Shape-Aware Human Pose and Shape Reconstruction Using Multi-View Images

Junbang Liang Ming C. Lin
University of Maryland, College Park
{liangjb,lin}@cs.umd.edu

Abstract

We propose a scalable neural network framework to reconstruct the 3D mesh of a human body from multi-view images, in the subspace of the SMPL model [23]. Use of multi-view images can significantly reduce the projection ambiguity of the problem, increasing the reconstruction accuracy of the 3D human body under clothing. Our experiments show that this method benefits from the synthetic dataset generated from our pipeline since it has good flexibility of variable control and can provide ground-truth for validation. Our method outperforms existing methods on real-world images, especially on shape estimations.

1. Introduction

Human body reconstruction, consisting of pose and shape estimation, has been widely studied in a variety of areas, including digital surveillance, computer animation, special effects, and virtual/augmented environments. Yet, it remains a challenging and popular topic of interest. While direct 3D body scanning can provide excellent and sufficiently accurate results, its adoption is somewhat limited by the required specialized hardware. We propose a practical method that can estimate body pose and shape directly from a small set of images (typically 3 to 4) taken at several different view angles, which can be adopted in many applications, such as Virtual Try-On. Compared to existing scanning-based reconstruction, ours is much easier to use. Compared to previous image-based estimation methods, ours has a higher shape estimation accuracy when the input human body is not within a normal range of body-mass index (BMI) and/or when the body is wearing loose clothing. Furthermore, our framework is flexible in the number of images used, which considerably extends its applicability.

In contrast to many existing methods, we use multi-view images as input. We use the word "multi-view" to refer photos taken of the same person with *similar* poses from different view angles. They can be taken using specialized multi-view cameras, but it is not necessary (Sec. 6.4). Single-view images often lack the necessary and complete information to infer the pose and shape of a human body, due to the nature of projection transformation. Although applying a predefined prior can alleviate this ambiguity, it is still insufficient in several cases, especially when a part of the body is occluded by clothing, or when the pose direction is perpendicular to the camera viewing plane. For example, when the human is walking towards the camera, it can be difficult to distinguish the difference between a standing vs. walking pose using a direct front-view image, while a side-view image could be more informative of the posture. By obtaining information from multiple view angles, the ambiguity from projection can be considerably reduced, and the body shape under loose garments can also be more accurately recovered.

Previous work on pose and shape estimation of a human body (see Sec. 2) mostly rely on optimization. One of the most important metrics used in these methods is the difference between the original and the estimated silhouette. As a result, these methods cannot be directly applied to images where the human wears loose garments, e.g. long coat, evening gown. The key insight of our method is: when estimating a person's shape, how the human body is interacting with the cloth, *e.g.* how a t-shirt is stretched out as pushed by the stomach or the chest, provides more information than the silhouette of the person. So image features, especially those on clothes, play an important role in the shape estimation. With recent advances in deep learning, it is widely believed that the deep Convolutional Neural Network (CNN) structure can effectively capture these subtle visual details as activation values. We propose a multi-view multi-stage network structure to effectively capture visual features on garments from different view angles to more accurately infer pose and shape information.

Given a limited number of images, we incorporate prior knowledge about the human body shape to be reconstructed. Specifically, we propose to use the Skinned Multi-Person Linear (SMPL) model [23], which uses Principal Component Analysis (PCA) coefficients to represent human body shapes and poses. In order to train the model to accurately output the coefficients for the SMPL model, a sufficient amount of data containing ground-truth information

is required. However, to the best of our knowledge, no such dataset exists to provide multiple views of a loosely clothed body with its ground-truth shape parameters (i.e. raw mesh). Previous learning-based methods do not address the shape (geometry) recovery problem [26] or only output one approximation close to the standard mean shape of the human body [19], which is insufficient when recovering human bodies with largely varying shapes. Taking advantage of physically-based simulation, we design a system pipeline to generate a large number of multi-view human motion sequences with different poses, shapes, and clothes. By training on the synthetic dataset with ground-truth shape data, our model is "shape-aware", as it captures the statistical correlation between visual features of garments and human body shapes. We demonstrate in the experiments that the neural network trained using additional simulation data can considerably enhance the accuracy of shape recovery.

To sum up, the key contributions of our work include:

- A learning-based *shape-aware* human body mesh reconstruction using SMPL parameters for both pose and shape estimation that is supervised directly on shape parameters.
- A scalable, end-to-end, multi-view multi-stage learning framework to account for the ambiguity of the 3D human body (geometry) reconstruction problem from 2D images, achieving improved estimation results.
- A large simulated dataset, including *clothed* human bodies and the corresponding ground-truth parameters, to enhance the reconstruction accuracy, especially in shape estimation, where no ground-truth or supervision is provided in the real-world dataset.
- Accurate *shape* recovery *under occlusion of garments* by (a) providing the corresponding supervision and (b) deepening the model using the multi-view framework.

2. Related Work

In this section, we survey recent works on human body pose and shape estimation, neural network techniques, and other related work that make use of synthetic data.

