images, they may require significantly expensive labor-intensive data during training. For example, most of the methods require 3D scanning [3, 4, 5, 6, 38] such as the recently popularized implicit function-based methods [6, 39, 40, 41] and few others require dense human pose estimation [4, 7] or depth data [42]. In our work, we aim at learning texture prediction without expensive 3D labels but by relying on image datasets [13, 11, 12].

2.1 Self-Supervised Learning with Differentiable Renderers

In the domain of 3D human texture estimation, recent methodologies have focused on training networks on image collections in a self-supervised manner using differentiable renderers [8, 2, 9, 10]. This approach involves predicting 3D body and pose models using advanced human mesh reconstruction methods and subsequently training a network to map these images into UV space. The technique of mapping human images into UV space has also found applications in dense human pose estimation [43], mesh recovery [44], and garment prediction [45], as well as in texture estimation [13, 11, 12].

For enhanced texture estimation, training data is often augmented with part segmentation models, which has proven to boost performance significantly [11, 12]. Despite these advancements, the field still faces the critical challenge of aligning input and output images more precisely. Texture prediction requires sampling from different locations for each example, necessitating varied approaches. Some methods learn texture flows to sample pixels directly from the input image [17, 11], while others focus on directly learning the texture maps in RGB values [13]. There are also methods that combine these techniques [12], merging the advantages of direct detail transfer from the input to the texture and the visually pleasing synthesis of invisible regions.

Our work extends these methodologies by incorporating deformable convolution [46]. This integration simplifies the prediction of texture maps, allowing the network to operate on adaptive offsets. Deformable convolution has been successfully implemented in a variety of computer vision tasks, including object detection, instance segmentation [47, 46], and texture synthesis [48]. We leverage this technique in our unique deformable-based design to achieve significant improvements in texture estimation, addressing the inherent challenges of alignment and sampling variability.

2.2 Improving Texture Predictions for Invisible Regions

A significant challenge in 3D texture estimation lies in accurately predicting textures for regions not visible in the source image. To address this, generative models, often classified under neural rendering methods, have been proposed [49, 50, 51]. Notably, StylePeople [49] introduces a generative model for creating full-body human avatars using a StyleGAN architecture [52]. This model synthesizes novel humans by sampling random codes and employs an encoder to estimate StyleGAN’s latent codes, thereby creating a reconstruction framework that outputs neural textures for any given input image. However, this process is not without its complexities, particularly due to the challenging nature of inverting images with an encoder, which necessitates a computationally intensive optimization step.

Furthermore, other approaches, such as the work by Zhao et al [11], leverage multi-view images to enforce cross-view consistency, thereby enhancing texture estimation for unseen regions. Our research extends these efforts by introducing a novel cycle consistency loss, aiming to refine the texture estimation process further.

Lastly, a notable obstacle in previous works has been the reluctance to employ reconstruction losses between the input and output images, primarily due to the observed degradation in performance [13, 12]. In contrast, our work proposes the use of a confidence-based reconstruction loss, designed to significantly improve texture estimation results by judiciously handling the inaccuracies and uncertainties inherent in the reconstruction process.

Chapter 3

Method

Our approach is centered on a setup where the 3D human texture estimation network is tasked with mapping spatially variant input images into predefined UV space coordinates. This task presents a unique challenge, as unlike most tasks in computer vision, the inputs (e.g., hands in various positions in the input images) and outputs (fixed coordinates in the parametric UV space) are not spatially aligned. This lack of alignment necessitates a carefully designed network architecture that can adeptly handle the disparity between the input and output spaces.

In practical terms, the pixels relevant to synthesizing a particular coordinate in the parametric UV space might be located at a considerable distance in the input space. While it is feasible to directly transfer the visible pixels from the input image to the UV maps, once the corresponding pixel locations are identified, the network must also be capable of synthesizing the invisible pixels. This synthesis is contingent upon the visible parts of the image, requiring the network to effectively ‘hallucinate’ the RGB values for these unseen areas.

1. Accurately determining the correct offsets for the visible regions, enabling the direct copying of RGB values from the input images.
2. Effectively hallucinating RGB values for the invisible pixels in the input

image, leveraging the visible parts as a basis for this synthesis.

This methodological approach addresses the inherent complexities of 3D human texture estimation, particularly the need for precise pixel-level manipulation and the creative reconstruction of unseen areas based on visible data.

3.1 Architecture

Our framework is specially designed to adapt to the geometric variations inherent in the input images. The primary objective is to determine the offsets of the associated pixels for each pixel within the parametric UV map. This process, which is akin to flow predictions, is crucial for aligning and mapping the textures accurately.

The framework is structured around two key goals:

1. **Identifying Offsets:** The initial step involves finding the offsets for the associated pixels. This task is important to predicting flows but with an added complexity. Our aim is not just to sample the input image with these offsets but to develop a robust architecture that can adaptively sample and process input images for enhanced texture fidelity.
2. **Processing for Texture Output:** Once the offsets are identified, the associated pixels are processed to generate texture predictions. This stage is pivotal in synthesizing both the visible and invisible regions of the texture map.

To achieve these objectives, we incorporate **deformable convolutions** into our framework. This approach differs from previous architectures that have used deformable convolutions. In our model, the pathway for predicting offsets is augmented with a deep convolutional neural network, as depicted in Fig. 3.1.

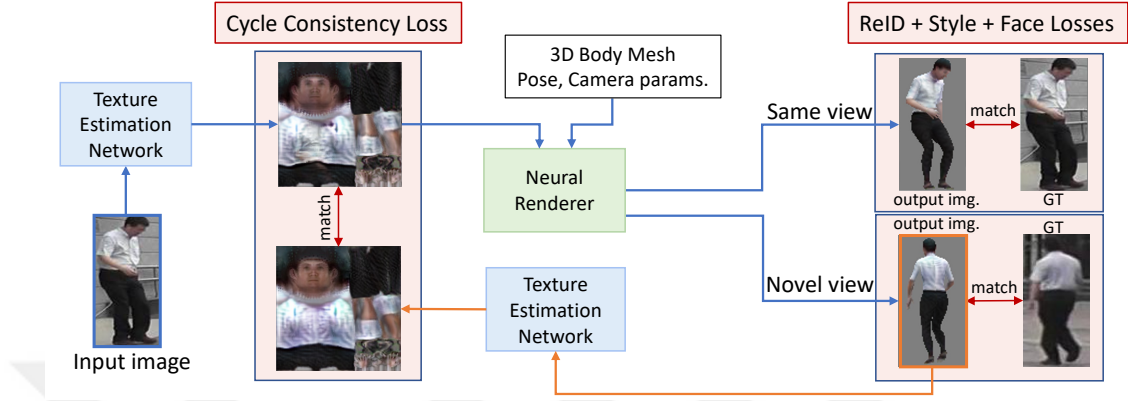


Figure 3.2: We predict a texture from an input image and render it for both the same and novel views. From the image rendered with a novel view, we estimate the texture again. We expect the estimated texture to match the texture predicted from the input image.

This network is responsible for outputting flow predictions and can be based on previous architectures proposed for texture estimation.

An attention-based architecture [12], as shown in Fig. 3.1, is utilized in our framework. This architecture leverages a color-encoded output UV space as the query and combines the 2D part-segmentation and input image as the key. The value is derived from the input image and the flow field of the image, represented as 2D coordinates for each pixel.

Multi-scale features are extracted from these inputs using separate encoder-decoder networks. The encoded value, key, and query features are then fed into attention blocks [53]. These blocks enable the key to correlate with the query elements, generating attention maps that provide global information for feature distribution from input to output.

The outputs from the attention blocks are channeled through a Unet-like architecture, as detailed in Fig. 3.1, resulting in the final flow and RGB predictions. This architecture is pivotal in our framework, facilitating the adaptive sampling and processing necessary for high-fidelity texture prediction in 3D human texture estimation.

3.1.1 The Refinement Module: Incorporating Deformable Convolutions

In our architecture, the component that incorporates deformable convolutions is referred to as the *refinement module*, as illustrated in Fig. 3.1. This module plays a critical role in enhancing the texture mapping from the input image to the UV space. The offsets, crucial for the deformable convolution operation, are calculated based on the flow predictions. These predictions are specifically trained to learn the mapping from the input image to the output UV space, focusing on understanding the absolute values of the input coordinates. Once the flow predictions are obtained, we transform these absolute coordinate values into offsets for pixel coordinates. This transformation involves applying a convolutional layer to the flow predictions, which then serve as the basis for the deformable convolution operation. The deformable convolution operation utilizes these offsets to sample pixels from distant coordinates in the input image. It then processes these samples to output the final texture maps. This process allows for a more flexible and accurate texture synthesis, particularly for regions not directly visible in the input.

The refinement module is designed to take an input image of size 128×64 and output features with a dimension of 128×128 , based on the offsets calculated for a spatial size of 128×128 . A key aspect of our implementation is the modification of the standard deformable convolution operation, which typically expects the offsets and input features to be of the same dimension. Our adaptation allows the module to handle arbitrary input sizes in relation to the offsets. This modification is significant as the offset size determines the output dimension, thereby providing a seamless and intuitive flow from the input image to the output texture maps. Consequently, our architecture effectively deforms from 128×64 images to 128×128 texture maps, enhancing the accuracy and fidelity of the texture estimation process.

