Deep Fashion3D: A Dataset and Benchmark for 3D Garment Reconstruction from Single Images – Supplemental Materials

Heming Zhu[1,2†], Yu Cao[1,3†], Hang Jin[1†], Weikai Chen[4], Dong Du[1,5],
Zhangye Wang[2], Shuguang Cui[1], and Xiaoguang Han[1*]

[1] Shenzhen Research Institute of Big Data,
The Chinese University of Hong Kong, Shenzhen
[2] State Key Lab of CAD&CG, Zhejiang University
[3] Xidian University
[4] Tencent America
[5] University of Science and Technology of China

In this supplemental material, we provide more details, results and analysis in the following aspects: 1) implementation details on the training settings of cloth classification, pose estimation, feature line regression and the design of the adaptable template; 2) experiments demonstrating the effectiveness of the adaptable template; 3) results reconstructed from the in-the-wild images using our baseline method.

1 Implementation Details

In the main paper, we have briefly introduced the losses adopted to train each stage for our proposed approach. In this section, we will describe the detailed training setting of cloth classification, pose estimation, implicit reconstruction along with the loss functions and the hyper parameters used in the feature line regression module.

1.1 Cloth Classification

We build a cloth classification network to infer the type of clothes presented in input images, which consists of a VGG16 feature extractor and a fully connected layer. The network is trained using a standard cross-entropy loss. To enhance the generalization performance of our model, the cloth classification module is trained on a hybrid source combining synthetic images rendered using the point clouds of our dataset, the real images selected from 1000 manually selected frontal images from DeepFashion [2] and DeepFashion2 [1], and the real multi-view images from our Deep Fashion3D. We use 90% of the data for training while leaving the other 10% for testing. Random resized crop as well as random

†The first three authors should be considered as joint first authors.

* Xiaoguang Han is the corresponding author. Email:hanxiaoguang@cuhk.edu.cn.

rotation are employed to augment the data during training. Furthermore, we balanced the number of images used for training by randomly selecting clothes from each category, so that for each cloth category, the number of images involved in the training is nearly the same.

1.2 Pose Estimation

As our adaptable template is built upon SMPL [3], we employ a subset of SMPL parameters to denote the pose of the clothes while setting the parameters irrelevant to the deformation of the clothes, such as global rotation, local rotation parameters for ankle, wrist, neck, etc., to zero. To estimate the partial pose parameters from the image, we build a pose estimation module consisting of a VGG16 feature extractor and a fully connected layer.

We train the pose estimation network by minimizing the loss function $\mathcal{L}_{pose}$ which consists of a cloth pose parameter loss $\mathcal{L}_{param}$ and a pose regularization loss $\mathcal{L}_{reg}$:

$$\mathcal{L}_{pose} = \mathcal{L}_{param} + \lambda_{reg}\ \mathcal{L}_{reg},$$

The pose parameter loss $\mathcal{L}_{param}$ is calculated as the MSE between the ground truth and the predicted pose parameters. We further introduce a regularization loss $\mathcal{L}_{reg}$ that is the squared sum of the pose parameters, aiming to eliminate unexpected rotations. During training, λ_{reg} is set to 1e-5.

1.3 Implicit Surface Reconstruction

We adopt OccNet [4] conditioned on image features for implicit reconstruction. The feature extraction module consists of a pre-trained ResNet-18 which is fine-tuned using the real and synthetic cloth images. A fully connected layer is used to predict whether the input 3D point is inside the surface or not given its coordinate and a latent code of the image. We exactly follow the settings of the original OccNet for training.

1.4 Feature Line Regression

As mentioned in the main paper, we proposed a novel feature line loss $\mathcal{L}_{line}$ as well as the edge length regularization loss $\mathcal{L}_{edge}$ to guide the feature line synthesis while reducing the zigzag artifacts:

$$\mathcal{L}_{fitting} = \mathcal{L}_{line} + \lambda_{edge}\ \mathcal{L}_{edge}.$$

The feature line loss $\mathcal{L}_{line}$ is calculated using Chamfer distance, which measures the average squared distance between the vertices on the predicted feature lines and the corresponding feature lines annotated on the ground-truth point clouds. The edge length regularization term is calculated as the average squared length of the edges on the selected feature lines. It is worth mentioning that the feature lines on the adaptable template are a series of closed polygonal curves while the ”feature lines” of the ground truth point clouds are a subset of the points that lie around the intended landmark regions (see Fig. 2 right). During training, λ_{edge} is set to 0.2 according to our experiment to strike a balance between reconstruction accuracy and smoothness of the generated feature lines.