2.1. Human Body Pose and Shape Recovering

Human body recovery has gained substantial interest due to its importance in a large variety of applications, such as virtual environments, computer animation, and garment modeling. However, the problem itself is naturally ambiguous, given limited input and occlusion. Previous works reduce this ambiguity using different assumptions and input data. They consist of four main categories: pose from images, pose and shape from images under tight clothing, scanned meshes, and images with loose clothing.

Pose From Images. Inferring 2D or 3D poses in images of one or more people is a popular topic in Computer Vision and has been extensively studied [31, 42, 43, 54, 55]. We refer to a recent work, VNect by Mehta *et al*. [26] that is able to identify human 3D poses from RGB images in real time using a CNN. By comparison, our method estimates the pose and shape parameters at the same time, recovering the entire human body mesh rather than only the skeleton.

Pose and Shape From Images under Tight Clothing. Previous work [3, 6, 10, 11, 12, 18] use the silhouette as the main feature or optimization function to recover the shape parameters. As a result, these methods can only be used when the person is wearing tight clothes, as shown in examples [41, 47]. By training on images with humans under various garments both in real and synthetic data, our method can learn to capture the underlying human pose and shape based on image features.

Pose and Shape From Scanned Meshes. One major challenge of recovering human body from scanned meshes is to remove the cloth mesh from the scanned human body wearing clothes [34]. Hasler *et al*. [13] used an iterative approach. They first apply a Laplacian deformation to the initial guess, before regularizing it based on a statistical human model. Wuhrer *et al*. [50] used landmarks of the scanned input throughout the key-frames of the sequences to optimize the body pose, while recovering the shape based on the 'interior distance' that helps constrain the mesh to stay under the clothes, with temporal consistency from neighboring frames. Yang *et al*. [51] applies a landmark tracking algorithm to prevent excessive human labor. Zhang *et al*. [53] took more advantages of the temporal information to detect the skin and cloth region. As mentioned before, methods based on scanned meshes are limited: the scanning equipment is expensive and not commonly used. Our method uses RGB images that are more common and thus much more widely applicable.

Pose and Shape from Images under Clothing. Bălan *et al*. [2] are the first to explicitly estimate pose and shape from images of clothed humans. They relaxed the loss on clothed regions and used a simple color-based skin detector as an optimization constraint. The performance of this method can be easily degraded when the skin detector is not helpful, *e.g*. when people have different skin colors or wear long sleeves. However, our method is trained on a large number of images, which does not require this constraint. Bogo *et al*. [4] used 2D pose machines to obtain joint positions and optimizes the pose and shape parameters based on joint differences and inter-penetration error. Lassner *et al*. [21] created a semi-automatic annotated dataset by incorporating a silhouette energy term on SMPLify [4]. They trained a Decision Forest to regress the parameter based on a much more dense landmark set provided by the SMPL model [23] during the optimization. Constraining the silhouette energy effect to a human body parameter subspace can reduce the negative impact from loose clothing, but their annotated data are from the optimization of SMPLify [4], which has

introduced errors inherently. In contrast, we generate a large number of human body meshes wearing clothes, with the pose and shape ground-truth, which can then train the neural network to be "*shape-aware*".

2.2. Learning-Based Pose/Shape Estimations

Recently a number of methods have been proposed to improve the 3D pose estimation with calibrated multi-view input, either using LSTM [46, 29], auto-encoder [36, 45] or heat map refinement [32, 44]. They mainly focus on 3D joint positions without parameterization, thus not able to articulate and animate. Choy *et al*. [7] proposed an LSTM-based shape recovery network for general objects. Varol *et al*. [48] proposed a 2-step estimation on human pose and shape. However, both methods are largely limited by the resolution due to the voxel representation. In contrast, our method outputs the entire body mesh with parameterization, thus is articulated with a high-resolution mesh quality. Also, our method does not need the calibration of the camera, which is more applicable to in-the-wild images. Kanazawa *et al*. [19] used an iterative correction framework and regularized the model using a learned discriminator. Since they do not employ any supervision other than joint positions, the shape estimation can be inaccurate, especially, when the person is relatively over-weighted. In contrast, our model is more shape-aware due to the extra supervision from our synthetic dataset. Recent works [30, 33, 20] tackle the human body estimation problem using various approaches; our method offers better performance in either single- or multi-view inputs by comparison (see Appendix C).

2.3. Use of Synthetic Dataset

Since it is often time- and labor-intensive to gather a dataset large enough for training a deep neural network, an increasing amount of attention is drawn to synthetic dataset generation. Recent studies [5, 52] have shown that using a synthetic dataset, if sufficiently close to the real-world data, is helpful in training neural networks for real tasks. Varol *et al*. [49] built up a dataset (SURREAL) which contains human motion sequences with clothing using the SMPL model and CMU MoCap data [8]. While the SURREAL dataset is large enough and is very close to our needs, it is still insufficient in that (a) the clothing of the human is only a set of texture points on the body mesh, meaning that it is a tight clothing, (b) the body shape is drawn from the CAESAR dataset [37], where the uneven distribution of the shape parameters can serve as a "prior bias" to the neural network, and (c) the data only consists of single view images, which is not sufficient for our training. Different from [5, 49], our data generation pipeline is based on physical simulation rather than pasting textures on the human body, enabling the model to learn from more realistic images where the human is wearing looser garments. Recent works [39, 1] also generate synthetic data to assist training, but their datasets have only very limited variance on pose, shape, and textures to prevent from overfitting. In contrast, our dataset consists of a large variety of different poses, shapes, and clothing textures.