3.1.2 Integration of Mask-Fusion Method

In our framework, we adopt the mask-fusion method to synergize the advantages of both RGB texture prediction and texture flow, as proposed by [12]. This method enables us to effectively merge the precision of direct RGB texture mapping with the flexibility offered by texture flow.

The output of our network comprises three key elements:

- 1. RGB Texture Map (T_{RGB}):** This component represents the direct texture prediction in RGB format. It provides the color and detail necessary for rendering realistic textures on the 3D model.

- 2. Texture Flow (F):** Texture flow maps the spatial relationship between the input image and the UV space. It serves as a guide for how textures should be transferred from the input image to the model.

- 3. Fusion Mask (M):** The fusion mask plays a pivotal role in blending the RGB texture map and the texture flow. It determines the contribution of each component in the final texture output, allowing for a seamless integration of both methods.

While the RGB texture map and texture flow are explicitly represented in the network’s output, the fusion mask M is not included in Fig. 3.1 for the sake of brevity. However, its role is integral to the texture synthesis process, as it ensures that the strengths of both RGB texture mapping and texture flow are effectively utilized.

This integration strategy, combining RGB texture prediction with texture flow through a fusion mask, enhances the capability of our network to produce high-fidelity textures. It allows for the nuanced handling of both visible and invisible regions, ensuring realistic and coherent texture mapping across the 3D model.

The mask-fusion process is as follows:

$$T = M \odot f_{sample}(F, I) + (1 - M) \odot T_{RGB}, \quad (3.1)$$

where f_{sample} refers to the bilinear sampling function that samples textures from the input image, I , by the flow predictions, F , and M is a binary mask. In this way, visible pixels can be taken by the more accurate $f_{sample}(F, I)$, and invisible pixels can be taken from T_{RGB} .

3.1.3 Supervision of the Framework through Rendering and Backpropagation

In our approach, the supervision of the framework is executed through a process of rendering the intermediate results generated by the deep network. This rendering is key to visualizing and assessing the progress of the network’s predictions at various stages. As illustrated in Fig. 3.1, these intermediate outputs are crucial for providing a clearer understanding of how the network is performing and where adjustments may be needed.

Following the rendering of these intermediate results, we employ backpropagation of losses to fine-tune the network’s performance. This backpropagation is needed for providing direct and immediate feedback to the network, particularly in terms of adjusting offsets. By doing so, offsets - which are critical for the alignment and accuracy of texture mapping - receive more focused and direct supervision. This step ensures that the deformable convolutions within the refinement module are correctly aligning and adapting to the spatial variations in the input images.

As the training progresses, we gradually shift the emphasis of supervision from these intermediate renderings to the final output of the network. This transition involves reducing, or turning off, the additional supervision provided by the intermediate stages and increasingly relying on the supervision received directly from the final output. This strategy allows the refinement module to

progressively adapt and become more attuned to the end goal of accurate texture mapping.

The gradual shift in the focus of supervision ensures that the network not only learns to correct its course in the early stages but also becomes proficient at achieving high-quality final outputs. This balanced approach to supervision - combining immediate feedback from intermediate renderings with the final goal-oriented supervision - is instrumental in optimizing the network’s performance for accurate and realistic 3D human texture estimation.

3.2 Loss Functions

The loss functions in our model are crucial for the self-supervised learning process, bridging the gap between the input and the rendered images. This section aims to detail these functions and their roles in enhancing the model’s accuracy and effectiveness.

Upon receiving an input image I , our model outputs a human texture $T(I)$. This texture, along with the mesh (M) and camera (C) predictions from the RSC-Net model [10], which is based on the SMPL model [54], is processed through a differentiable renderer [2]. The renderer’s output is an image I^r , providing a visual representation of how the estimated texture maps onto a 3D model.

In line with previous research [12], we utilize established loss functions for comparing I and I^r . These functions are integral to ensuring the rendered image retains the core characteristics of the input while accurately depicting the texture mapping.

Additionally, we adopted *Uncertainty-Based Reconstruction Loss* that focuses on minimizing the reconstruction error between I and I^r . The incorporation of uncertainty allows the model to weigh errors differently based on the confidence level of the predictions, leading to a more nuanced and effective learning process.

Finally, we use *Cycle Consistency Loss* To ensure the model’s generalizability and robustness, the cycle consistency loss is employed. This loss encourages the model to maintain consistent texture mapping even when the input image is modified or viewed from different angles, enhancing the model’s applicability in diverse scenarios.

The subsequent sections will delve into the specifics of each loss function, explaining their individual contributions and their collective impact on the model’s performance.

3.2.1 The Overall Base Loss

The overall base loss serves as the foundational objective function for our model’s training. It encapsulates the various facets of texture estimation, from maintaining identity consistency to ensuring stylistic and structural accuracy. This holistic approach to loss formulation ensures that the model is comprehensively trained, leading to high-quality, realistic, and accurate texture estimations.

In summary, the overall base loss, as defined in Eq. 3.2, represents the amalgamation of the key loss functions adopted from previous works, setting a solid foundation for the model’s learning and optimization process. This comprehensive base loss ensures that the model effectively addresses the multiple challenges associated with 3D human texture estimation.

$$\mathcal{L}_{base} = \lambda_1 \times \mathcal{L}_{reid} + \lambda_2 \times \mathcal{L}_{style} + \lambda_3 \times \mathcal{L}_{face}, \quad (3.2)$$

where

1. L_{reid} is the re-identification loss, ensuring identity consistency between the input and rendered images,
2. L_{style} is the part-style loss, focusing on the stylistic similarity of individual body parts,

3. L_{face} is the face structure loss, targeting the accuracy of facial features, and
4. $\lambda_1 = 5000$, $\lambda_2 = 0.4$, and $\lambda_3 = 0.01$. We take these parameters from the previous work [12] and do not tune them.

3.2.1.1 Re-Identification Loss

The re-identification loss designed to maintain the identity characteristics present in the input image. This loss function is particularly important in scenarios where the texture mapping must preserve distinct features of the input, such as in pedestrian re-identification tasks.

As defined by [13], the re-identification loss minimizes the difference in features extracted from a pedestrian re-identification network, denoted as (Φ) , as proposed by Sun et al. [16]. This network extracts features at various layers, with each layer (j) capturing different aspects of the input image.

The re-identification loss is mathematically formulated as:

$$\mathcal{L}_{reid} = \|\Phi_j(I) - \Phi_j(I^r)\|_2^2. \quad (3.3)$$

This loss function ensures that the rendered image, I^r , maintains the core identity features of the input image, I . It is a critical aspect of ensuring the realism and consistency of the texture mapping, especially in applications where the accurate representation of individual characteristics is essential.

3.2.1.2 Part-Style Loss

The part-style loss function is integral to ensuring that each specific body part of the rendered human closely resembles the corresponding part in the input image. This loss function, as proposed by [12], is vital for maintaining the integrity and realism of individual body parts in the texture mapping process.

The part-style loss enforces similarity based on the Gram-matrix [55], a method initially popularized in style transfer tasks. The Gram-matrix provides a way to measure the stylistic similarity between features extracted from different images.

$$\mathcal{L}_{style} = ||G(M_p \odot \Phi_1(I)) - G(M'_p \odot \Phi_1(I^r))||_2^2, \quad (3.4)$$

where M_p and M'_p are the human part segmentation masks from the 2D human parsing model [56] and 3D mesh, respectively, p indicates the body part, and this loss is calculated for each body part separately and summed together. G stands for Gram-matrix, features are again encoded by the same pedestrian re-identification network (Φ), and only the first layer is used. By focusing on the style and appearance of each part, the model can produce more lifelike and convincing renderings in the final output.

3.2.1.3 The Face Structure Loss

The face structure loss [12], is specifically designed to ensure the accuracy and reliability of face textures in the rendered output. Considering the uniformity in human facial structure, this loss function is crucial in maintaining the similarity between the estimated textures and synthetically generated textures, particularly for the face.

The loss function is based on the premise that each region of the parametric UV map is predetermined, and consequently, the location of the face on this texture map is constant. This consistency allows for a focused comparison of facial features by applying a fixed mask to the face region.

$$\mathcal{L}_{face} = -\frac{1}{N} \sum_{i=1}^N s(M_{face} \odot T(I), M_{face} \odot F_{syn}^i), \quad (3.5)$$

where $\{F_{syn}^i\}_{i=1}^N$ is a set of synthetically generated human textures obtained from the synthetic human dataset [57], M_{face} is a predefined binary mask that indicates

the face region on the texture map, and s is a structure-similarity function [14] that calculates face similarities. The loss function in Eq. 3.5 optimizes the network to output face texture predictions with similar structures as F_{syn} . This results in the generation of plausible face textures while retaining the colors of the input human. The losses we use from previous works are our base losses, and Eq. 3.2 gives the overall base loss.

3.2.2 Uncertainty-based Reconstruction Loss

Traditional pixel-level reconstruction losses often struggle with generating details, primarily due to the misalignment issues that arise from inaccuracies in human body pose and shape estimation. This misalignment between rendered and input images can significantly degrade the performance of such loss functions. Consequently, previous 3D human texture estimation models have predominantly relied on re-identification losses and style losses, which don't require strict spatial correspondence.

Despite these challenges, pixel-level reconstruction losses have been shown to improve results significantly when alignment between generated and output images is achieved [58, 59, 60, 61]. They are particularly effective in enhancing the detail and realism of the generated textures.

