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

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

Abstract. High-fidelity clothing reconstruction is the key to achieving photorealism in a wide range of applications including human digitization, virtual try-on, etc. Recent advances in learning-based approaches have accomplished unprecedented accuracy in recovering unclothed human shape and pose from single images, thanks to the availability of powerful statistical models, e.g. SMPL, learned from a large number of body scans. In contrast, modeling and recovering clothed human and 3D garments remains notoriously difficult, mostly due to the lack of large-scale clothing models available for the research community. We propose to fill this gap by introducing Deep Fashion3D, the largest collection to date of 3D garment models, with the goal of establishing a novel benchmark and dataset for the evaluation of image-based garment reconstruction systems. Deep Fashion3D contains 2078 models reconstructed from real garments, which covers 10 different categories and 563 garment instances. It provides rich annotations including 3D feature lines, 3D body pose and the corresponded multi-view real images. In addition, each garment is randomly posed to enhance the variety of real clothing deformations. To demonstrate the advantage of Deep Fashion3D, we propose a novel baseline approach for single-view garment reconstruction, which leverages the merits of both mesh and implicit representations. A novel adaptable template is proposed to enable the learning of all types of clothing in a single network. Extensive experiments have been conducted on the proposed dataset to verify its significance and usefulness.

1 Introduction

Human digitization is essential to a variety of applications ranging from visual effects, video gaming, to telepresence in VR/AR. The advent of deep learning techniques has achieved impressive progress in recovering unclothed human

†The first three authors should be considered as joint first authors.
* Xiaoguang Han is the corresponding author. Email:hanxiaoguang@cuhk.edu.cn.

shape and pose simply from multiple [30, 63] or even single [45, 57, 5] images. However, these leaps in performance come only when a large amount of labeled training data is available. Such limitation has led to inferior performance of reconstructing clothing – the key element of casting a photorealistic digital human, compared to that of naked human body reconstruction. One primary reason is the scarcity of 3D garment datasets in contrast with large collections of naked body scans, e.g. SMPL [39], SCAPE [6], etc. In addition, the complex surface deformation and large diversity of clothing topologies have introduced additional challenges in modeling realistic 3D garments.

Fig. 1: We present Deep Fashion3D, a large-scale repository of 3D clothing models reconstructed from real garments. It contains over 2000 3D garment models, spanning 10 different cloth categories. Each model is richly labeld with ground-truth point cloud, multi-view real images, 3D body pose and a novel annotation named feature lines. With Deep Fashion3D, inferring the garment geometry from a single image becomes possible.

To address the above issues, there is an increasing need of constructing a high-quality 3D garment database that satisfies the following properties. First of all, it should contain a large-scale repository of 3D garment models that cover a wide range of clothing styles and topologies. Second, it is preferable to have models reconstructed from the real images with physically-correct clothing wrinkles to accommodate the requirement of modeling complicated dynamics and deformations caused by the body motions. Lastly, the dataset should be labeled with sufficient annotations to provide strong supervision for deep generative models.

Multi-Garment Net (MGN) [7] introduces the first dataset specialized for digital clothing obtained from real scans. The proposed digital wardrobe contains 356 digital scans of clothed people which are fitted to pre-defined parametric cloth templates. However, the digital wardrobe only captures 5 garment categories, which is quite limited compared to the large variety of garment styles. Apart from 3D scans, some recent works [61, 26] propose to leverage synthetic data obtained from physical simulation. However, the synthetic models lack real-

ism compared to the 3D scans and cannot provide the corresponding real images, which are critical to generalizing the trained model to images in the wild.

In this paper, we address the lack of data by introducing Deep Fashion3D, the largest 3D garment dataset by far, that contains thousands of 3D clothing models with comprehensive annotations. Compared to MGN, the collection of Deep Fashion3D is one order of magnitude larger – including 2078 3D models reconstructed from real garments. It is built from 563 diverse garment instances, covering 10 different clothing categories. Annotation-wise, we introduce a new type of annotation tailored for 3D garment – 3D feature lines. The feature lines denote the most prominent geometrical features on garment surfaces (see Fig. 3), including necklines, cuff contours, hemlines, etc, which provide strong priors for 3D garment reconstruction. Apart from feature lines, our annotations also include calibrated multi-view real images and the corresponded 3D body pose. Furthermore, each garment item is randomly posed to enhance the dataset capacity of modeling dynamic wrinkles.

To fully exploit the power of Deep Fashion3D, we propose a novel baseline approach that is capable of inferring realistic 3D garments from a single image. Despite the large diversity of clothing styles, most of the existing works are limited to one fixed topology [19, 33]. MGN [7] introduces class-specific garment network – each deals with a particular topology and is trained by one-category subset of the database. However, given the very limited data, each branch is prone to having overfitting problems. We propose a novel representation, named adaptable template, that can scale to varying topologies during training. It enables our network to be trained using the entire dataset, leading to stronger expressiveness. Another challenge of reconstructing 3D garments is that clothing model is typically a shell structure with open boundaries. Such topology can hardly be handled by the implicit or voxel representation. Yet, the methods based on deep implicit functions [43, 48] have shown their ability of modeling fine-scale deformations that the mesh representation is not capable of. We propose to connect the good ends of both worlds by transferring the high-fidelity local details learnt from implicit reconstruction to the template mesh with correct topology and robust global deformations. In addition, since our adaptable template is built upon the SMPL topology, it is convenient to repose or animate the reconstructed results. The proposed framework is implemented in a multi-stage manner with a novel feature line loss to regularize mesh generation.