3. Overview

In this section, we give an overview of our approach. First, we define the problem formally. Then, we introduce the basic idea of our approach.

Problem Statement: Given a set of multi-view images, $\mathbf{I}_1 \dots \mathbf{I}_n$, taken for the same person with the same pose, recover the underlying human body pose and shape.

In the training phase, we set $n = 4$, *i.e*. by default we take four views of the person: front, back, left and right, although the precise viewing angles and their orders are not required, as shown in Sec. 4.3. To extend our framework to be compatible with single view images, we copy the input image four times as the input. For more detail about image ordering and extensions to other multi-view input, please refer to Sec. 4.3. We employ the widely-used SMPL model [23] as our mesh representation, for its ability to express various human bodies using low dimensional parametric structures.

As mentioned before, this problem suffers from ambiguity issues because of the occlusions and the camera projection. Directly training on one CNN as the regressor can easily lead to the model getting stuck in local minima, and it cannot be adapted to an arbitrary number of input images. Inspired by the residual network structure [15], we propose a multi-view multi-stage framework (Sec. 4) to address this problem. Since real-world datasets suffer from limited foreground/background textures and ground-truth pose and shape parameters, we make use of synthetic data as additional training samples (Sec. 5) so that the model can be trained to be more shape-aware.

4. Model Architecture

In this section, we describe the configuration of our network model. As shown in Fig. 1, we iteratively run our model for several stages of error correction. Inside each stage, the multi-view image input is passed on one at a time. At each step, the shared-parameter prediction block computes the correction based on the image feature and the input guesses. We estimate the camera and the human body parameters at the same time, projecting the predicted 3D joints back to 2D for loss computation. The estimated pose and shape parameters are shared among all views, while each view maintains its camera calibration and the global rotation. The loss at each step is the sum of the joint loss and the human body parameter loss:

$$L_i = \lambda_0 L_{2Djoint} + \lambda_1 L_{3Djoint} + L_{SMPL} \qquad (1)$$

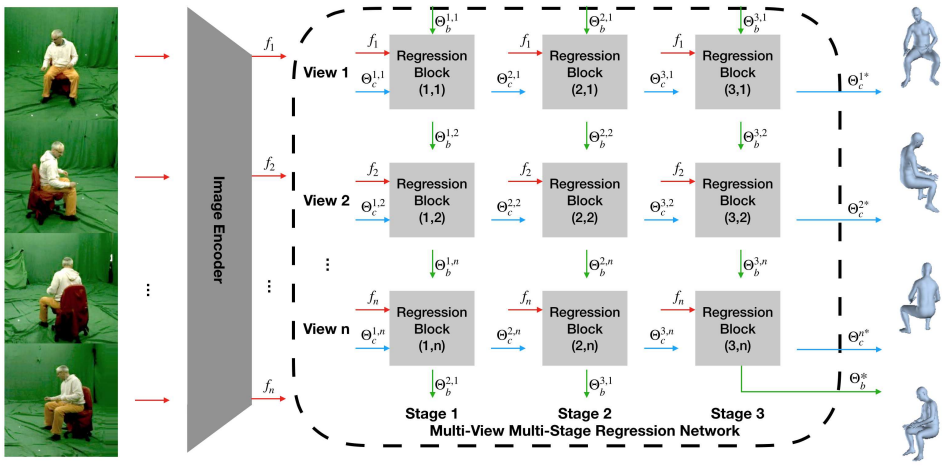

Figure 1: The network structure. Multi-view images are first passed through an image encoder to get feature vectors $f_1, ..., f_n$. With initial guesses of the camera parameters $\Theta_c^{1,i}$ and the human body parameters $\Theta_b^{1,1}$, the network starts to estimate the parameters stage by stage and view by view. Each regression block at the i^{th} stage and the j^{th} view regresses the corrective values from image feature f_j (red) and previous guesses $\Theta_c^{i,j}$ (blue) and $\Theta_b^{i,j}$ (green). The results will be added up to the input values and passed to future blocks. While the new human body parameters (green) can be passed to the next regression block, the view-specific camera parameters (blue) can only be passed to the next stage of the same view. Finally, the predictions of the n views in the last stage are outputted to generate the prediction.

where λ_0 and λ_1 scale the units and control the importance of each term. We use L1 loss on 2D joints and L2 loss on others. L_{SMPL} is omitted if there is no ground-truth.

4.1. 3D Body Representation

We use the Skinned Multi-Person Linear (SMPL) model [23] as our human body representation. It is a generative model trained from human mesh data. The pose parameters are the rotations of 23 joints inside the body, and the shape parameters are extracted from PCA. Given the pose and shape parameter, the SMPL model can then generate a human body mesh consisting of 6980 vertices:

$$\mathbf{X}(\theta, \beta) = \mathbf{W}\mathbf{G}(\theta)(\mathbf{X_0} + \mathbf{S}\beta + \mathbf{P}\mathbf{R}(\theta)) \quad (2)$$

where $\mathbf{X} \in \mathbb{R}^{6980} \times \mathbb{R}^3$ is the computed vertices, $\theta \in \mathbb{R}^{72}$ are the rotations of each joint plus the global rotation, $\beta \in \mathbb{R}^{10}$ are the PCA coefficients, $\mathbf{W}, \mathbf{S}$ and $\mathbf{P}$ are trained matrices, $\mathbf{G}(\theta)$ is the global transformation, $\mathbf{X_0}$ are the mean body vertices, and $\mathbf{R}(\theta)$ is the relative rotation matrix.