To harness the advantages of pixel-level reconstruction loss while being resilient to alignment issues, we propose incorporating an uncertainty-based approach. This method involves estimating a confidence map, σ , which dynamically adjusts the reconstruction loss objective based on the level of confidence in alignment at each pixel.

The ground-truth output image, I , and the rendered image, I^r , are compared via the loss given in Eq. 3.6 as was also defined in [62].

$$\mathcal{L}_{url} = - \sum_{x,y} \ln \left(\frac{1}{\sqrt{2\sigma_{x,y}^2}} \exp - \frac{\sqrt{2}|I_{x,y} - I_{x,y}^r|}{\sigma_{x,y}} \right). \quad (3.6)$$

In this loss objective, the role of the confidence map, σ , is to estimate the aleatoric uncertainty of the model [63]. This uncertainty is the noise associated with the data collection that cannot be reduced with more data.

In our setting, the uncertainty is the noise caused by the misalignment of our data. (x, y) are the spatial pixel coordinates of I . The objective is the negative log-likelihood of a factorized Laplacian distribution with the mean predicted by the model and σ predicted by the confidence model. The confidence map estimates which pixels from the rendered image will not be aligned with the ground truth.

By implementing this loss function, our model can intelligently calibrate itself, focusing more on aligning those pixels where it has higher confidence and thus minimizing the overall reconstruction loss. This approach leads to more accurate texture predictions, particularly in the presence of misalignments and other uncertainties inherent in the data.

The confidence model itself utilizes a *Unet* architecture [64]. It processes the input image I and outputs the confidence map σ . This model is specifically used during the training phase to optimize the reconstruction loss by calibrating the model’s confidence in its predictions.

It’s important to note that the confidence model is only utilized during the training phase. During inference, this component is not needed, simplifying the model’s deployment and usage in real-world applications.

3.2.3 Cycle Consistency Loss

The cycle consistency loss is designed to enforce consistency in the textures estimated by the model, ensuring that textures remain coherent and accurate when viewed from different perspectives. This loss function is particularly important for maintaining the integrity of the texture mapping across various camera views.

In the cycle process:

1. The model initially generates a texture map, uv_{map} , from the input image I using the texture estimation network T .
2. This texture map is then used to render a new image for a different view, utilizing another image of the same person. This rendering involves the renderer R , body mesh parameters m , and camera parameters c .
3. The model re-estimates the texture from this newly rendered image, completing the cycle.

The cycle consistency loss is calculated as a pixel-wise L2 loss between the two estimated textures:

$$\mathcal{L}_{cyc} = ||T(I) - T(R(T(I), m, c))||_2. \quad (3.7)$$

Fig. 3.2 explains the detailed structure of the cycle consistency. In previous methods, the texture map is estimated from the input image, and it is rendered for the same and novel views to enforce multi-view consistency and minimize base losses, as described in the previous subsection. We also estimate the texture from the novel view rendering to obtain additional guidance for the texture estimation network. This additional loss also helps us to achieve our second objective, where we can compare images at the pixel level.

3.2.4 Total Loss

The total loss function in our framework is a composite measure that combines several individual loss components to optimize the texture estimation process. This total loss ensures both single-view fidelity and multi-view consistency in texture mapping.

3.2.4.1 Baseline Losses for Different Rendered Images

Baseline Loss for Single-View ($L_{base-sv}$): This loss is calculated between the input image and an image rendered using the texture generated from the input image, along with its camera parameters. Also, it ensures that the model accurately captures the texture details present in the input image’s view.

Baseline Loss for Novel-View ($L_{base-nv}$): This loss is computed between a ground-truth image from a different view and an image rendered using the texture generated from the input image, but with the camera parameters of the novel-view ground-truth image. Additionally, this loss component contributes to multi-view consistency, ensuring that the model accurately predicts the texture of unseen parts.

3.2.4.2 Incorporation of New Losses

Uncertainty-Based Reconstruction Loss (L_{url}): This loss addresses the inaccuracies and uncertainties in texture prediction, especially in the presence of misalignments between the input and rendered images.

Cycle Consistency Loss (L_{cyc}): This loss ensures that the textures estimated for different views are consistent, enhancing the model’s ability to generalize across various viewing angles.

3.2.4.3 Formulation of the Overall Loss Function

The overall loss function that drives the training of our framework is formulated by combining these components. It can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{base-sv} + \mathcal{L}_{base-nv} + \lambda_4 * \mathcal{L}_{cyc} + \lambda_5 * \mathcal{L}_{url}, \quad (3.8)$$

where $\lambda_4 = 0.1$, and $\lambda_5 = 10^{-3}$. These values determined through empirical experimentation.

Chapter 4

Experiments

This section outlines the experimental setup used to evaluate the performance of our 3D human texture estimation framework. We detail the datasets used, the architecture of our model, and the specifics of our training process.

4.1 Dataset Selection and Utilization

4.1.1 Market-1501 Dataset

Source: The Market-1501 dataset [65] is a widely recognized dataset in the domain of human image analysis.

Composition: It comprises images of 1501 person identities, captured in a variety of poses and from multiple camera views.

Training-Testing Split: We adhere to the same training-testing split as utilized in previous works [13, 12], ensuring consistency and comparability in our evaluations.

4.1.2 DeepFashion

Source: The DeepFashion dataset [66] is a rich repository of in-shop clothes images, known for its diversity in poses and viewing angles.

Relevance: The large pose and view variations in this dataset make it particularly suitable for testing the robustness and versatility of our texture estimation framework.

Data Handling: Following the methodology of previous research [11], we use their curated subset of the DeepFashion dataset. This subset excludes images containing only small body parts, resulting in a total of 20,185 training images and 6,639 testing images.



Figure 4.1: Input/same view and novel view ground-truth images are provided in the first two columns. Other columns show the baseline-trained results with multi-view consistency and trained without multi-view consistency (with single-view images). Blue arrows show that models trained with single-view images better reconstruct the same view but not the novel one. Therefore, instead of only evaluating the models based on the reconstruction of the same view, which misses the point of 3D models, we additionally evaluate the models from a novel view.

4.2 The Overall Architecture

Our model’s architecture is a well-orchestrated system, combining deformable and traditional convolution layers with an attention-based encoder-decoder structure. The design facilitates precise texture mapping, robust feature extraction, and effective multi-scale processing, crucial for high-quality 3D human texture estimation.

The detailed structure, starting from initial offset determination to the final texture and mask predictions, highlights the model’s sophistication and its ability to handle the complexities inherent in human texture estimation tasks.

4.2.1 Deformable Convolution Layers

Initial Offsets: The deformable convolution layers receive initial offsets from a deep attention-based network. These offsets are crucial for accurately adapting the convolution process to the specific needs of the input image.

Refinement Module: The process begins with the refinement module, where the deformable convolutions first apply these offsets, setting the stage for precise texture mapping.

4.2.2 Attention-Based Network Structure

Encoder-Decoder Architecture: The network is designed with an encoder-decoder architecture. At each scale of this architecture, an attention block, as introduced by [12], is employed to enhance feature extraction and processing.

Convolution Layers: There are 6 convolution layers, each having a filter size of 3×3 and a channel size of 128.

Downsampling and Upsampling: Downsampling occurs after the second and

third convolution layers, while upsampling is implemented after the fourth and fifth layers.

Encoder-Decoder Division: The first three convolution layers form the encoder, and the remaining three constitute the decoder.

Skip Connections: There are skip connections between the encoder and decoder at each attention block, integrating multi-scale features effectively.

4.2.3 Deformable Convolution Specifics

Convolution Layer for Offsets: The initial offsets are processed through a convolution layer to produce 18 ($2 \times 3 \times 3$) channels (corresponding to the x, y offsets for each pixel in a 3×3 kernel), which are then utilized in the deformable convolution layer.

Dimensions of Offsets and Input: The offsets have dimensions of $b \times 18 \times 128 \times 128$, and the input has dimensions of $b \times 4 \times 128 \times 64$, where 4 represents the RGB image with part segmentation concatenated.

Output Feature Maps: From the input image with a spatial dimension of 128×64 , the model outputs feature maps with a spatial dimension of 128×128 .

4.2.4 Additional Architectural Features

Convolutional Layers Specifications: Convolution layers use a kernel size of 3, stride of 1, and padding of 1, with the channel dimensions transitioning as $\{4, 128, 128, 128, 6\}$.

Skip Connection from Deep Network: There is a skip connection from the output of the deep attention network to the refinement module. We concatenate the predictions from the deep network with the deformable convolution’s output.

Final Channel Size and Activation Functions: The final output has a channel size of 6, corresponding to RGB predictions (3 channels), UV flow predictions (2 channels), and fusion mask prediction (1 channel). The RGB and UV flow predictions utilize a *tanh* activation function, while mask predictions employ a *sigmoid* activation function.

Post-Convolution Processes: Following each convolution layer, batch normalization and ReLU layers are applied.