We have conducted extensive benchmarking and ablation analysis on the proposed dataset. Experimental results demonstrate that the proposed baseline model trained on Deep Fashion3D sets new state of the art on the task of single-view garment reconstruction. Our contributions can be summarized as follows:

- We build Deep Fashion3D, a large-scale, richly annotated 3D clothing dataset reconstructed from real garments. To the best of our knowledge, this is the largest dataset of its kind.
- We introduce a novel baseline approach that combines the merits of mesh and implicit representation and is able to faithfully reconstruct 3D garment from a single image.

- We propose a novel representation, called adaptable template, that enables encoding clothing of various topologies in a single mesh template.
- We first present the feature line annotation specialized for 3D garments, which can provide strong priors for garment reasoning related tasks, e.g., 3D garment reconstruction, classification, retrieval, etc.
- We build a benchmark for single-image garment reconstruction by conducting extensive experiments on evaluating a number of state-of-the-art single-view reconstruction approaches on Deep Fashion3D.

2 Related Work

3D Garment Datasets. While most of existing repositories focus on naked [6, 8, 39, 9] or clothed [68] human body, datasets specially tailored for 3D garment is very limited. BUFF dataset [67] consists of high-resolution 4D scans of clothed human with very limited ammount. In addition, it fails to provide separated models for body and clothing. Segmenting garment models from the 3D scans remains extremely laborious and often leads to corrupted surfaces due to occlusions. To address this issue, Pons-Moll et al. [49] propose an automatic solution to extract the garments and their motion from 4D scans. Recently, a few datasets specialized for 3D garment are proposed. Most of the works [25, 61] propose to synthetically generate garment dataset using physical simulation. However, the quality of the synthetic data is not on par with that of real data. In addition, it remains difficult to generalize the trained model to real images as only synthetic images are available. MGN [7] introduces the first garment dataset obtained from 3D scans. However, the dataset only covers 5 cloth categories and is limited to a few hundreds of samples. In contrast, Deep Fashion3D collects more than two thousand clothing models reconstructed from real garments, which covers a much larger diversity of garment styles and topologies. Further, the novel annotation of feature lines provides stronger and more accurate supervision for reconstruction algorithms, which is demonstrated in Section 5.

Performance capture. Over the past decades, progress [59, 44, 42] has been made to capture cloth surface deformation in motion. Vision-based approaches strive to leverage the easily accessible RGB data and develop frameworks either based on texture pattern tracking [62, 53], shading cues [69] or calibrated silhouettes obtained from multi-view videos [12, 55, 37, 11]. However, without dense correspondences or priors, the silhouette-based approaches cannot fully recover the fine details. To improve the reconstruction quality, stronger prior knowledge, including the clothing type [20], pre-scanned template model [27], stereo [10] and photometric [29, 58] constraints, has been considered in recent works. With the advances of fusion-based solutions [32, 46], template model can be eliminated as the surface geometry can be progressively fused on the fly [18, 21] with even a single depth camera [66, 65, 64]. Yet, most of the existing works estimate body and clothing jointly and thus cannot obtain a separated cloth surface from the output. Chen et al. [15] propose to model 3D garment from a single depth camera by fitting deformable templates to the initial mesh generated by KinectFusion.

Single-view garment reconstruction. Inferring 3D cloth from a single image is highly challenging due to the scarcity of the input and the enormous searching space. Statistical model has been introduced for such ill-posed problem to provide strong priors. However, most models [6, 39, 28, 50, 34] are restricted to capturing human body only. Attempts have been made to jointly reconstruct body and clothing from videos [3, 4] and multi-view images [30, 63]. Recent advances in deep learning based approaches [45, 57, 52, 5, 36, 2, 51, 14, 56] have achieved single-view clothed body reconstruction. However, for all these methods, tedious manual post-processing is required to extract the clothing surface. And yet, the reconstructed clothing lacks realism. DeepWrinkles [35] synthesizes faithful clothing wrinkles onto a coarse garment mesh following a given pose. Jin et al. [33] leverage similar idea with [31], which encodes detailed geometry deformations in the uv space. However, the method is limited to a fixed topology and cannot scale well to large deformations. Daněřek et al. [19] propose to use physics based simulations as supervision for training a garment shape estimation network. However, the quality of their results is limited to that of the synthetic data and thus cannot achieve high photo-realism. Closer to our work, Multi-Garment Net [7] learns per-category garment reconstruction using scanned data. Nonetheless, their method typically requires 8 frames as input while our approach only consumes a single image. Further, since MGN relies on pre-trained parametric models, it cannot deal with out-of-scope deformations, especially the clothing wrinkles that are dependent on body poses. In contrast, our approach is blendshape-free and is able to faithfully capture multi-scale shape deformations.

3 Dataset Construction

Despite the rapid evolution of 2D garment image datasets from DeepFashion [38] to DeepFashion2 [23] and FashionAI [70], large-scale collection of 3D clothing is very rare. The digital wardrobe released by MGN [7] only contains 356 scans and is limited to only 5 garment categories, which is not sufficient for training an expressive reconstruction model. To fill this gap, we build a more comprehensive dataset named Deep Fashion3D, which is one order larger than MGN, richly annotated and covers a much larger variations of garment styles. We provide more details on data collection and statistics in the following context.

Type	Number	Type	Number
Long-sleeve coat	157	Long-sleeve dress	18
Short-sleeve coat	98	Short-sleeve dress	34
None-sleeve coat	35	None-sleeve dress	32
Long trousers	29	Long skirt	104
Short trousers	44	Short skirt	48

Table 1: Statistics of the each clothing categories of Deep Fashion3D.