For the camera model, we use orthogonal projection since it has very few parameters and is a close approximation to real-world cameras when the subject is sufficiently far away, which is mostly the case. We project the computed 3D body back to 2D for loss computation:

$$\mathbf{x} = s\mathbf{X}(\theta, \beta)\mathbf{R}^T + \mathbf{t} \quad (3)$$

where $\mathbf{R} \in \mathbb{R}^2 \times \mathbb{R}^3$ is the orthogonal projection matrix, s and $\mathbf{t}$ are the scale and the translation, respectively.

4.2. Scalable Multi-View Framework

Our proposed framework uses a recurrent structure, making it a universal model applicable to the input of any number of views. At the same time, it couples the shareable information across different views so that the human body pose and shape can be optimized using image features from all views. As shown in Fig. 1, we use a multi-view multi-stage framework to couple multiple image inputs, with shared parameters across all regression blocks. Since the information from multiple views can interact with each other multiple times, the regression needs to run for several iterative stages. We choose to explicitly express this shared information as the predicted human body parameter since it is meaningful and also contains all of the information of the human body. Therefore the input of a regression block is the corresponding image feature vector and the predicted camera and human body parameters from the previous block. Inspired by the residual networks [15], we predict the corrective values instead of the updated parameters at each regression block to prevent gradient vanishing.

We have n blocks at each stage, where n is the number of views. Since all the input images contain the same human body with the same pose, these n blocks should output the same human-specific parameters but possibly different camera matrices. Thus we share the human parameter output across different views and the camera transformation across different stages of the same view. More specifically, the regression block at the i^{th} stage and the j^{th} view takes an input of $(f_j, \Theta_c^{i,j}, \Theta_b^{i,j})$, and outputs the correction $\Delta\Theta_c^{i,j}, \Delta\Theta_b^{i,j}$, where f_j denotes the j^{th} image feature vector, $\Theta_c^{i,j}$ is the camera matrices and $\Theta_b^{i,j}$ is the human parameters. After that, we pass $\Theta_c^{i+1,j} = \Theta_c^{i,j} + \Delta\Theta_c^{i,j}$ to the next **stage** of the block at the same view, while we pass $\Theta_b^{i,j+1} = \Theta_b^{i,j} + \Delta\Theta_b^{i,j}$ to the next **block** of the chain

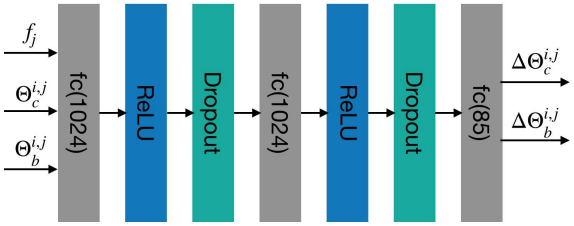

Figure 2: Detailed network structure of the regression block at the i^{th} stage and the j^{th} view. f_j denotes the image feature of the j^{th} view, $\Theta_c^{i,j}$ denotes the camera parameters, and $\Theta_b^{i,j}$ denotes the human body parameters.

(Fig. 1). At last, we compute the total loss as the average of the prediction of all n views in the final stage. Different from static multi-view CNNs which have to fix the number of inputs, we make use of the RNN-like structure in a cyclic form to accept any number of views, and avoid the gradient vanishing by using the error correction framework.

4.3. Training and Inferring

Intuitively we use $n = 4$ in our training process, since providing front, back, left, and right views can often give sufficient information about the human body. We choose a random starting view from the input images to account for the potential correlation between the first view and the initial guess. A specific order of the input views is not required since (a) the network parameters of each regression block are identical, and (b) none of the camera rotation information are shared among different views. To make use of large public single-view datasets, we copy each instance to 4 identical images as our input.

During inference, our framework can adapt to images with any number of views n as shown below. If $n \leq 4$, we use the same structure as used for training. We can pad any of the input images to fill up the remaining views. As each view is independent in terms of global rotation, the choice of which view to pad does not matter. If $n > 4$, we extend our network to n views. Since this is an error-correction structure, the exceeded values introduced by extra steps can be corrected back. Note that the number of camera parameter corrections of each view always remains the same, which is the number of stages.