4.2.5 The Confidence Model

The confidence model consists of six convolution layers, each having a filter size of 3×3 and a channel size of 128. This configuration enables the model to capture detailed features relevant for estimating confidence levels. Following each convolution layer, batch normalization and ReLU activation layers are applied. These steps ensure stable training and non-linear processing of features. Similar to the main architecture, the confidence model employs an encoder-decoder structure. The first three convolution layers serve as the encoder, and the subsequent three layers function as the decoder. Downsampling occurs after the second and third layers, reducing the spatial dimensions and increasing the representational power of the network. Upsampling is performed after the fourth and fifth layers, restoring the spatial dimensions. Skip connections between the encoder and decoder are implemented at each scale. These connections ensure that features from the encoder are effectively integrated with the decoder, enhancing the fidelity of the reconstructed confidence map. The final convolution layer reduces the channel size to 1, preparing the output for confidence estimation. The softplus activation function is then applied, ensuring that the output confidence values are positive. This network is specifically utilized during the training phase and is not required during inference.

4.2.6 Training Specifications

The model is trained using the Adam optimizer. This optimizer is chosen for its efficiency and effectiveness in handling large datasets and complex architectures.

Training Parameters:

1. **Batch Size:** The training is conducted with a batch size of 16, balancing computational efficiency and memory usage.
2. **Learning Rate:** A constant learning rate of 1×10^{-3} is maintained throughout the training process.
3. **Betas:** The betas parameters for the Adam optimizer are set to 0.9 and 0.999, which are standard values that provide stable convergence.

The model is trained for a total of 200 epochs, providing sufficient time for the network to learn and adapt to the complexities of the texture estimation task. There is no learning rate scheduling employed. The learning rate remains constant throughout the training duration to maintain consistency and simplicity in the training process.

4.3 Evaluation Metrics

In line with previous research in the field, we employ a variety of metrics to evaluate the performance of our 3D human texture estimation model. These metrics are designed to assess different aspects of the model’s output in comparison to the ground truth.

SSIM (Structural Similarity Index Measure): SSIM [14] is used to assess the pixel-level similarity between the output and ground-truth images. It measures the structural integrity and visual quality of the rendered textures.

LPIPS (Learned Perceptual Image Patch Similarity): LPIPS [15] evaluates the feature-level similarity, offering an insight into the perceptual resemblance between the model’s output and the ground truth.

Cosine Similarities of Person ReID Features: This metric calculates the high-level semantic similarity, specifically focusing on whether the rendered human is recognized as the same person in the ground-truth image. We compute this metric using two different networks: PCB [16] and TorchReid [67], resulting in two separate metrics: *CosSim* and *CosSim-R*.

4.3.1 Evaluation Across Different Views

4.3.1.1 Traditional Evaluation Protocol

Previously, these metrics have been reported only between the input and rendered images from the same view (SV). This approach primarily evaluates the visible texture estimates, as it measures the texture generated from the same viewpoint as the input.

4.3.1.2 Extension to Novel View Evaluation (NV)

To address the limitation of the traditional evaluation protocol, we also measure these metrics from a novel view (NV) for which we have ground-truth data. This novel approach involves rendering the estimated 3D human texture based on camera views and pose estimates different from the input image. For each person in the dataset, we evaluate the model’s output across all other available views, providing a more comprehensive assessment of the model’s ability to accurately estimate textures from various perspectives.

4.3.1.3 Evaluating Reliability of Novel View Results

We experiment with the reliability of novel view results by comparing baseline methods trained with and without multi-view consistency. Multi-view consistency is broadly used on datasets that contain multiple images taken from different views of the same object [11, 12]. One view is input in these settings, and loss functions are calculated on the same input view and a novel view. The multi-view consistency improves texture estimation quality for the invisible regions because networks receive gradients from the novel image target for invisible parts in the input image.

Table 4.1 compares our baseline models, trained with multi-view and single-view data, which refers to using multi-view consistency and not, respectively. The traditional evaluation results on the same view show that the single-view baseline model achieves significantly better scores. This is because it is tuned to estimate the visible areas perfectly, producing better results on the same but worse on a novel view, as seen in Fig. 4.1. The quality of renderings in novel views is usually more valuable and the primary purpose of 3D models. Table 4.1 shows the evaluation of the quality of novel view renderings. This setup indicates the limitation of previously used evaluation protocol by 3D human texture estimation models. In this work, we are interested in improving the quality of both the same and novel views, and we keep track of both of these metrics.

This evaluation methodology, encompassing both SV and NV assessments, offers a more holistic view of the model’s performance. It allows us to evaluate not only the accuracy of texture estimates for visible parts of the image but also the model’s capability to reliably predict textures for unseen regions. This dual approach ensures a thorough and rigorous evaluation of the model’s effectiveness in various real-world scenarios.

	SSIM $\uparrow$		LPIPS $\downarrow$		CosSim $\uparrow$		CosSim-R $\uparrow$	
	SV	NV	SV	NV	SV	NV	SV	NV
Multi-view	0.7422	0.6535	0.1154	0.2040	0.5747	0.4943	0.5422	0.4736
Single-view	0.7706	0.6494	0.0963	0.2150	0.5823	0.4809	0.5496	0.4585

Table 4.1: Comparisons of models trained with multi-view consistency and without multi-view consistency (single-view images). Results in bold indicate the best in each column.

	SSIM $\uparrow$		LPIPS $\downarrow$		CosSim $\uparrow$		CosSim-R $\uparrow$	
Method	SV	NV	SV	NV	SV	NV	SV	NV
HPBTT [11]	0.7380	0.6496	<u>0.1148</u>	0.2156	0.5336	0.4697	0.5077	0.4508
RSTG [13]	0.6735	0.6283	0.1778	0.2421	0.5282	0.4717	0.4924	0.4454
TexGlo [68]	0.6658	-	0.1776	-	0.5408	-	0.5048	-
Texformer [12]	<u>0.7422</u>	<u>0.6535</u>	0.1154	<u>0.2040</u>	<u>0.5747</u>	<u>0.4943</u>	<u>0.5422</u>	<u>0.4736</u>
Ours	0.7611	0.6544	0.1003	0.2040	0.5858	0.4963	0.5538	0.4758

Table 4.2: Evaluation results of our method on Market-1501 dataset and state-of-the-art competing methods. Results in bold indicate the best in each column and underscored results indicate the second.

4.4 Comparison with the State-of-the-art

We benchmark our method against leading state-of-the-art 3D human texture estimation methods, including HPBTT [11], RSTG [13], TexGlo [68], and Texformer [12]. These methods represent the latest advancements in the field and provide a comprehensive context for evaluating our approach.

We present quantitative results for both Same-view (SV) and Novel-view (NV) scenarios in Table 4.2. For NV results, we utilize the released models of RSTG and Texformer to ensure a fair comparison. We conduct comparisons on the Market-1501 dataset, ensuring that all methods are evaluated under consistent conditions. For HPBTT, we retrain the method using our data split protocol for consistency, utilizing their publicly released code.

4.4.1 Performance Analysis of Competing Methods

HPBTT [11] outputs texture flow with a regular convolutional neural network that takes body segmentation and pose as input. Since the method outputs texture flow, the colors match the input image, which results in a relatively good performance on SSIM and LPIPS metrics for the same view. On LPIPS, HPBTT even achieves the second-best result. However, its results are poor when measured with CosSim and CosSim-R metrics due to the artifacts in the final renderings. RSTG [13] and TexGlo [68] generate results in fewer artifacts, which leads to relatively better performance on CosSim and CosSim-R metrics. On the other hand, their results do not match the input images in colors and fine details, which causes significantly worse SSIM and LPIPS scores. Texformer [12] is closest to our work with good metrics overall.

Our proposed method outperforms the competing methods across all metrics, both for SV and NV evaluations. This performance is achieved with slightly more parameters, see Table 4.3, than Texformer and significantly fewer parameters than the other benchmarked methods. We effectively balance texture fidelity, color accuracy, and the reduction of artifacts, leading to a comprehensive improvement in both visible and invisible texture estimations. This superior performance, particularly in novel view evaluations, underscores the effectiveness of our approach in accurately estimating 3D human textures across different viewing conditions.

Method	Params. (M)
HPBTT	42.3
RSTG	13.4
TexGlo	16.1
Texformer	7.6
Ours	8.2

Table 4.3: Comparison of Parameter Sizes Across Different Methods

In Fig. 4.2, we showcase the qualitative results of our method against leading 3D human texture estimation methods. The HPBTT model tends to produce results with significant artifacts, particularly in regions not visible in the input

image. In contrast, RSTG reduces these artifacts but often outputs blurry predictions lacking fine details. Texformer shows an improvement over both HPBTT and RSTG, yet it sometimes struggles with high fidelity in complex patterns, like checkered shirts. Our method, on the other hand, excels in producing high-quality results with fine details, accurately reproducing complex patterns and maintaining color fidelity. This superiority is especially evident in challenging scenarios where other methods falter, demonstrating our approach’s advancements in capturing intricate details and achieving high realism in texture estimations.

4.4.2 User Study

In our user study, we evaluated the effectiveness of our method compared to Texformer on the Market-1501 dataset. We engaged 20 users to assess 30 samples, using an A/B test format. Participants were provided with input images and their corresponding predictions, which were rotated 360° along the azimuth in GIF format. This approach allowed users to easily spot any inconsistencies in the textures.

To ensure unbiased evaluations, the left-right order of our method and Texformer in the presentation was randomized. Participants were asked to judge which result was superior in terms of fidelity to the input image and overall realism and quality.

Our method was favored by users over Texformer 71.25% of the time, where a 50% selection rate would indicate no preference. These user study results align with the metrics and qualitative evaluations reported earlier, reinforcing the effectiveness of our approach in 3D human texture estimation.