Fig. 2: Example garment models of Deep Fashion3D.

Cloth Capture. To model the large variety of real-world clothing, we collect a large number of garments, consisting of 563 diverse items that covers 10 clothing categories. The detailed numbers for each category are shown in Table 1. We adopt the image-based reconstruction software [1] to reconstruct high-resolution garment models from multi-view images in the form of dense point cloud. In particular, the input images are captured in a multi-view studio with of 50 RGB cameras and controlled lighting. To enhance the expressiveness of the dataset, each garment item is randomly posed on a dummy model or real human to generate a large variety of real deformations caused by body motion. The body parts are manually removed from reconstructed point clouds. With the augmentation of poses, 2078 3D garment models in total are reconstructed from our pipeline.

Annotations. To facilitate future research on 3D garment reasoning, apart from the calibrated multi-view images, we provide additional annotations for Deep Fashion3D. In particular, we introduce *feature line* annotation which is specially tailored for 3D garments. Akin to facial landmarks, the feature lines denote the most prominent features, e.g. the open boundaries, the neckline, cuff, waist, etc, that could provide strong priors for faithful garment reconstruction. The details of feature line annotations are provided in Table 2 and visualized in Figure 3. We will show in method section that feature line labels can supervise the learning of 3D key lines prediction, which provide explicit constraints for mesh generation.

Furthermore, each reconstructed model is labeled with 3D pose represented by SMPL [39] coefficients. The pose is obtained by fitting the SMPL model to the reconstructed dense point cloud. Due to the highly coupled nature between human body and clothing, we believe the labeled 3D pose could be beneficial to infer the global shape and pose-dependent deformations of the garment model.

Data Statistics. To the best of our knowledge, among existing works, there are only three publicly available datasets specialized for 3D garments: Wang et.

Fig. 3: Visualization of feature line annotations. Different feature lines are highlighted using different colors.

Cloth Category	Feature line Positions
long-sleeve coat	ne, wa, sh, el, wr
short-sleeve coat	ne, wa, sh, el
none-sleeve coat	ne, wa, sh
long-sleeve dress	ne, wa, sh, el, wr, he
short-sleeve dress	ne, wa, sh, el, he
none-sleeve dress	ne, wa, sh, he
long/short trousers	wa, kn, an/ wa, kn
long/short skirt	wa, he/ wa, he

Table 2: Feature line positions for each cloth category. The meanings for the abbreviations are: 'ne'-neck, 'wa'-waist, 'sh'-shoulder, 'el'-elbow, 'wr'-wrist, 'kn'-knee, 'an'-ankle, 'he'-'hemline'.

	Wang et al. [61]	GarNet [26]	MGN [7]	**Deep Fashion3D**
# Models	2000	600	712	2078
# Categories	3	3	5	10
Real/Synthetic	synthetic	synthetic	real	real
Method	simulation	simulation	scanning	multi-view stereo
Annotations	input 2D sketch 3D body pose		vertex color 3D body pose	multi-view real images 3D feature lines 3D body pose

Table 3: Comparisons with other 3D garment datasets.

al [61], GarNet [26] and MGN [7]. In Table 3, we provide detailed comparisons with these datasets in terms of the number of models, categories, data modality, production method and data annotations. Scale-wise, Deep Fashion3D and Wang et al. [61] are one order larger than the other counterparts. However, our dataset covers much more garment categories compared to Wang et al. [61]. Apart from our dataset, only MGN collects models reconstructed from real garments while the other two are fully synthetic. Regarding data annotations, Deep Fashion3D provides the richest data labels. In particular, multi-view real images are only available in our dataset. In addition, we present a new form of garment annotation, the 3D feature lines, which could offer important landmark information for a variety of 3D garment reasoning tasks including garment reconstruction, segmentation, retrieval, etc.

4 A Baseline Approach for Single-view Reconstruction

To demonstrate the usefulness of Deep Fashion3D, we propose a novel baseline approach for single-view garment reconstruction. Specifically, taking a single image I of a garment as input, we aim to reconstruct its 3D shape represented as a triangular mesh. Although recent advances in 3D deep learning techniques have achieved promising progress in single-view reconstruction on general objects, we

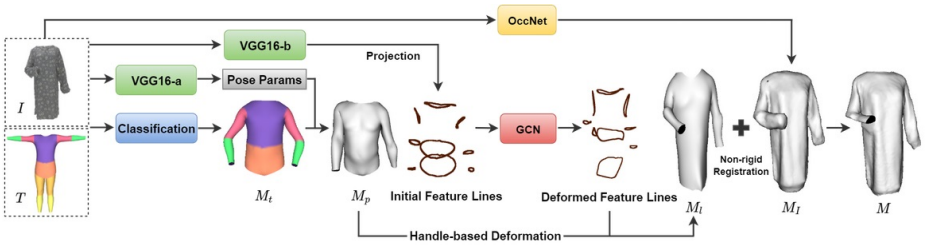

Fig. 4: The pipeline of our proposed approach.

found all existing approaches have difficulty scaling to cloth reconstruction. The main reasons are threefolds: (1) *Non-closed surfaces.* Unlike the general objects in ShapeNet [13], the garment shape typically appears as a thin layer with open boundary. While implicit representation [43, 48] can only model closed surface, voxel based approach [16] is not suited for recovering shell-like structure like the garment surface. (2) *Complex shape topologies.* As all existing mesh-based approaches [24, 60, 47] rely on deforming a fixed template, they fail to handle the highly diversified topologies introduced by different clothing categories. (3) *Complicated geometric details.* While general man-made objects typically consist of smooth surfaces, the clothing dynamics often introduces intricate high-frequency surface deformations that are challenging to capture.