4.4. Implementation Details

During training, besides our synthetic dataset for enhancing the shape estimation (detailed discussion in Sec. 5), we train on MS-COCO [22], MPI_INF_3DHP [24] and Human3.6M [17] datasets. Each mini-batch consists of half single view and half multi-view samples. Different from HMR [19], we do not use the discriminator. This is because (a) we initialized our parameters as the trained model of HMR [19], (b) the ground-truth given by our dataset serves as the regularization to prevent unnatural pose not captured by joint positions (*e.g.* foot orientations), and most importantly, (c) the ground-truth SMPL parameters from their training dataset does not have sufficient shape variety. Enforcing the discriminator to mean-shape biased dataset will prevent the model to predict extreme shapes. We use 50-layer ResNet-v2 [16] for image feature extraction. The detailed structure inside the regression block is shown in Fig. 2. We fix the number of stages as 3 throughout the entire training and all testing experiments. The learning rate is set to 10^{-5}, and the training lasts for 20 epochs. Training on a GeForce GTX 1080 Ti GPU takes about one day. Our synthetic dataset will be released with the paper.

5. Data Preparation

To the best of our knowledge, there is no public real-world dataset that captures motion sequences of human bodies, annotated with pose and shape (either using a parametric model or raw meshes), with considerable shape variation and loose garments. This lack of data, in turn, forces most of the previous human body estimations to focus only on joints. The most recent work [19] that recovers both pose and shape of human body does not impose an explicit shape-related loss function, so their model is not aware of varying human body shapes. In order to make our model shape-aware under clothing, we need data with ground-truth human body shapes where the garments should be dressed rather than pasted on the skin. A large amount of data is needed for training; sampling real-world data that captures the ground-truth shape parameters is both challenging and time-consuming. We choose an alternate method — using synthesized data. In this section, we propose an automatic pipeline to generate shape-aware training data, to enhance the shape estimation performance.

5.1. Parameter Space Sampling

We employ the SMPL model [23], which contains pose and shape parameters for human body. Pose parameters are rotation angles of joints. To sample meaningful human motion sequences in daily life, we use the CMU MoCap dataset [8] as our pose subspace. The shape parameters are principle component weights. It is not ideal to sample the shape parameters using Gaussian distribution; otherwise there will be many more mean-shape values than extreme ones, resulting in an unbalanced training data. To force the model to be more shape-aware, we choose to uniformly sample values at $[\mu - 3\sigma, \mu + 3\sigma]$ instead, where μ and σ represent the mean value and standard deviation of the shape parameters.

5.2. Human Body Motion Synthesis

After combining CMU MoCap pose data with the sampled shape parameters, it is likely that the human mesh generated by the SMPL model has inter-penetration due to the shape difference. Since inter-penetration is problematic

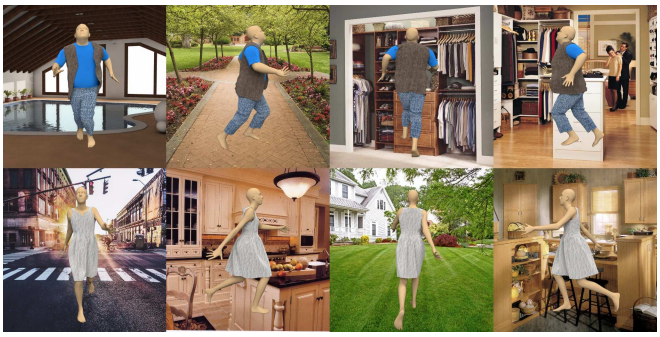

Figure 3: Examples of rendered synthetic images. We use a large number of real-world backgrounds and cloth textures so that the rendered images are realistic and diverse.

for cloth simulation, we design an optimization scheme to avoid it in a geometric sense:

$$\min \|\mathbf{x} - \mathbf{x_0}\| \quad s.t. \quad g(\mathbf{x}) + \epsilon \leq 0 \tag{4}$$

where $\mathbf{x}$ and $\mathbf{x}_0$ stand for the vertex positions, $g(\mathbf{x})$ is the penetration depth, and ϵ is designed to reserve space for the garment. The main idea here is to avoid inter-penetrations by popping vertices out of the body, but at the same time keeping the adjusted distance as small as possible, so that the body shape does not change much. This practical method works sufficiently well in most of the cases.

5.3. Cloth Registration and Simulation

Before we can start to simulate the cloth on each body generated, we first need to register them to the initial pose of the body. To account for the shape variance of different bodies, we first manually register the cloth to one of the body meshes. We mark the relative rigid transformation T of the cloth. For other body meshes, we compute and apply the global transformation, including both the transformation T and the scaling between two meshes.At last, we use the similar optimization scheme described in Sec. 5.2 to avoid any remaining collisions since it can be assumed that the amount of penetration after the transformation is small.

We use ArcSim [28] as the cloth simulator. We do not change the material parameters during the data generation. However, we do randomly sample the tightness of the cloth. We generally want both tight and loose garments in our training data.

5.4. Multi-View Rendering

We randomly apply different background and cloth textures in different sets of images. We keep the same cloth textures but apply different background across different views. We use the four most common views (front, back, left, and right), which are defined w.r.t. the initial human body orientation and fixed during the rendering. We sample 100 random shapes and randomly apply them to 5 pose sequences in the CMU MoCap dataset (slow and fast walking, running, dancing, and jumping). After resolving collisions described in 5.3, we register two sets of clothes on it, one with a dress and the other with a t-shirt, pants, and jacket (Fig. 3). The pose and garment variety is arguably sufficient because (a) they provide most commonly seen poses and occlusions, and (b) it is an auxiliary dataset providing shape ground-truth which is jointly trained with real-world datasets that have richer pose ground-truth. We render two instances of each of the simulated frames, with randomly picked background and cloth textures. Given an average of 80 frames per sequence, we have generated 32,000 instances, with a total number of 128,000 images. We set the first 90 shapes as the training set and the last 10 as the test set. We ensure the generalizability across pose and clothing by coupling our dataset with other datasets with joint annotations (Sec. 4.4).