4.4.3 Comparison on DeepFashion

On the DeepFashion dataset, we focused on comparing HPBTT [11], Texformer [12], and our method, as they emerged as the top-3 methods from the

	SSIM $\uparrow$		LPIPS $\downarrow$		CosSim $\uparrow$		CosSim-R $\uparrow$	
Method	SV	NV	SV	NV	SV	NV	SV	NV
HPBTT [11]	0.6506	0.6391	0.3527	0.3669	0.6193	<u>0.5808</u>	0.6122	<u>0.5717</u>
Texformer [12]	<u>0.7354</u>	0.5479	<u>0.1173</u>	<u>0.3160</u>	<u>0.7595</u>	0.5536	<u>0.7429</u>	0.5451
Ours	0.8358	<u>0.5995</u>	0.0871	0.2689	0.7728	0.6027	0.7582	0.5897

Table 4.4: Evaluation results of our method on DeepFashion dataset and state-of-the-art competing methods. Results in bold indicate the best in each column and underscored results indicate the second.

Market-1501 dataset analysis. In Table 4.4, it’s evident that our method demonstrates a more substantial improvement on this dataset compared to others. Notably, HPBTT shows a slightly better score in the NV SSIM metric. This is consistent with its performance on the Market-1501 dataset, where HPBTT’s texture flow approach results in pixel colors closely matching the input, favoring SSIM scores. However, it’s important to note that SSIM can sometimes prefer blurred images, which, while matching well at the pixel level, may not align with human perception. LPIPS is considered a more accurate metric in this regard [15]. Despite its SSIM performance, HPBTT suffers in other metrics due to artifacts in its final renderings. This issue is apparent in Fig. 4.3, where HPBTT’s outputs for novel views appear blurred. In contrast, our method not only achieves significantly better LPIPS scores but also shows marked improvements over Texformer. The superior performance of our method on the DeepFashion dataset could be attributed to its more varied poses, highlighting the effectiveness of our approach in handling such challenges. This result underscores our method’s adaptability and robustness in 3D human texture estimation across diverse datasets.

4.4.4 Comparison on Pose-guided Human Image Generation

We conducted a comparative analysis with CoordInpaint [1], a method specializing in coordinate-based texture inpainting for pose-guided human image generation. CoordInpaint operates through two distinct pipelines. First pipeline begins with estimating dense poses using the Dense Pose method [43], which are then

converted to SMPL coordinates. This pipeline focuses on obtaining a complete body texture through an inpainting network, resulting in a texture map that can render humans in various poses. The second pipeline involves rendering images for a target pose and refining the results further in the image space. Unlike the first, this pipeline does not produce a 3D model but instead focuses on image space refinement.

Our method’s comparison is pertinent to the output of CoordInpaint’s first pipeline, as it aligns with our focus on 3D texture mapping. The comparative results are illustrated in Fig. 4.4. Given that CoordInpaint achieves its notable results primarily through its second pipeline, we chose not to include it in our main comparative analysis. Our focus is on models directly outputting 3D textures, making the first pipeline of CoordInpaint a more relevant point of comparison.

4.5 Ablation Study

In this section, we present an ablation study to elucidate the impact of each proposed contribution within our framework. The results of this study are detailed in Table 4.5 and visually represented in Fig. 4.5. Our baseline model (BL) serves as the foundational comparison point. This model incorporates solely the deep network depicted in Fig. 3.1, but it excludes the final branch, namely, the deformable convolution-based module. Furthermore, BL is exclusively trained using the base losses as delineated in Sec. 3.2. Subsequently, we assess the effectiveness of each proposed modification in isolation to quantify their individual contributions to the overall performance of our model.

4.5.1 Refinement Module (BL+Refine)

Incorporating the deformable-based refinement module leads to improvements across almost all metrics as seen in Table 4.5. The refinement module particularly excels in complex scenarios, such as accurately reproducing shirt stripes

	SSIM $\uparrow$		LPIPS $\downarrow$		CosSim $\uparrow$		CosSim-R $\uparrow$	
	SV	NV	SV	NV	SV	NV	SV	NV
BL(Baseline)	0.7422	0.6535	0.1154	0.2040	0.5747	0.4943	0.5422	0.4736
BL + Refine	0.7490	0.6511	0.1038	0.2046	0.5887	0.4976	0.5568	0.4775
BL + URL	0.7560	0.6540	0.1053	0.2020	0.5813	0.4802	0.5481	0.4767
BL + Cycle	0.7448	0.6503	0.1107	0.2027	0.5837	0.4968	0.5502	0.4761

Table 4.5: Ablation Study on Market-1501 dataset.

(Fig. 4.5). Fig. 4.6 illustrates the adaptive nature of deformable convolution through offsets visualization for selected UV map points, underscoring its suitability for the texture estimation task.

4.5.2 Uncertainty-Based Reconstruction Loss (BL+URL)

The addition of URL notably enhances SSIM and LPIPS metrics, see Table 4.5, indicating its effectiveness in providing pixel-level reconstruction loss. This results in superior color fidelity, as observed in the second example of Fig. 4.5.

4.5.3 Cycle Consistency Loss (BL+Cycle)

The integration of cycle consistency loss significantly boosts the model’s ability to predict invisible regions (Table 4.5). A notable example includes a more realistically rendered backpack from merely its straps (illustrated in the third row of Fig. 4.5).

Through this ablation study, we demonstrate the individual contributions of our proposed modifications, confirming that each element plays a pivotal role in enhancing the overall model performance. The study not only validates the effectiveness of each component but also highlights the synergy achieved when these elements are combined in our texture estimation framework.

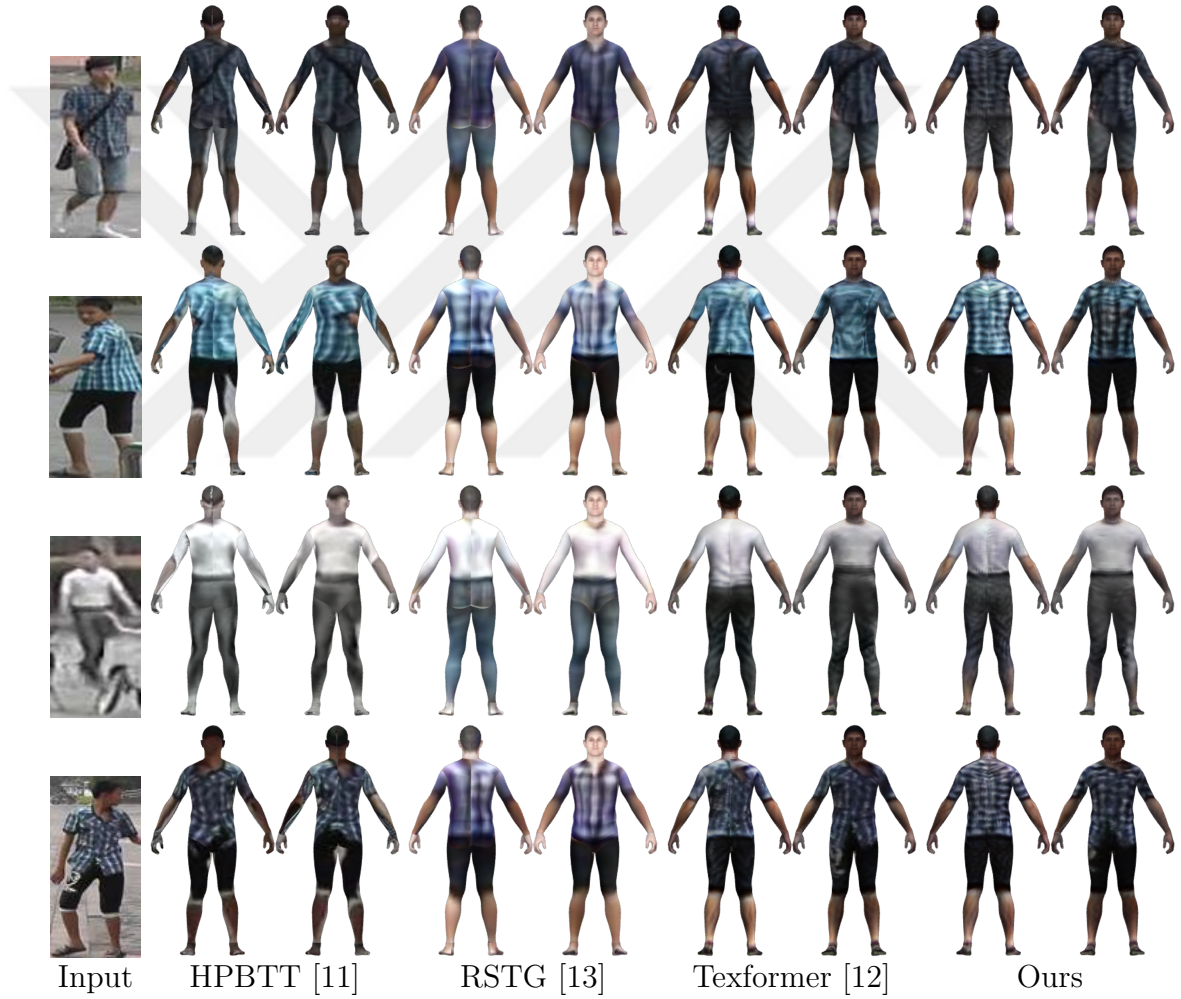


Figure 4.2: Qualitative results of our method and state-of-the-art competing methods on the Market-1501 dataset.