Overview. To address the above issues, we propose to employ a hybrid representation that leverages the merits of each embedding. In particular, we harness both the capability of implicit surface of modeling fine geometric details and the flexibility of mesh representation of handling open surfaces. Our method starts with generating a template mesh M_t which can automatically adapt its topology to fit the target clothing category in the input image. It is then deformed to M_p according to estimated 3D pose. By treating the feature lines as a graph, we then apply image-guided graph convolutional network (GCN) to capture the 3D feature lines, which later trigger handle-based deformation and generates mesh M_l. To exploit the power of implicit representation, we first employ OccNet [43] to generate a mesh model M_I and then adaptively register M_l to M_I by incorporating the learned fine surface details from M_I while discarding its outliers and noises caused by enforcement of close surface. The proposed pipeline is illustrated in Figure 4.

4.1 Template Mesh Generation

Adaptable template. We propose *adaptable template*, a new representation that is scalable to different cloth topologies, enabling the generation of all types of cloth available in the dataset using a single network. The adaptable template is built on the SMPL [39] model by removing the head, hands and feet regions. As seen in Figure 4, it is then segmented into 6 semantic regions: torso, waist, and

upper/lower limbs/legs. During training, the entire adaptable template is fed into the pipeline. However, different semantic regions are activated according to the estimated cloth topology. We denote the template mesh as $M_t = (V, E, B)$, where $V = \{v_i\}$ and E are the set of vertices and edges respectively, and $B = \{b_i\}$ is a per-vertex binary activation mask. v_i will only be activated if $b_i = 1$; otherwise v_i will be detached during the training and removed in the output. The activation mask is determined by the estimated cloth category, where regions of vertices are labeled as a whole. For instance, to model a short-sleeve dress, vertices belonging to the regions of lower limbs and legs are deactivated. Note that in order to adapt the waist region to large deformations for modeling long dresses, we densify its triangulation accordingly using mesh subdivisions.

Cloth classification. We build a cloth classification network based on a pre-trained VGGNet. The classification network is trained using both real and synthetic images. The synthetic images are used in order to provide augmented lighting conditions to the training images. In particular, we render each garment model under different global illuminations in 5 random views. We generate around 10,000 synthetic images, 90% of which is used for training while the rest is reserved for testing. Our classification network can achieve an accuracy of 99.3%, leading to an appropriate template at both train and test time.

4.2 Learning Surface Reconstruction

To achieve a balanced trade-off between mesh smoothness and accuracy of reconstruction, we propose a multi-stage pipeline to progressively deforming M_t to fit the target shape.

Feature line-guided Mesh Generation. It is well understood that, the feature lines, such as necklines, hemlines, etc, play a key role in casting the shape contours of the 3D clothing. Therefore, we propose to first infer the 3D feature lines and then deform M_t by treating the feature lines as deformation handles.

Pose Estimation. Due to the large degrees of freedom of 3D lines, directly regressing their positions is highly challenging. To reduce the searching space, we first estimate the body pose and deform M_t to M_p which provides an initialization $\{l_i^p\}$ of 3D feature lines. Here, the pose of 3D garment is represented with SMPL pose parameters θ [39], which are regressed by a pose estimation network.

GCN-based Feature line regression. We represent the feature lines $\{l_i^p\}$ as polygons during pose estimation. This enables us to treat it as a graph and further employ an image-guided GCN to regress the vertex-wise displacements. We employ another VGG module to extract image features and leverage a similar learning strategy with Pixel2Mesh [60] to infer deformation of feature lines. Note that all of the feature lines predefined on the template are fed into the network, but only the activated subset of the feature lines are adopted to update network parameters.

Handle-based deformation. We denote the output feature lines of the above steps as $\{l_i^o\}$. M_l is obtained by deforming M_p so that its feature lines $\{l_i^p\}$ fit our prediction $\{l_i^o\}$. We use the handle-based Laplcacian deformation [54] by setting the alignment between $\{l_i^p\}$ and $\{l_i^o\}$ as hard constrains while optimizing the displacements of the remaining vertices to achieve smooth and visually pleasing deformations. Note that the explicit handle-based deformation can quickly lead to a result that is close to the target surface, which alleviates the difficulty of regressing of a large number of vertices.

Surface Refinement by Fitting Implicit Reconstruction. After obtaining M_l, a straightforward way to obtain surface details is to apply Pixel2Mesh [60] by taking M_l as input. However, as illustrated in Fig. 5, this method fails probably due to the inherent difficulty of learning the high-frequency details while preserving surface smoothness. In contrast, our empirical results indicate that the implicit surface based methods, such as OccNet [43], can faithfully recover the details but only generate closed surface. We therefore perform an adaptive non-rigid registration from M_l to OccNet output for transferring surface details.

Learning implicit surface. We directly employ OccNet [43] for learning the implicit surface. Specifically, the input image is first encoded into a latent vector using ResNet-18. For each 3D point in the space, a MLP layer consumes its coordinate and the latent code to predict if the point is inside or outside the surface. Note that we convert all the data into closed meshes using Poisson reconstruction in MeshLab [17]. With the trained network, we first generate an implicit field and then extract the reconstructed surface M_I using marching cube algorithm [40].

