Chapter 7: Conclusion

In this dissertation, I proposed various methods for reconstructing human body and garment from images or videos, as well as synthesizing new garments given the body mesh. These methods enables shape-aware body recovery from multi-view images, accurate material retrieval using optimization, faithful garment reconstruction together with body and material estimation from videos, efficient and scalable distributed simulation, and fast and realistic garment prediction on human bodies. I have shown their high effectiveness and accuracy in extensive experiments. The proposed work can greatly boost the efficiency and realism of simulation-based virtual try-on systems.

7.1 Summary of Results

To train a shape-aware body estimation model, Chapter 2 introduces a novel multi-view multi-stage framework. This framework is scalable and can take arbitrary number of views as input. Moreover, it describes a training data generation pipeline using physically-based simulation. It is critical for shape estimation and regularization of end-effectors that real-world datasets cannot provide. Experiments show that this method benefits from the simulated dataset generated from this pro-

posed pipeline and outperforms existing methods on real-world images, especially on shape estimations.

Material cloning from real world to virtual world is critical to attaining realism of virtual try-on systems. To enable a first-order estimation of garment materials, Chapter 3 presents a differentiable cloth simulator that can provide the gradients of the cloth simulation function. The gradient of the dynamic collision handling is explicitly derived. In order to obtain gradients of large formulated functions, implicit differentiation is used instead of backpropagating step by step. Experiments show that my backpropagation is two orders of magnitude faster than the baseline method. The proposed differentiable simulation can be used in a number of inverse problems, when the environmental setup is known.

To further combine the two correlated objectives above for an end-to-end method with higher estimation accuracy, Chapter 4 proposes an end-to-end learning model for estimating body and garment material from RGB videos. In order to maximize the multi-tasking benefits, human body and the garment shape are jointly learned as features for the material prediction. To reduce the degree of freedom of cloth due to its highly dynamic and deformable nature of cloth, a two-level auto-encoder to represent garments has been proposed in a hierarchical structure. Using this network, it also becomes possible to smoothly transition the geometry between different garment topologies. During the estimation, a feedback loop is introduced to correctly refine the body estimation using the garment prediction. Experiments show that this proposed system has the highest estimation accuracy and can be easily generalized to unseen input.

Cloth simulations, widely used in computer animation and apparel design, can be computationally expensive for real-time applications. Some parallelization techniques have been proposed for visual simulation of cloth using CPU or GPU clusters and often rely on parallelization using spatial domain decomposition techniques that have a large communication overhead. To avoid a large overhead, Chapter 5 introduces a novel temporal-domain parallelization method for performance-demanding tasks. I parallelize the cloth simulation in temporal domain by using a faster simulation result on coarser meshes as an initialization, resulting in accelerated computation within each temporal computational block and minimal communication overhead. An iterative detail recovery algorithm is designed to minimize the visual artifacts due to the inconsistency between two resolutions. This proposed method has nearly linear scaling on manycore clusters and achieves more runtime performance advantages, when compared to previous parallel methods with a higher number of computing cores.

Finally, to train an efficient and accurate prediction model of garments to provide the final try-on results, Chapter 6 proposes a novel semi-supervised model to learn physically-correct garment draping on a large variation of human body shapes. Several loss functions are incorporated to increase the physical awareness of the network, and a new GCN decoder is used for higher expressiveness. Moreover, a self-correcting optimization method is adopted to minimize the potential energy of the prediction and remove the remaining collision with the human body. Retraining this proposed network with the optimized samples can yield better predictions. Experiments show that this novel method can provide better draping predictions

than previous works in terms of simulated wrinkles and folds, it can also cover a large distribution of body shapes, while achieving higher speed and accuracy compared to physics-based cloth simulation.

7.2 Limitations

There are four main limitations in my proposed methods. First, there is a sim-to-real gap that can hinder the virtual world to reproduce exactly the same visual effect as those in the real world. Common sources of this gap include lighting, human skin texture, mesh resolution, and spatial relation with the background and other objects in the scene. Simulation settings, such as the discretization scheme, the physical governing laws, and the material model, can also affect the final results. When the simulated data is used for training network models, this gap may potentially affect its generalization to unseen real-world images or videos. While higher-quality training data with more complex settings can certainly reduces the gap, the labor cost of doing so increases as well. Although the current results show that models can already perform well when trained with the simulation data, the potential improvements given by more realistic training data is not yet quantified.

Second, the high computational cost of cloth simulation is also a challenge. When simulating garments for learning purposes or optimizations, I want the simulation to be as fast as possible so that I can obtain the maximum amount of training data or the maximum number of iterations within a given amount of time. When the simulation is used for visual demonstration, real-time response from the user

interaction is critical. Although the proposed method in Chapter 5 achieves near real-time performance, the single-frame simulation runtime is not reduced in this framework, leading to a high latency feedback.