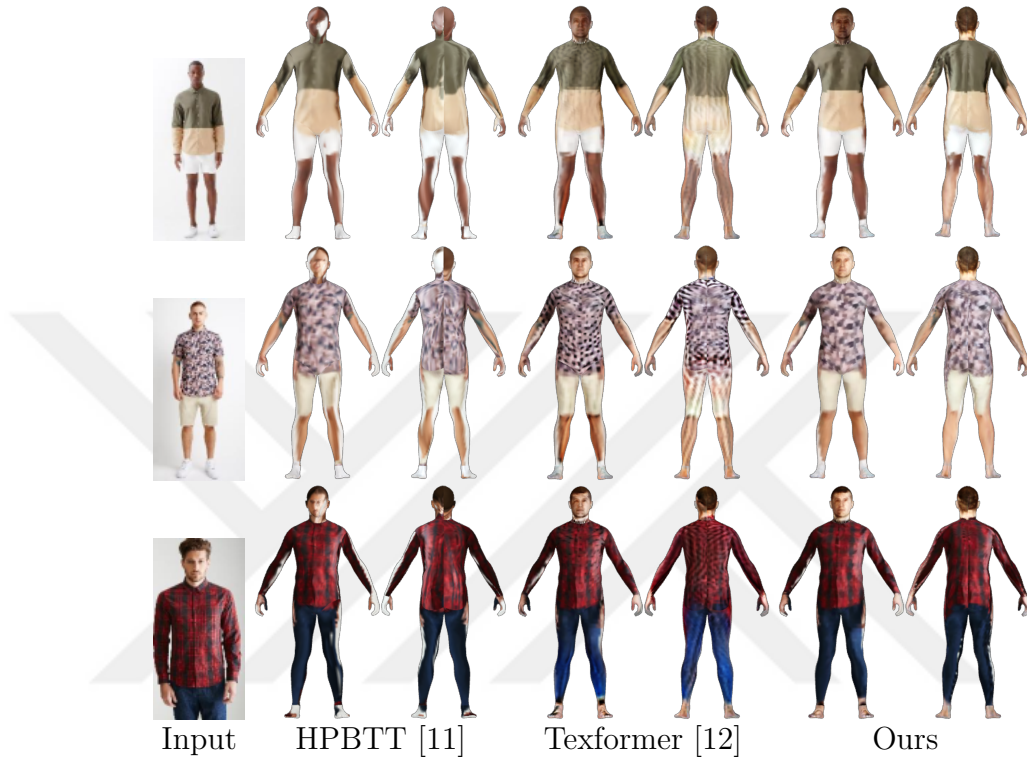


Figure 4.3: Qualitative results of our method and state-of-the-art competing methods on the DeepFashion dataset.

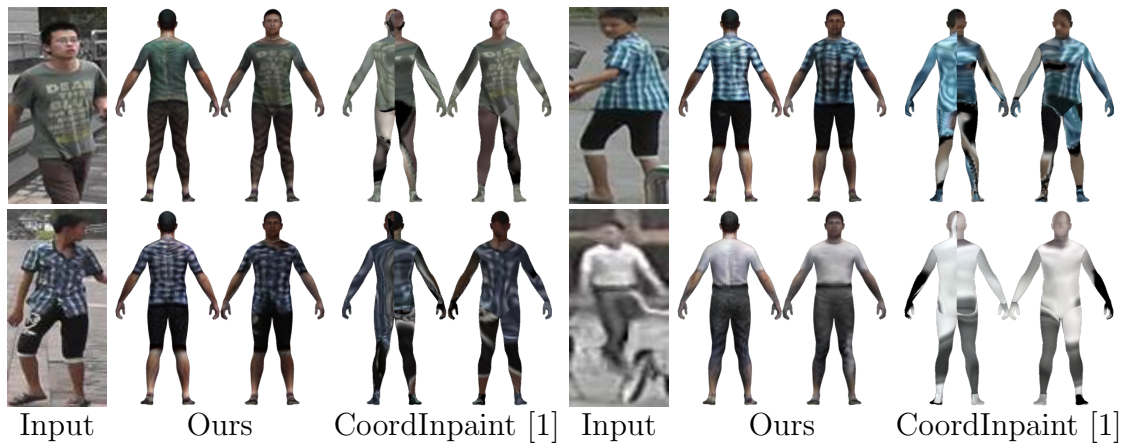


Figure 4.4: Qualitative results of our method and Coordinate-based texture inpainting [1] on the Market-1501 dataset.

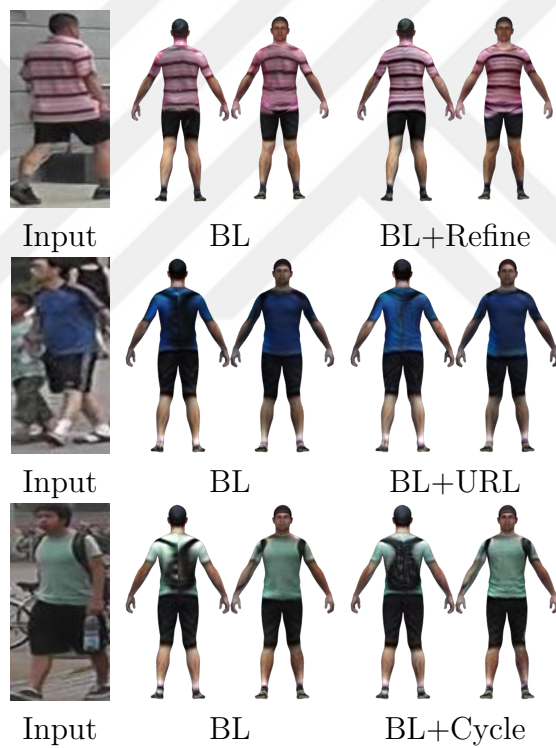


Figure 4.5: Qualitative results of Ablation Study.

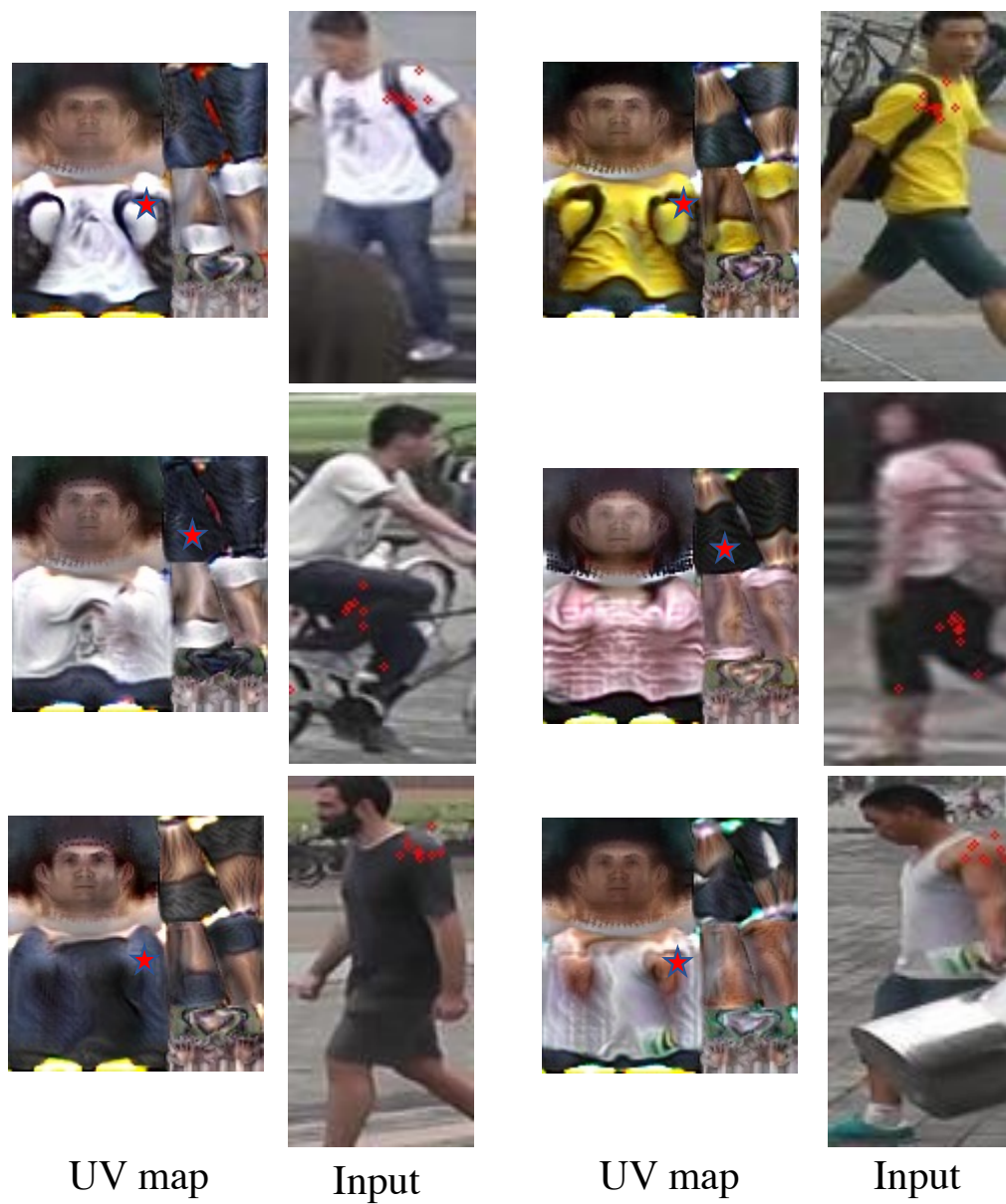


Figure 4.6: Visualization of offsets in input images.