6. Results

We use the standard test set in Human3.6M and the validation set of MPI_INF_3DHP to show the performance gain by introducing multi-view input. Since no publicly available dataset has ground-truth shape parameters or mesh data, or data contains significantly different shapes from those within the normal range of BMI (*e.g.* overweight or underweight bodies), we test our model against prior work (as the baseline) using the synthetic test set. Also, we test on real-world images to show that our model is more *shape-aware* than the baseline method – qualitatively using online images and quantitatively using photographs taken with hand-held cameras.

Our method does not assume prior knowledge of the camera calibration so the prediction may have a scale difference compared to the ground-truth. There is also extra translation and rotation due to image cropping. To make a fair comparison against other methods, we report the metrics after a rigid alignment, following [19]. We also report the metrics before rigid alignment in the appendix.

6.1. Ablation Study

We conduct an ablation study to show the effectiveness of our model and the synthetic dataset. In the experiments, HMR [19] is fine-tuned with the same learning setting.

6.1.1 Pose Estimation

We tested our model on datasets using multi-view images to demonstrate the strength of our framework. We use *Mean Per Joint Position Error* (MPJPE) of the 14 joints of the body, as well as *Percentage of Correct Keypoints* (PCK) at the threshold of 150mm along with *Area Under the Curve* (AUC) with threshold range 0-150mm [25] as our metrics. PCK gives the fraction of keypoints within an error threshold, while AUC computes the area under the PCK curve, presenting a more detailed accuracy within the threshold.

We use the validation set of MPI_INF_3DHP [19] as an additional test dataset since it provides multi-view input. It is not used for validation during our training. We also evaluated the original test set, which consists of single-view images. Please refer to our appendix in the supplementary document for this comparison result.

Comparison: As shown in Table 1 and 2, under the same training condition, our model in single-view has similar, if not better, results in all experiments. Meanwhile, our model in multi-view achieves much higher accuracy.

Method	MPJPE w/ syn. training	MPJPE w/o syn. training
HMR	60.14	58.1
Ours (single)	58.55	59.09
Ours (multi)	**45.13**	**44.4**

Table 1: Comparison results on Human3.6M using MPJPE. Smaller errors implies higher accuracy.

Method	PCK/AUC/MPJPE w/ syn. training	PCK/AUC/MPJPE w/o syn. training
HMR	86/49/89	88/52/83
Ours (single)	88/52/84	87/52/85
Ours (multi)	**95/63/62**	**95/65/59**

Table 2: Comparison results on MPI_INF_3DHP in PCK/AUC/MPJPE. Better results have higher PCK/AUC and lower MPJPE.

6.1.2 Shape Estimation

To the best of our knowledge, there is no publicly available dataset that provides images with the captured human body mesh or other representation among a sufficiently diverse set of human shapes. Since most of the images-based datasets are designed for joint estimation, we decide to use our synthetic test dataset for large-scale statistical evaluation, and later compare with [19] using real-world images.

Other than MPJPE for joint accuracy, we use the Hausdorff distance between two meshes to capture the shape difference to the ground-truth. The Hausdorff distance is the maximum shortest distance of any point in a set to the other set, defined as follows:

$$d(V_1, V_2) = \max(\hat{d}(V_1, V_2), \hat{d}(V_2, V_1)) \quad (5)$$

$$\hat{d}(V_1, V_2) = \max_{\mathbf{u} \in V_1} \min_{\mathbf{v} \in V_2} \|\mathbf{u} - \mathbf{v}\|^2 \quad (6)$$

where V_1 and V_2 are the vertex set of two meshes in the same ground-truth pose, in order to negate the impact of different poses. Intuitively a Hausdorff distance of d means that by moving each vertex of one mesh by no more than d away, two meshes will be exactly the same.

As shown in Table 3, our model with multi-view input achieves the smallest error values, when compared to two other baselines. After joint-training with synthetic data, all models perform better in shape estimation, while maintaining similar results using other metrics (Table 1 and 2), i.e. they do not overfit. The joint errors of the HMR [19] are fairly good, so they can still recognize the synthesized human in the image. However, a larger Hausdorff distance indicates that they lose precision on the shape recovery.

Method	MPJPE/HD w/ syn. training	MPJPE/HD w/o syn. training
HMR	42/83	89/208
Ours (single)	44/65	102/283
Ours (multi)	**27/53**	**84**/273

Table 3: Comparison results on our synthetic dataset in MPJPE/Hausdorff Distance(HD). Better results have lower values.

Adding our synthetic datasets for training can effectively address this issue and thereby provide better shape estimation. We achieved a much smaller Hausdorff distance (with syn. training) even only using single view. This is because our refinement framework is effectively deeper, aiming at not only the pose but also the shape estimation, which is much more challenging than the pose-only estimation. With the same method, multi-view inputs can further improve the accuracy of shape recovery compared to results using only one single-view image.