Next, a parametric generation model for universal garments is still missing. An ideal garment generation model should be able to account for different variations, including sizes and lengths, topology, fitness to any given body shape, sewing patterns, or even accessory locations and geometry. Chapter 4 introduces the first generation model that unites garments of different topology and sizes using point clouds. However, this model cannot fully represent garments that have multiple layers or self-folding, because it does not store connectivity information between points. Moreover, it does not currently support multiple separated garments and sewing seams for the same reason. However, such geometric information is critical, as it directly affects the physical behavior of the reconstructed and learned garment models.

At last, there remains a large visual disparity and physical difference between simulation results and network prediction results when it comes to garment synthesis. Although the proposed method in Chapter 6 successfully connects simulation realism to the network prediction by introducing dynamical constraints in the training for more realistic predictions, the coverage of the current model is still limited, when compared to all possible combinations of human poses and shapes, garment sizes, topology and materials. Given a specific type of garment, the best model I can get in theory will be able to present different wrinkle formations according to different input of the human body and the garment material. However, it requires

an unrealistic amount of captured or simulated data used for training to cover the entire spectrum spanned by the control variables. As it is practically impossible to include every sample in the training datasets, an effective and general decoupling algorithm between different factors is necessary. But, this problem is currently under-explored.

7.3 Future Work

While the proposed methods have been proven to be effective in the reconstruction and the synthesis tasks regarding garments dressed on human bodies, there are still several potential improvements and challenges that need to be addressed. Below I suggest a few possible future directions regarding each proposed limitation described in Section 7.2.

Training data. More research can be conducted to minimize or mitigate the sim-to-real gap in training data distributions. Chapter 2 and 4 use simulated data for training the models due to ease of collecting ground-truth labels. While cloth simulation provides a large diversity of training samples with ground-truth parameters, a more systematic garment design and registration method are needed to minimize the visual perception difference between real-world images and synthetic ones. In addition, other variables such as hair, skin color, and 3D backgrounds can also influence the perceived realism of the synthetic data, but implementing them requires a more complex data generation pipeline. With the recent progress in image style transfer and translation using GAN [81], a promising direction is to transfer the

appearance of realistic images to the rendered simulation models to further improve the learning results.

Learning algorithms and architectures. A lot of options are adopted regarding network architectures being used in different tasks. Chapter 2 introduces a shared-weight prediction correction framework to integrate multi-view information, while Chapter 4 jointly learns the body and the garment shape together with a feed-back loop and feeds the single-frame features to an LSTM for material estimation. One key problem remains to be solved is to what extent do different network and framework choices affect the final result, and whether or not there is a specialized architecture that is generally beneficial to the capturing task regarding garment properties.

Parametric garment models. More improvements can be made regarding the parametric garment modeling introduced in Chapter 4. First, encoding multi-layer information can be achieved by adding structural prior to the garment model so that more complex structures of garments can be supported. For example, one can encode a tree structure to represent different garment parts from bodice to accessories. Second, adding curvature information such as normal in addition to 3D location can enable support of multi-fold features. A normal vector is extremely helpful to differentiate two pieces of cloth stacked together with different orientations. Taking a step further, encoding the texture space coordinates together can even provide the connectivity information between vertices, potentially allowing easier reconstruction from point clouds to meshes and smoother results.

Human body representation. Currently, the parametric human body model in Chapter 2, 4, and 6 uses only linear blend shapes and principal components to represent thehuman body. An intrinsic drawback of this approach is that it cannot model deformation of the body when encountering geometric constraints such as collisions. It would be of great value to investigate more complex and realistic body models as a basic building block of all estimation and prediction modules.

Visual rendering and synthesis. To further speed up the cloth simulation, I can either introduce vectorization or implement multiple parallelization schemes at the same time. The current simulation implementation in Chapter 3 is not optimized for large-scale vectorized operations, leading to imperfect runtime performance. This issue can be improved by a specialized, optimized simulation system based solely on large tensor operations. Chapter 5 also proposes a method to accelerate the computation, which is the time-domain parallelization. This method can naturally be combined with GPU-based spatial-domain parallelization to reduce the per-frame computation time, which further decreases the latency and increases the scalability in terms of the number of computational units.

Generalization and robustness. Disentanglement of multi-dimensional input factors is the key to boost the generalization ability for the method proposed in Chapter 6. Regarding data representation, principal component analysis (PCA) or tangent space displacement representation can be potential directions to disentangle body pose and shape variations. Another promising future improvement can be training refinement networks conditioned on body pose parameters or the corre-

sponding global transformations. While single-frame draping prediction is far from being solved, extending the current work to learning continuous garment motion is the ultimate goal towards real-time animated virtual try-on. This will require not only the encoding of the geometric constraints, but also the capturing of different temporal patterns dependent on the fabric material type, and can eventually become a differentiable simulation in a learning-based way.