Detail transfer with adaptive registration. Though OccNet can synthesize high-quality geometric details, it may also introduce outliers due to its enforcement of generating closed surface. To improve robustness and convergence in conventional non-rigid ICP, we impose normal and distance constraints to filter out wrong correspondences so that only the correct high-frequency details are transferred: (1) the two points of a valid correspondence should have consistent normal direction (i.e., the angle of the two normal directions should be smaller than a threshold which is set as 60°). (2) the bi-directional Chamfer distance between the corresponded points should be less than a preset threshold σ (σ is set as 0.01). The adaptive registrations helps to remove erroneous correspondences and produces our final output M_r.

4.3 Training

There are four sub-networks need to be trained: cloth classification, pose estimation, GCN-based feature line fitting and the implicit reconstruction. Each of the sub-networks is trained independently. In the following subsections, we will provide the details on training data preparation and loss functions.

Training Data Generation

Pose estimation. We obtain the 3D pose of the garment model by fitting the SMPL model to the reconstructed dense point cloud. The data processing procedures are as follows: 1) for each annotated feature line, we calculate its center point as the its corresponding skeleton joint; 2) we use the joints in the torso region to align all the point clouds to ensure a consistent orientation and scale. 3) lastly, we compute the SMPL pose parameters for each model by fitting the joints and point cloud. The obtained pose parameters will be used for supervising the pose estimation module in Section 4.2.

Image rendering. We augment the input with synthetic images. In particular, for each model, we generate rendered images by randomly sampling 3 viewpoints and 3 different lighting environments, obtaining 9 images in total. Note that we only sample viewpoints from the front viewing angles as we only focus on front-view reconstruction in this work. However, our approach can scale to side or back view prediction by providing corresponding training images.

Loss functions The training of cloth classification, pose estimation and implicit reconstruction exactly follows the mainstream protocols. Hence, due to the page limit, we only focus on the part of feature line regression here while leaving other details in the appendix.

Feature line regression. Our training goal is to minimize the average distance between the vertices on the obtained feature lines and the ground-truth annotations. Therefore, our loss function is a weighted sum of a distance metric (we use Chamfer distance here) and an edge length regularization loss [60], which helps to smooth the deformed feature lines (more details can be found in supplementals).

5 Experimental Results

Implementation details. The whole pipeline proposed is implemented using PyTorch. The initialized learning rate is set to 5e-5 and with the batch size of 8. It takes about 30 hours to train the whole network using Adam optimization for 50 epochs using a NVIDIA TITAN XP graphics card.

5.1 Benchmarking on Single-view Reconstruction

Methods. We compare our method against six state-of-the-art single-view reconstruction approaches that use different 3D representations: 3D-R2N2 [16], PSG(Point Set Generation) [22], MVD (generating multi-view depth maps) [41], Pixel2Mesh [60], AtlasNet [24], MGN [7] and OccNet [43]. For AtlasNet, we have experimented it using both sphere template and patch template, which are denoted as "Atlas-Sphere" and "Atlas-Patch". To ensure fairness, we train all the algorithms, except MGN, on our dataset. In particular, training MGN requires

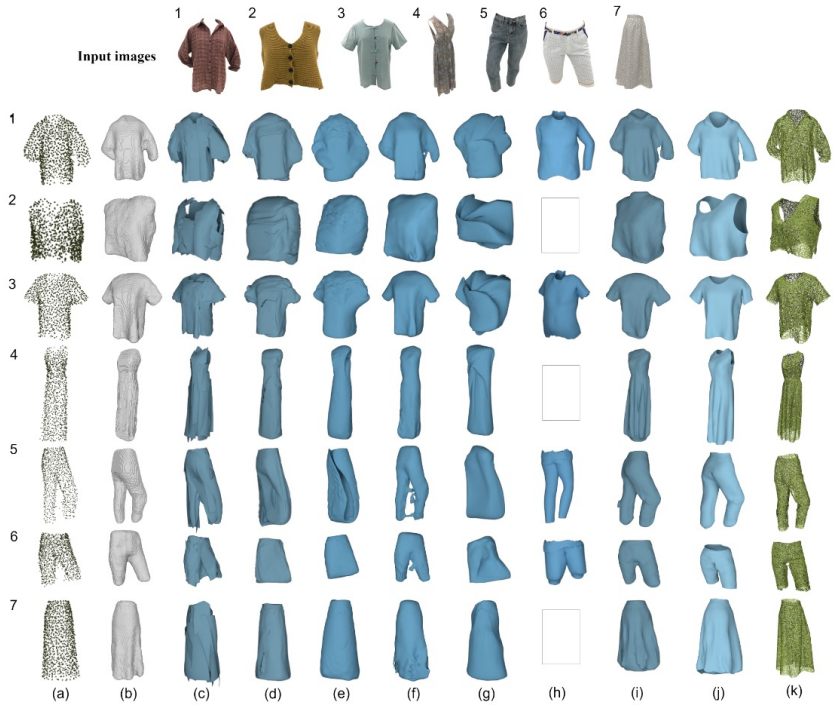

Fig. 5: Experiment results against other methods. Given an image, results are followed with (a) PSG (Point Set Generation) [22]; (b) 3D-R2N2 [16]; (c) AtlasNet [24] with 25 square patches; (d) AtlasNet [24] whith a sphere template; (e) Pixel2Mesh [60]; (f) MVD [41] (multi-view depth generation); (g) TMN [47] (topology modification network); (h) MGN (Multi-Garment Network) [7]; (i) OccNet [43]; (j) Ours; (k) The groundtruth point clouds. The input images on the top. The null means the method fails to generate a result.

ground-truth parameters for their category-specific cloth template, which is not applicable in our dataset. It is worth mentioning that, the most recent algorithm MGN can only handle 5 cloth categories and fails to produce reasonable results for out-of-scope classes, e.g., dress, as demonstrated in Fig. 5. To obtain the results of MGN, we manually prepared input data to fulfill the requirements of its released model, that is trained on digital wardrobe [7].