6.2. Comparisons with Multi-View Methods

Since other multi-view methods only estimate human poses but not the entire body mesh, we compare the pose estimation results to them in Human3.6M. As shown in Table 4, we achieved state-of-the-art performance even when camera calibration is unknown and no temporal information is provided. As stated in Sec. 6, unknown camera parameters result in a scaling difference to the ground-truth, so the joint error would be worse than what it actually is. After the Procrustes alignment that accounts for this effect, our method achieves the best MPJPE compared to other methods. Another potential source of the error is that our solution is constrained in a parametric subspace, while other methods output joint positions directly. In contrast, our method computes the entire human mesh in addition to joints and the result can be articulated and animated directly.

6.3. Real-World Evaluations

We first conduct a study on how our method performs differently with either single- or multi-view inputs under various conditions. Our test subjects have two poses: standing and sitting, and the model is additionally tested on two sets of variants from the images. One is slightly dimmed, and the other has a large black occlusion at the center of the first image. We use the percentage of errors from common body measurements used by tailors (*i.e.* lengths of neck, arm, leg, chest, waist, and hip), which is obtained using direct tape measurements on the subjects. We report the av-

Method	MPJPE	Known Camera?	Run Time	Temporal Opt?	Articulated?	Shape?
Rhodin *et al*. [35]	-	Yes	0.025fps	Yes	No	Mix-Gaussian
Rhodin *et al*. [36]	98.2	Yes	-	Yes	No	No
Pavlakos *et al*. [32]	56.89	Yes	-	No	No	No
Trumble *et al*. [46]	87.3	Yes	25fps	Yes	No	No
Trumble *et al*. [45]	62.5	Yes	3.19fps	Yes	No	Volumetric
Núñez *et al*. [29]	54.21	Yes	8.33fps	Yes	No	No
Tome *et al*. [44]	52.8	Yes	-	No	No	No
Ours	79.85	No	33fps	No	Yes	Parametric
Ours (PA)	**45.13**					

Table 4: Comparison on Human3.6M with other multi-view methods. Our method has comparable performance with previous work even without the assistance of camera calibration or temporal information. PA stands for Procrustes Aligned results for ours.

Method	Standing	Sitting
HMR [19]	7.72%	7.29%
BodyNet [48]	13.72%	29.30%
Ours (single)	6.58%	10.18%
Ours (multi)	**6.23%**	**5.26%**

Table 5: Comparison results on tape-measured data using average relative errors (lower the better).

(a) The input image. (b) Our result. (c) HMR.

Figure 4: Prediction results compared to HMR. Our model can better capture the shape of the human body. The recovered legs and chest are closer to the person in the image.

erage relative error in Table 5. The detailed errors of each measurement are also provided in the appendix. It is observed that single-view results are affected by the "occluded sitting" case, while the multi-view input can largely reduce the error. The reason why HMR is not impacted is that they uniformly output average human shapes for all input images. We also report results from BodyNet [48]. BodyNet outputs voxelized mesh and needs a time-consuming optimization to output the SMPL parameters. Its accuracy largely depends on the initial guess. Therefore, it resulted in a large amount of errors on the "sitting" case.

We also tested our model on other online images, where no such measurement can be done. As shown in Fig. 4, HMR [19] can predict the body pose but fails on inferring the person's shape. On the contrary, our model not only refines the relative leg orientations but also largely respects and recovers the original shape of the body. More examples are shown in our supplemental document and video.

6.4. Multi-View Input in Daily Life

It is often difficult to have multiple cameras from different view angles capturing a subject simultaneously. Our model has the added benefit of not requiring the multi-view input be taken with the exact same pose. As the model has an error correction structure, it can be applied as long as the poses of the four views are not significantly different. We do not impose any assumptions on the background, so the images can be even taken with a fixed camera and a "rotating" human subject, which is the typically case when the method is used in applications like virtual try-on.

7. Conclusion and Future Work

We proposed a novel multi-view multi-stage framework for pose and shape estimation. The framework is trained on datasets with at most 4 views but can be naturally extended to an arbitrary number of views. Moreover, we introduced a physically-based synthetic data generation pipeline to enrich the training data, which is very helpful for shape estimation and regularization of end effectors that traditional datasets do not capture. Experiments have shown that our trained model can provide equally good pose estimation as state-of-the-art using single-view images, while providing considerable improvement on pose estimation using multi-view inputs and a better shape estimation across all datasets.

While synthetic data improves the diversity of human bodies with ground-truth parameters, a more convenient cloth design and registration are needed to minimize the performance gap between real-world images and synthetic data. In addition, other variables such as hair, skin color, and 3D backgrounds are subtle elements that can influence the perceived realism of the synthetic data at the higher expense of a more complex data generation pipeline. With the recent progress in image style transfer using GAN [27], a promising direction is to transfer the synthetic result to more realistic images to further improve the learning result.

Acknowledgement: This work is supported by National Science Foundation and Elizabeth S. Iribe Professorship.