Chapter 5

Analysis and Discussions

Our research presents a novel framework for refining 3D human texture estimation from a single image, marking a significant advancement in this field. The core of our methodology is the integration of a deformable convolution-based refinement module, which adaptively samples input images to optimize texture quality. We also innovatively employ an uncertainty-based reconstruction loss and a novel cycle consistency loss, which collectively contribute to high-fidelity texture estimation.

5.1 Comparative Analysis with State-of-the-Art Methods

Through extensive comparative analysis against state-of-the-art methods like HPBTT, RSTG, TexGlo, and Texformer, our framework demonstrated superior performance. Notably, it excelled in challenging scenarios, such as accurately reproducing complex patterns like checkered shirts, which were problematic for existing methods. The qualitative and quantitative improvements observed in our results underscore our framework’s efficacy in generating detailed and realistic textures.

Comprehensive Performance: Across all metrics and evaluations, our method demonstrated superior performance, achieving high-quality texture estimations in both visible and invisible regions.

Handling of Complex Patterns: One of the most notable strengths of our method was its ability to handle complex patterns and textures, which has been a persistent challenge in the field.

Qualitative Improvements: Visually, our method produced more realistic and detailed textures, as evidenced in qualitative comparisons. This was particularly noticeable in scenarios where other methods struggled, such as in rendering intricate clothing patterns and textures.

5.2 Ablation Study Insights

The ablation study we conducted shed light on the individual contributions of each component in our framework. The addition of the deformable convolution-based refinement module led to enhancements across all metrics. Particularly, it improved the rendering of complex scenarios, highlighting the module’s effectiveness in handling intricate textures. The uncertainty-based reconstruction loss and the cycle consistency loss significantly improved color fidelity and the prediction of invisible regions, respectively.

5.3 Limitations and Future Directions

5.3.1 Limitations

1. **SMPL Model Dependency:** Our framework currently relies on the SMPL human body model [54], which, while robust, has limitations in accurately representing loose-fitting clothing such as long skirts. This reliance

may restrict the model’s applicability to a range of clothing styles.

2. **Challenge with Loose-Fitting Clothes:** The current model’s performance in accurately rendering loose-fitting clothing is an area that requires further refinement. The dynamics and fluidity of such garments present unique challenges that our model may not fully capture.

5.3.2 Future Directions

1. **Integration with Advanced Body Models:** To address the limitation with loose-fitting clothing, future iterations of our method could explore integrating more advanced body models. These models, capable of better handling a variety of clothing styles, would significantly enhance the versatility of our texture estimation framework.
2. **Broader Applicability:** Expanding the model’s capability to accurately represent a wider range of clothing styles, including both tight-fitting and loose garments, would increase its applicability in diverse real-world scenarios, such as fashion design and virtual try-on applications.
3. **Ethical Considerations and Safeguards:** As with any image manipulation technology, there is a potential for misuse. Going forward, it will be essential to develop ethical guidelines and safeguards to prevent the use of such technologies in creating misleading or false information, particularly in sensitive contexts.
4. **Beyond Texture Estimation:** While our current focus is on texture estimation, the methodologies and insights gained from this research could be valuable in related areas, such as facial reconstruction or full-body 3D modeling, encouraging a holistic approach to human image generation.
5. **Interdisciplinary Collaboration:** Future research might benefit from interdisciplinary collaborations, merging insights from fashion, computer graphics, and human-computer interaction. This collaboration could lead

to more nuanced models that account for the complexities of human appearance and garment interaction.

5.4 Discussions

Our research introduces number of advancements in 3D human texture estimation, marked by the integration of a deformable convolution-based refinement module and novel loss functions. This approach has proven effective in accurately rendering complex patterns and textures, outperforming existing state-of-the-art methods in both qualitative and quantitative assessments. The framework’s adaptability in complex scenarios, such as rendering detailed clothing patterns, demonstrates its potential for applications in virtual reality and fashion technology. Additionally, the success in novel view estimations suggests broader applicability in dynamic 3D modeling and augmented reality. While our method represents a significant stride in texture estimation, it also highlights the need for further advancements, particularly in handling diverse clothing styles like loose-fitting garments. This challenge presents an opportunity for future research to integrate more sophisticated body models and fabric dynamics. Furthermore, the ethical implications of image manipulation technology underline the importance of responsible usage, especially in contexts where authenticity is crucial. As the technology evolves, so does the need for ethical guidelines to govern its application. In conclusion, our study not only advances the field of human texture estimation but also sets the stage for future innovations that balance technical excellence with ethical responsibility. The potential for interdisciplinary applications, particularly in fashion and interactive media, opens up exciting possibilities for real-world implementations of this technology.

Chapter 6

Conclusion

Our work presents a pioneering framework that significantly propels the field of 3D human texture estimation from single images. The core innovation lies in the incorporation of a deformable convolution-based refinement module, strategically designed to dynamically enhance texture quality through adaptive sampling of input images. The integration of novel loss functions, specifically the uncertainty-based reconstruction loss and a unique cycle consistency loss, further distinguishes our framework by fostering high-fidelity texture estimation. Our comprehensive evaluation, encompassing both qualitative and quantitative assessments, unequivocally demonstrates superior performance compared to existing state-of-the-art methods. However, it is important to acknowledge a notable limitation tied to our framework’s reliance on the SMPL human body model [54], which may pose challenges in accurately capturing loose-fitting clothing, such as long skirts.

Looking ahead, future iterations of our method hold the potential for refinement through the integration of more advanced body models, thus addressing this constraint and expanding the applicability of our framework to a broader spectrum of clothing styles. In doing so, we aspire to continually enhance the robustness and versatility of our approach in pushing the boundaries of 3D human texture estimation.

Bibliography

- [1] A. Grigorev, A. Sevastopolsky, A. Vakhitov, and V. Lempitsky, “Coordinate-based texture inpainting for pose-guided human image generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’19, pp. 12135–12144, 2019.
- [2] H. Kato, Y. Ushiku, and T. Harada, “Neural 3D mesh renderer,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3907–3916, 2018.
- [3] Z. Huang, Y. Xu, C. Lassner, H. Li, and T. Tung, “ARCH: Animatable Reconstruction of Clothed Humans,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’20, pp. 3093–3102, 2020.
- [4] V. Lazova, E. Insafutdinov, and G. Pons-Moll, “360-degree textures of people in clothing from a single image,” in *Proceedings of the International Conference on 3D Vision*, 3DV ’19, pp. 643–653, IEEE, 2019.
- [5] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima, “SiCloPe: Silhouette-Based Clothed People,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’19, pp. 4480–4490, 2019.
- [6] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, “PiFu: Pixel-aligned implicit function for high-resolution clothed human digitization,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, CVPR ’19, pp. 2304–2314, 2019.

- [7] N. Neverova, R. A. Guler, and I. Kokkinos, “Dense pose transfer,” in *Proceedings of the European Conference on Computer Vision, ECCV ’18*, Lecture Notes in Computer Science, LNIP, Vol. 11219, pp. 123–138, 2018.
- [8] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR ’18*, pp. 7122–7131, 2018.
- [9] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, “Learning to reconstruct 3D human pose and shape via model-fitting in the loop,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV ’19*, pp. 2252–2261, 2019.
- [10] X. Xu, H. Chen, F. Moreno-Noguer, L. A. Jeni, and F. D. l. Torre, “3D human shape and pose from a single low-resolution image with self-supervised learning,” in *Proceedings of the European Conference on Computer Vision, ECCV ’20*, Lecture Notes in Computer Science, LNIP, Vol. 12354, pp. 284–300, 2020.
- [11] F. Zhao, S. Liao, K. Zhang, and L. Shao, “Human parsing based texture transfer from single image to 3D human via cross-view consistency,” in *Proceedings of the International Conference on Neural Information Processing Systems, NeurIPS ’20*, pp. 14326–14337, 2020.
- [12] X. Xu and C. C. Loy, “3D human texture estimation from a single image with transformers,” in *Proceedings of the IEEE International Conference on Computer Vision, ICCV ’21*, pp. 13829–13838, 2021.
- [13] J. Wang, Y. Zhong, Y. Li, C. Zhang, and Y. Wei, “Re-identification supervised texture generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR ’19*, pp. 11846–11856, 2019.
- [14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

- [15] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR ’18, pp. 586–595, 2018.
- [16] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, “Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline),” in *Proceedings of the European Conference on Computer Vision*, ECCV ’18, Lecture Notes in Computer Science, LNIP, Vol. 11219, pp. 480–496, 2018.
- [17] A. Kanazawa, S. Tulsiani, A. A. Efros, and J. Malik, “Learning category-specific mesh reconstruction from image collections,” in *Proceedings of the European Conference on Computer Vision*, ECCV ’18, Lecture Notes in Computer Science, LNIP, Vol. 11219, p. 386–402, 2018.
- [18] W. Chen, H. Ling, J. Gao, E. Smith, J. Lehtinen, A. Jacobson, and S. Fidler, “Learning to predict 3D objects with an interpolation-based differentiable renderer,” in *Proceedings of the Advances in Neural Information Processing Systems*, NeurIPS ’19, pp. 9609–9619, 2019.
- [19] S. Goel, A. Kanazawa, and J. Malik, “Shape and viewpoint without keypoints,” in *Proceedings of the European Conference on Computer Vision*, ECCV ’20, Lecture Notes in Computer Science, LNIP, Vol. 12359, pp. 88–104, Springer, 2020.
- [20] X. Li, S. Liu, K. Kim, S. D. Mello, V. Jampani, M.-H. Yang, and J. Kautz, “Self-supervised single-view 3d reconstruction via semantic consistency,” in *Proceedings of the European Conference on Computer Vision*, ECCV ’20, Lecture Notes in Computer Science, LNIP, Vol. 12359, pp. 677–693, Springer, 2020.
- [21] A. Bhattad, A. Dundar, G. Liu, A. Tao, and B. Catanzaro, “View generalization for single image textured 3D models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’21, pp. 6081–6090, 2021.