Quantitative results. Since the approaches leverage different 3D representations, we convert the outputs into point cloud for fair comparison. We then compute the Chamfer distance (CD) and Earth Mover's distance (EMD) between the outputs and the ground-truth for quantitative measurements. Table 4 shows the performance of different methods on our testing dataset. Our approach achieves the highest reconstruction accuracy compared to the other approaches.

Method	CD($\times 10^{-3}$)	EMD ($\times 10^{2}$)
3D-R2N2 (128^3) [16]	1.264	3.609
MVD [41]	1.047	4.058
PSG [22]	1.065	4.675
Pixel2Mesh [60]	0.782	9.078
AtlasNet(sphere) [24]	0.855	6.193
AtlasNet(patch) [24]	0.908	9.428
TMN [47]	0.865	8.580
OccNet (256^3) [43]	0.960	3.431
Ours	**0.679**	**2.942**

Table 4: The prediction errors of different methods evaluated on our testing data.

Qualitative results. In Figure 5, we also provide qualitative comparisons by randomly selecting some samples from different garment categories in arbitrary poses. Compared to the other methods, our approach provides more accurate reconstructions that are closer to ground truths. The reasons are: 1) 3D representations like point set [22], voxel [16] or multi-view depth maps [41] are not suitable for generating a clean mesh. 2) Although template-based methods [24, 60, 47] are designed for mesh generation, it is hard to use a fixed template for fitting diverse shape complexity of clothing. 3) As shown in the results, method based on implicit function [43] is able to synthesis rich details. However, it can only generate closed shapes, making it difficult to handle garment reconstruction, which typically consists of multiple open boundaries. By explicitly combining the merits of template-based methods and implicit ones, the proposed approach can not only capture the global shape but also generate faithful geometric details.

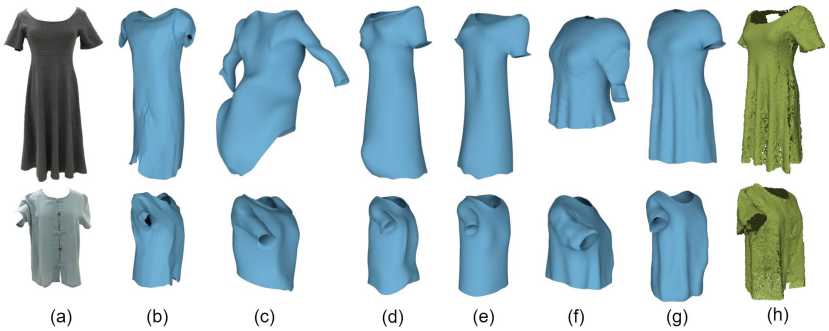

Fig. 6: Results of ablation studies. (a) input images; (b) results of M_t+GCN; (c) results of M_p+GCN; (d) results of M_l+GCN. (e) results of our approach without surface refinement, i.e., M_l. (f) M_t+Regis. (g) results of our full approach. (h) groundtruth point clouds.

5.2 Ablation Analysis

We further validate the effectiveness of each algorithmic component by selectively applying them in different settings: 1) Directly applying GCN on the generated template mesh M_t to fit the target shape, termed as M_t+GCN; 2) Applying GCN on M_p (obtained by deforming M_t with estimated SMPL pose) to fit the target shape, termed as M_p+GCN; 3) Applying GCN on the resulted mesh after feature line-guided deformation, i.e. M_l. This is termed as M_l+GCN; 4) Directly performing registration from M_t to M_I for details transferring, which is termed as M_t+Regis. Figure 6 shows the qualitative comparisons between these settings and the proposed one. As seen, the baseline approach produce the best results.

As observed from the experiments, it is difficult for GCN to learn geometric details. There are two possible reasons: 1) It is inherently difficult to synthesize high-frequency signals while preserving surface smoothness; 2) GCN structure might be not suitable for a fine-grained geometric learning task as graph is a sparse and crude approximation of a surface. We also found that the feature lines are much easier to learn and explicit handle-based deformation works surprisingly well. The deeper study in this regard is left as one of our further works.

6 Conclusions and Discussions

We have proposed a new dataset called Deep Fashion3D for image-based garment reconstruction, which is by far the largest 3D garment collection reconstructed from real clothing images. In particular, it consists of over 2000 highly diversified garment models covering 10 clothing categories and 563 distinct garment items. In addition, each model of Deep Fashion3D is richly labeled with 3D body pose, 3D feature lines and multi-view real images. We also presented a baseline approach for single-view reconstruction to validate the usefulness of the proposed dataset. It uses a novel representation, called adaptable template, to learn a variety of clothing types in a single network. We have performed extensive benchmarking on our dataset using a variety of recent methods. We found that single-view garment reconstruction is an extremely challenging problem with ample opportunity for improved methods. We hope Deep Fashion3D and our baseline approach will bring some insight to inspire future research in this field.