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 3, 14

[2] Alexandru O Bălan and Michael J Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision*, pages 15–29. Springer, 2008. 2

[3] Alexandru O Balan, Leonid Sigal, Michael J Black, James E Davis, and Horst W Haussecker. Detailed human shape and pose from images. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 2

[4] 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 *European Conference on Computer Vision*, pages 561–578. Springer, 2016. 2

[5] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 479–488. IEEE, 2016. 3

[6] Yu Chen, Tae-Kyun Kim, and Roberto Cipolla. Inferring 3d shapes and deformations from single views. In *European Conference on Computer Vision*, pages 300–313. Springer, 2010. 2

[7] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016. 3

[8] CMU. Carnegie-mellon mocap database. created with funding from nsf eia- 0196217, 2003. 3, 5

[9] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 668–683, 2018. 13

[10] Endri Dibra, Himanshu Jain, Cengiz Öztireli, Remo Ziegler, and Markus Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 108–117. IEEE, 2016. 2

[11] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1381–1388. IEEE, 2009. 2

[12] Nils Hasler, Hanno Ackermann, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1823–1830. IEEE, 2010. 2

[13] Nils Hasler, Carsten Stoll, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. Estimating body shape of dressed humans. *Computers & Graphics*, 33(3):211–216, 2009. 2

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. 12

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 5

[17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2014. 5

[18] Arjun Jain, Thorsten Thormählen, Hans-Peter Seidel, and Christian Theobalt. Moviereshape: Tracking and reshaping of humans in videos. In *ACM Transactions on Graphics (TOG)*, volume 29, page 148. ACM, 2010. 2

[19] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 2, 3, 5, 6, 7, 8, 12, 13, 14

[20] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019. 3, 13

[21] 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 Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 3, 2017. 2

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[23] 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. 1, 2, 3, 4, 5

[24] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 5

[25] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 International Conference on*, pages 506–516. IEEE, 2017. 6, 13

[26] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel,

Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):44, 2017. 2, 13

[27] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–59, 2018. 8

[28] Rahul Narain, Armin Samii, and James F O'Brien. Adaptive anisotropic remeshing for cloth simulation. *ACM transactions on graphics (TOG)*, 31(6):152, 2012. 6

[29] Juan Carlos Núñez, Raúl Cabido, José F Vélez, Antonio S Montemayor, and Juan José Pantrigo. Multiview 3d human pose estimation using improved least-squares and lstm networks. *Neurocomputing*, 323:335–343, 2019. 3, 8

[30] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 International Conference on 3D Vision (3DV)*, pages 484–494. IEEE, 2018. 3, 13

[31] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1263–1272. IEEE, 2017. 2, 13

[32] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6988–6997, 2017. 3, 8

[33] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018. 3, 13

[34] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):73, 2017. 2

[35] Helge Rhodin, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *European conference on computer vision*, pages 509–526. Springer, 2016. 8

[36] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–767, 2018. 3, 8

[37] Kathleen M Robinette, Sherri Blackwell, Hein Daanen, Mark Boehmer, and Scott Fleming. Civilian american and european surface anthropometry resource (caesar), final report. volume 1. summary. Technical report, SYTRONICS INC DAYTON OH, 2002. 3

[38] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR 2017-IEEE Conference on Computer Vision & Pattern Recognition*, 2017. 13

[39] Hosnieh Sattar, Gerard Pons-Moll, and Mario Fritz. Fashion is taking shape: Understanding clothing preference based on body shape from online sources. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 968–977. IEEE, 2019. 3

[40] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, page 7, 2017. 13

[41] J Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *BMVC*, volume 3, page 6, 2017. 2

[42] Bugra Tekin, Pablo Marquez Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *International Conference on Computer Vision (ICCV)*, number EPFL-CONF-230311, 2017. 2

[43] Denis Tome, Christopher Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *CVPR 2017 Proceedings*, pages 2500–2509, 2017. 2, 13

[44] Denis Tome, Matteo Toso, Lourdes Agapito, and Chris Russell. Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture. In *2018 International Conference on 3D Vision (3DV)*, pages 474–483. IEEE, 2018. 3, 8

[45] Matthew Trumble, Andrew Gilbert, Adrian Hilton, and John Collomosse. Deep autoencoder for combined human pose estimation and body model upscaling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–800, 2018. 3, 8

[46] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. Total capture: 3d human pose estimation fusing video and inertial sensors. In *Proceedings of 28th British Machine Vision Conference*, pages 1–13, 2017. 3, 8

[47] 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 5236–5246, 2017. 2

[48] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 20–36, 2018. 3, 8

[49] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2017. 3

[50] Stefanie Wuhrer, Leonid Pishchulin, Alan Brunton, Chang Shu, and Jochen Lang. Estimation of human body shape and posture under clothing. *Computer Vision and Image Understanding*, 127:31–42, 2014. 2

[51] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Estimation of human body

shape in motion with wide clothing. In *European Conference on Computer Vision*, pages 439–454. Springer, 2016. 2

[52] Shan Yang, Junbang Liang, and Ming C Lin. Learning-based cloth material recovery from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4383–4393, 2017. 3

[53] Chao Zhang, Sergi Pujades, Michael Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017. 2

[54] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Weaklysupervised transfer for 3d human pose estimation in the wild. In *IEEE International Conference on Computer Vision*, volume 206, page 3, 2017. 2

[55] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *European Conference on Computer Vision*, pages 186–201. Springer, 2016. 2, 13