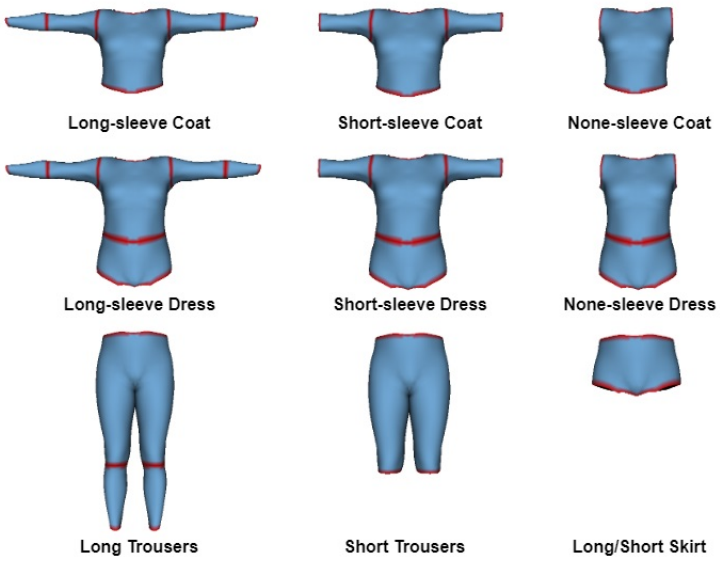

Fig. 1: Activated semantic regions and feature lines of adaptable template for each type of clothes. The feature lines are colored in red. Long and short skirts share the same activated patterns.

1.5 Adaptable Template

In this section, we provide more details regarding the generation of the adaptable template. Fig. 1 shows the activated semantic regions of our adaptable template when dealing with different types of garments. The curves colored in red are the activated feature lines for each scenario.

2 Evaluation of Adaptable Template

In this section, we further evaluate the effectiveness of the proposed adaptable template by comparing it with an alternative method that uses type-specific template. In particular, the comparing method leverages different models for reconstructing different types of clothing – each model is trained using one type-specific cloth template and its corresponding subset of the 3D data and training images. In contrast, as our adaptable template can handle all available clothing categories in Deep Fashion3D, our approach is trained on the entire training set. Note that, to ensure fair comparison, the training of these two approaches share the identical settings, network structure, and losses except the above-mentioned differences in the training data.

We compare the performance of two approaches in terms of feature line prediction on novel images as shown in Figure 2. Our approach can generate much more plausible feature lines compared to that of the approach which relies on category-specific templates. In addition, our predicted features lines are very close to the ground truth despite the different data modalities. The stronger generalization performance indicates the advantages of our proposed adaptable template which is able to harness all kinds of training data at a single network to prevent overfitting.

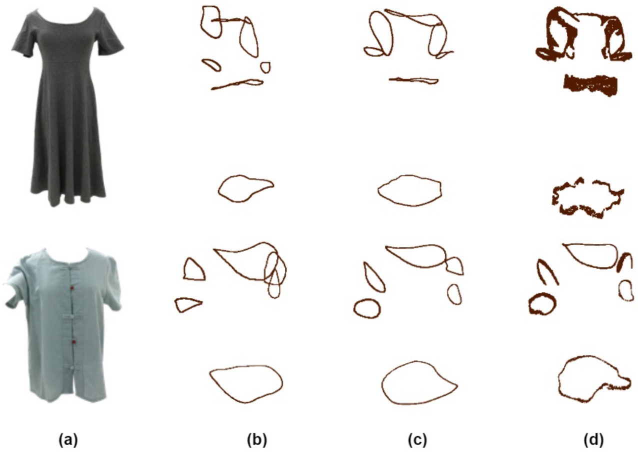

Fig. 2: Comparisons with category-specific models in terms of feature line prediction. Given the input images (a) on the left, we show (b) feature lines generated by method using type-specific template and trained with type-specific data; (c) feature lines generated using our approach with adaptable template and trained with full training data covering all clothing categories; (d) ground truth. Note that the ground-truth feature lines are labeled on reconstructed point clouds which are also point clouds as shown here.

3 More Results on In-the-wild Images

In this section, we show more results generated by our baseline approach trained on our Deep Fashion3D dataset in Figure 3. Note that all the images adopted for testing are in-the-wild images from the Internet, which are not seen during training. It is also worth mentioning that we mannually mask out the non-garment reginons when testing with the images in the wild and will leave cloth

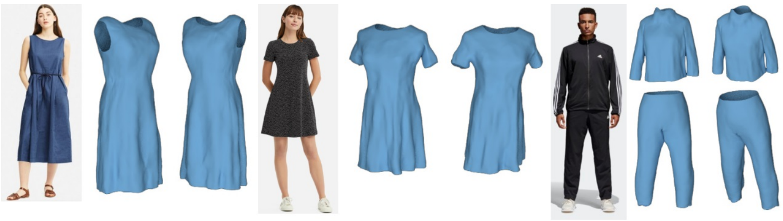

Fig. 3: Reconstruction from in-the-wild images using our approach. For each result, we visualize the obtained reconstructed meshes in two different views.

segementataion for the future work. As seen from the results, our approach can well capture the clothing topologies given a variety of input styles while faithfully recovering the geometric details.

References

1. Ge, Y., Zhang, R., Wang, X., Tang, X., Luo, P.: Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5337–5345 (2019)
2. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
3. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics **34**(6), 248:1–248:16 (2015)
4. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4460–4470 (2019)