Currently, our pipeline does not support end-to-end training and requires some offline processing steps. We believe it would be an interesting future avenue to investigate an end-to-end pipeline to enable more accurate reconstruction.

Acknowledgment

The work was supported in part by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the National Key R&D Program of China with grant No. 2018YFB1800800, by Natural Science Foundation of China with grant NSFC-61629101 and 61902334, by Guangdong Research Project No. 2017ZT07X152, and by Shenzhen Key Lab Fund No.ZDSYS201707251409055. The authors would thank Yuan Yu for her early efforts on dataset construction.

References

1. Agisoft: Mentashape. https://www.agisoft.com/ (2019)
2. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single RGB camera. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (jun 2019)
3. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Detailed human avatars from monocular video. In: International Conference on 3D Vision (3DV) (sep 2018)
4. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
5. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2019)
6. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: SCAPE: shape completion and animation of people. ACM Transactions on Graphics **24**(3), 408–416 (2005)
7. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3d people from images. In: IEEE International Conference on Computer Vision (ICCV). IEEE (oct 2019)
8. Bogo, F., Romero, J., Loper, M., Black, M.J.: FAUST: Dataset and evaluation for 3D mesh registration. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE, Piscataway, NJ, USA (Jun 2014)
9. Bogo, F., Romero, J., Pons-Moll, G., Black, M.J.: Dynamic FAUST: Registering human bodies in motion. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017. IEEE, Piscataway, NJ, USA (Jul 2017)
10. Bradley, D., Popa, T., Sheffer, A., Heidrich, W., Boubekeur, T.: Markerless garment capture. In: ACM Transactions on Graphics (TOG). vol. 27, p. 99. ACM (2008)
11. Cagniart, C., Boyer, E., Ilic, S.: Probabilistic deformable surface tracking from multiple videos. In: European conference on computer vision. pp. 326–339. Springer (2010)
12. Carranza, J., Theobalt, C., Magnor, M.A., Seidel, H.P.: Free-viewpoint video of human actors, vol. 22. ACM (2003)
13. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
14. Chen, X., Guo, Y., Zhou, B., Zhao, Q.: Deformable model for estimating clothed and naked human shapes from a single image. The Visual Computer **29**(11), 1187–1196 (2013)
15. Chen, X., Zhou, B., Lu, F.X., Wang, L., Bi, L., Tan, P.: Garment modeling with a depth camera. ACM Trans. Graph. **34**(6), 203–1 (2015)
16. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016)
17. Cignoni, P., Callieri, M., Corsini, M., Dellepiane, M., Ganovelli, F., Ranzuglia, G.: Meshlab: an open-source mesh processing tool. In: Eurographics Italian chapter conference. vol. 2008, pp. 129–136. Salerno (2008)
18. Collet, A., Chuang, M., Sweeney, P., Gillett, D., Evseev, D., Calabrese, D., Hoppe, H., Kirk, A., Sullivan, S.: High-quality streamable free-viewpoint video. ACM Transactions on Graphics (ToG) **34**(4), 69 (2015)

19. Daněřek, R., Dibra, E., Öztireli, C., Ziegler, R., Gross, M.: Deepgarment: 3d garment shape estimation from a single image. In: Computer Graphics Forum. vol. 36, pp. 269–280. Wiley Online Library (2017)
20. De Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video, vol. 27. ACM (2008)
21. Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S.R., Kowdle, A., Escolano, S.O., Rhemann, C., Kim, D., Taylor, J., et al.: Fusion4d: Real-time performance capture of challenging scenes. ACM Transactions on Graphics (TOG) **35**(4), 114 (2016)
22. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
23. 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)
24. Groueix, T., Fisher, M., Kim, V.G., Russell, B., Aubry, M.: AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2018)
25. Gundogdu, E., Constantin, V., Seifoddini, A., Dang, M., Salzmann, M., Fua, P.: Garnet: A two-stream network for fast and accurate 3d cloth draping. arXiv preprint arXiv:1811.10983 (2018)
26. Gundogdu, E., Constantin, V., Seifoddini, A., Dang, M., Salzmann, M., Fua, P.: Garnet: A two-stream network for fast and accurate 3d cloth draping. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8739–8748 (2019)
27. Habermann, M., Xu, W., Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Livecap: Real-time human performance capture from monocular video. ACM Transactions on Graphics (TOG) **38**(2), 14 (2019)
28. Hasler, N., Stoll, C., Sunkel, M., Rosenhahn, B., Seidel, H.P.: A statistical model of human pose and body shape. In: Computer graphics forum. vol. 28, pp. 337–346. Wiley Online Library (2009)
29. Hernández, C., Vogiatzis, G., Brostow, G.J., Stenger, B., Cipolla, R.: Non-rigid photometric stereo with colored lights. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8. IEEE (2007)
30. Huang, Z., Li, T., Chen, W., Zhao, Y., Xing, J., LeGendre, C., Luo, L., Ma, C., Li, H.: Deep volumetric video from very sparse multi-view performance capture. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 336–354 (2018)
31. Huynh, L., Chen, W., Saito, S., Xing, J., Nagano, K., Jones, A., Debevec, P., Li, H.: Mesoscopic facial geometry inference using deep neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
32. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., et al.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th annual ACM symposium on User interface software and technology. pp. 559–568. ACM (2011)
33. Jin, N., Zhu, Y., Geng, Z., Fedkiw, R.: A pixel-based framework for data-driven clothing. arXiv preprint arXiv:1812.01677 (2018)

34. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8320–8329 (2018)
35. Lahner, Z., Cremers, D., Tung, T.: Deepwrinkles: Accurate and realistic clothing modeling. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 667–684 (2018)
36. Lazova, V., Insafutdinov, E., Pons-Moll, G.: 360-degree textures of people in clothing from a single image. In: International Conference on 3D Vision (3DV) (sep 2019)
37. Leroy, V., Franco, J.S., Boyer, E.: Multi-view dynamic shape refinement using local temporal integration. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3094–3103 (2017)
38. 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)
39. 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)
40. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. ACM siggraph computer graphics **21**(4), 163–169 (1987)
41. Lun, Z., Gadelha, M., Kalogerakis, E., Maji, S., Wang, R.: 3d shape reconstruction from sketches via multi-view convolutional networks. In: 2017 International Conference on 3D Vision (3DV). pp. 67–77. IEEE (2017)
42. Matsuyama, T., Nobuhara, S., Takai, T., Tung, T.: 3D video and its applications. Springer Science & Business Media (2012)
43. 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)
44. Miguel, E., Bradley, D., Thomaszewski, B., Bickel, B., Matusik, W., Otaduy, M.A., Marschner, S.: Data-driven estimation of cloth simulation models. In: Computer Graphics Forum. vol. 31, pp. 519–528. Wiley Online Library (2012)
45. Natsume, R., Saito, S., Huang, Z., Chen, W., Ma, C., Li, H., Morishima, S.: Siclope: Silhouette-based clothed people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4480–4490 (2019)
46. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 343–352 (2015)
47. Pan, J., Han, X., Chen, W., Tang, J., Jia, K.: Deep mesh reconstruction from single rgb images via topology modification networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9964–9973 (2019)
48. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 165–174 (2019)
49. Pons-Moll, G., Pujades, S., Hu, S., Black, M.: ClothCap: Seamless 4D clothing capture and retargeting. ACM Transactions on Graphics (SIGGRAPH) **36**(4) (2017)
50. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.J.: Dyna: A model of dynamic human shape in motion. ACM Transactions on Graphics (TOG) **34**(4), 120 (2015)

51. Pumarola, A., Sanchez, J., Choi, G., Sanfeliu, A., Moreno-Noguer, F.: 3DPeople: Modeling the Geometry of Dressed Humans. In: International Conference on Computer Vision (ICCV) (2019)
52. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. arXiv preprint arXiv:1905.05172 (2019)
53. Scholz, V., Stich, T., Keckeisen, M., Wacker, M., Magnor, M.: Garment motion capture using color-coded patterns. In: Computer Graphics Forum. vol. 24, pp. 439–447. Wiley Online Library (2005)
54. Sorkine, O., Cohen-Or, D., Lipman, Y., Alexa, M., Rössl, C., Seidel, H.P.: Laplacian surface editing. In: Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing. pp. 175–184. ACM (2004)
55. Starck, J., Hilton, A.: Surface capture for performance-based animation. IEEE computer graphics and applications **27**(3), 21–31 (2007)
56. Tang, S., Tan, F., Cheng, K., Li, Z., Zhu, S., Tan, P.: A neural network for detailed human depth estimation from a single image. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7750–7759 (2019)
57. Varol, G., Ceylan, D., Russell, B., Yang, J., Yumer, E., Laptev, I., Schmid, C.: Bodynet: Volumetric inference of 3d human body shapes. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 20–36 (2018)
58. Vlasic, D., Peers, P., Baran, I., Debevec, P., Popović, J., Rusinkiewicz, S., Matusik, W.: Dynamic shape capture using multi-view photometric stereo. In: ACM Transactions on Graphics (TOG). vol. 28, p. 174. ACM (2009)
59. Wang, H., O'Brien, J.F., Ramamoorthi, R.: Data-driven elastic models for cloth: modeling and measurement. In: ACM Transactions on Graphics (TOG). vol. 30, p. 71. ACM (2011)
60. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: ECCV (2018)
61. Wang, T.Y., Ceylan, D., Popovic, J., Mitra, N.J.: Learning a shared shape space for multimodal garment design. ACM Trans. Graph. **37**(6), 1:1–1:14 (2018). https://doi.org/10.1145/3272127.3275074
62. White, R., Crane, K., Forsyth, D.A.: Capturing and animating occluded cloth. In: ACM Transactions on Graphics (TOG). vol. 26, p. 34. ACM (2007)
63. Xu, Y., Yang, S., Sun, W., Tan, L., Li, K., Zhou, H.: 3d virtual garment modeling from rgb images. arXiv preprint arXiv:1908.00114 (2019)
64. Yu, T., Guo, K., Xu, F., Dong, Y., Su, Z., Zhao, J., Li, J., Dai, Q., Liu, Y.: Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 910–919 (2017)
65. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7287–7296 (2018)
66. Yu, T., Zheng, Z., Zhong, Y., Zhao, J., Dai, Q., Pons-Moll, G., Liu, Y.: Simulcap: Single-view human performance capture with cloth simulation. arXiv preprint arXiv:1903.06323 (2019)
67. Zhang, C., Pujades, S., Black, M.J., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3d scan sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4191–4200 (2017)

68. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
69. Zhou, B., Chen, X., Fu, Q., Guo, K., Tan, P.: Garment modeling from a single image. In: Computer graphics forum. vol. 32, pp. 85–91. Wiley Online Library (2013)
70. Zou, X., Kong, X., Wong, W., Wang, C., Liu, Y., Cao, Y.: Fashionai: A hierarchical dataset for fashion understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)