- [22] A. Dundar, J. Gao, A. Tao, and B. Catanzaro, “Fine detailed texture learning for 3D meshes with generative models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [23] A. Dundar, J. Gao, A. Tao, and B. Catanzaro, “Progressive learning of 3d reconstruction network from 2D GAN data,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, In press.
- [24] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, “Detailed human avatars from monocular video,” in *Proceedings of the International Conference on 3D Vision*, 3DV ’18, pp. 98–109, IEEE, 2018.
- [25] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, “Learning to reconstruct people in clothing from a single RGB camera,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’19, pp. 1175–1186, 2019.
- [26] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, “Video based reconstruction of 3D people models,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR ’18, pp. 8387–8397, 2018.
- [27] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll, “Multi-Garment Net: Learning to dress 3D people from images,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV ’19, pp. 5420–5430, 2019.
- [28] A. Mir, T. Alldieck, and G. Pons-Moll, “Learning to transfer texture from clothing images to 3D humans,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’20, pp. 7023–7034, 2020.
- [29] T. Zhi, C. Lassner, T. Tung, C. Stoll, S. G. Narasimhan, and M. Vo, “TexMesh: Reconstructing detailed human texture and geometry from RGB-D video,” in *Proceedings of the European Conference on Computer Vision*, ECCV ’20, Lecture Notes in Computer Science, LNIP, Vol. 12359, pp. 492–509, Springer, 2020.

- [30] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou, “Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’21, pp. 9054–9063, 2021.
- [31] S.-Y. Su, F. Yu, M. Zollhoefer, and H. Rhodin, “A-NeRF: Articulated Neural Radiance Fields for learning human shape, appearance, and pose,” in *Proceedings of the Advances in Neural Information Processing Systems*, NeurIPS ’21, 2021.
- [32] W. Jiang, K. M. Yi, G. Samei, O. Tuzel, and A. Ranjan, “NeuMan: Neural human radiance field from a single video,” *arXiv preprint arXiv:2203.12575*, 2022.
- [33] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “NeRF: Representing scenes as neural radiance fields for view synthesis,” *Communications of the ACM*, vol. 65, no. 1, p. 99–106, 2021.
- [34] R. Martin-Brualla, N. Radwan, M. S. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, “Nerf in the wild: Neural radiance fields for unconstrained photo collections,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’21, pp. 7210–7219, 2021.
- [35] C. Guo, T. Jiang, X. Chen, J. Song, and O. Hilliges, “Vid2Avatar: 3D avatar reconstruction from videos in the wild via self-supervised scene decomposition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’23, pp. 12858–12868, 2023.
- [36] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, “Tex2Shape: Detailed full human body geometry from a single image,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV ’19, pp. 2293–2303, 2019.
- [37] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, “DeepHuman: 3D human reconstruction from a single image,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, CVPR ’19, pp. 7739–7749, 2019.

- [38] D. Svitov, D. Gudkov, R. Bashirov, and V. Lempitsky, “Dinar: Diffusion inpainting of neural textures for one-shot human avatars,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, CVPR ’23, pp. 7062–7072, 2023.
- [39] S. Saito, T. Simon, J. Saragih, and H. Joo, “PiFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’20, pp. 84–93, 2020.
- [40] T. He, J. Collomosse, H. Jin, and S. Soatto, “Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction,” in *Proceedings of Advances in Neural Information Processing Systems*, NeurIPS ’20, pp. 9276–9287, 2020.
- [41] Y. Cao, G. Chen, K. Han, W. Yang, and K.-Y. K. Wong, “JIFF: Jointly-aligned Implicit Face Function for high quality single view clothed human reconstruction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’22, 2022.
- [42] A. Pang, X. Chen, H. Luo, M. Wu, J. Yu, and L. Xu, “Few-shot neural human performance rendering from sparse RGBD videos,” *arXiv preprint arXiv:2107.06505*, 2021.
- [43] R. A. Güler, N. Neverova, and I. Kokkinos, “DensePose: Dense human pose estimation in the wild,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR ’18, pp. 7297–7306, 2018.
- [44] Y. Wang, Q. Sun, W. Wang, J. Ling, Z. Cai, R. Xie, and L. Song, “Learning dense uv completion for human mesh recovery,” *arXiv preprint arXiv:2307.11074*, 2023.
- [45] Y. Jafarian, T. Y. Wang, D. Ceylan, J. Yang, N. Carr, Y. Zhou, and H. S. Park, “Normal-guided garment uv prediction for human re-texturing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’23, pp. 4627–4636, 2023.

- [46] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, “Deformable convolutional networks,” in *Proceedings of the IEEE International Conference on Computer Vision, ICCV ’17*, pp. 764–773, 2017.
- [47] X. Zhu, H. Hu, S. Lin, and J. Dai, “Deformable ConvNets V2: More deformable, better results,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR ’19*, pp. 9308–9316, 2019.
- [48] M. Mardani, G. Liu, A. Dundar, S. Liu, A. Tao, and B. Catanzaro, “Neural FFTs for universal texture image synthesis,” in *Advances in Neural Information Processing Systems*, vol. 33 of *NeurIPS ’00*, pp. 14081–14092, 2020.
- [49] A. Grigorev, K. Isakov, A. Ianina, R. Bashirov, I. Zakharkin, A. Vakhitov, and V. Lempitsky, “Stylepeople: A generative model of fullbody human avatars,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR ’21*, pp. 5151–5160, 2021.
- [50] Y. Lan, X. Meng, S. Yang, C. C. Loy, and B. Dai, “Self-supervised geometry-aware encoder for style-based 3d gan inversion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR ’23*, pp. 20940–20949, 2023.
- [51] J. Gao, T. Shen, Z. Wang, W. Chen, K. Yin, D. Li, O. Litany, Z. Gobic, and S. Fidler, “Get3d: A generative model of high quality 3d textured shapes learned from images,” in *Advances In Neural Information Processing Systems*, vol. 35 of *NeurIPS ’22*, pp. 31841–31854, 2022.
- [52] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR ’19*, pp. 4401–4410, 2019.
- [53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoeit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proceedings of the 31st Conference on Neural Information Processing Systems, NIPS ’17*, p. 6000–6010, 2017.

- [54] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “SMPL: A skinned multi-person linear model,” *ACM Transactions on Graphics*, vol. 34, no. 6, pp. Article no. 248, 16 pages, 2015.
- [55] L. A. Gatys, A. S. Ecker, and M. Bethge, “Image style transfer using convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR ’16, pp. 2414–2423, 2016.
- [56] H. Huang, W. Yang, X. Chen, X. Zhao, K. Huang, J. Lin, G. Huang, and D. Du, “Eanet: Enhancing alignment for cross-domain person re-identification,” *arXiv preprint arXiv:1812.11369*, 2018.
- [57] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, “Learning from synthetic humans,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR ’17, pp. 4627–4635, 2017.
- [58] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’17, pp. 5967–5976, 2017.
- [59] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, “High-resolution image synthesis and semantic manipulation with conditional gans,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR ’18, pp. 8798–8807, 2018.
- [60] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, “Semantic image synthesis with spatially-adaptive normalization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’19, pp. 2337–2346, 2019.
- [61] A. Dundar, K. Sapra, G. Liu, A. Tao, and B. Catanzaro, “Panoptic-based image synthesis,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’20, pp. 8070–8079, 2020.

- [62] S. Wu, C. Rupprecht, and A. Vedaldi, “Unsupervised learning of probably symmetric deformable 3D objects from images in the wild,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR ’20, pp. 1–10, 2020.
- [63] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS 17, pp. 5580–5590, 2017.
- [64] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, MICCAI 15, Lecture Notes in Computer Science, LNIP, Vol. 9351, pp. 234–241, Springer, 2015.
- [65] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable person re-identification: A benchmark,” in *Proceedings of the IEEE International Conference on Computer Vision*, CVPR ’15, pp. 1116–1124, 2015.
- [66] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “Deepfashion: Powering robust clothes recognition and retrieval with rich annotations,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR ’16, pp. 1096–1104, 2016.
- [67] K. Zhou and T. Xiang, “Torchreid: A library for deep learning person re-identification in PyTorch,” *arXiv preprint arXiv:1910.10093*, 2019.
- [68] X. Xu, H. Chen, F. Moreno-Noguer, L. A. Jeni, and F. De la Torre, “3D human pose, shape and texture from low-resolution images and videos,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4490–4504, 2022.