GarmentNets: Category-Level Pose Estimation for Garments via Canonical Space Shape Completion (Supplementary Material)

Cheng Chi Shuran Song
Columbia University
https://garmentnets.cs.columbia.edu

1. Network Architecture Details

Canonical Coordinate Prediction Network To predict per-point canonical coordinate, we used a PointNet++ [5] network in Multi-Resolution Grouping (MRG) configuration. The network consists of 3 Set Abstraction layers and 3 Feature Propagation layers. Detailed parameters are shown in Tab. 1. The final per-point 128 dimensional feature vector is transformed with a 3-layer MLP to perform 3×64 way classification with Cross Entropy Loss.

Layer	SA Radius	SA Ratio	SA Features	FP k	FP Features
1	0.05	0.5	128	3	128
2	0.1	0.25	256	3	128
3	Inf	1	1024	1	256

Table 1. **PointNet++ Parameters.** SA: parameters for Set Abstraction layers. FP: parameters for Feature Propagation layers.

Feature Completion Network (3D CNN) To transform the sparse feature volume scattered from per-point features to a dense feature volume, we used a symmetrical 3D UNet [2] architecture with 4 levels of encoder/decoder pairs. Each level of encoder/decoder has 32 feature maps.

Shape Completion Network To predict Winding Number Field (WNF), the interpolated features from dense feature volume is transformed using a 3 layer MLP with feature dimensions [512, 512, 1].

Warp Field Network Similar to the Shape Completion Network, the interpolated features are transformed using a 3 layer MLP with feature dimensions [512 ,512, 3].

2. Additional Results

Fig. 2 and 3 show additional results on real world and simulated data respectively. The real world point cloud are collecting using an iPhone 12 Pro Max.

Garment Category Classification Our algorithm described in the paper assumes known garment category for the input point cloud. When dealing with a mixed pile of garments, we assume that the category can be inferred using a classifier.

To validate this assumption, we trained a simple image classifier using only RGB images. The model uses an ImageNet [3] pre-trained ResNet-50 [4] backbone to extract a 2048 dimensional feature. The feature is then transformed using a 3-layer MLP to perform 6 way classification with Cross Entropy Loss.

The classifier is trained on each view independently. During prediction, we use the majority ensemble of all 4 views. This simple model yields 93.85% prediction accuracy on the test set. The confusion matrix is shown in Tab. 2.

	Dress	Jumpsuit	Skirt	Top	Trousers	Tshirt
Dress	0.966	0.020	0.003	0.001	0.005	0.005
Jumpsuit	0.003	0.956	0.000	0.000	0.010	0.008
Skirt	0.162	0.010	0.778	0.013	0.030	0.006
Top	0.001	0.004	0.002	0.979	0.009	0.004
Trousers	0.010	0.025	0.005	0.009	0.944	0.008
Tshirt	0.021	0.026	0.003	0.026	0.058	0.866

Table 2. **Confusion Matrix for Image Classification.**

Failure mode analysis Fig. 4 shows various failure cases on unseen simulation data. Due to low sharpness in the predicted winding number field, the canonical reconstruction for the Top and Shirt example have missing faces around the shoulder area. The Jumpsuit, Skirt and Pants example have over-smoothed warp field prediction, resulting in inaccurate task space mesh. The Dress example has missing shoulder strap due to winding number field's inability to represent wire-like structure.

Error distribution and correlation As shown in Fig. 1, the correspondence error is highly correlated to the canonical coordinate error. This suggests that jointly optimizing for both metrics might yield performance improvement.

Training Testing Split We use CLOTH3D dataset [1] for data generation. Tab. 3 shows the number of garment instances in training testing split for each category.

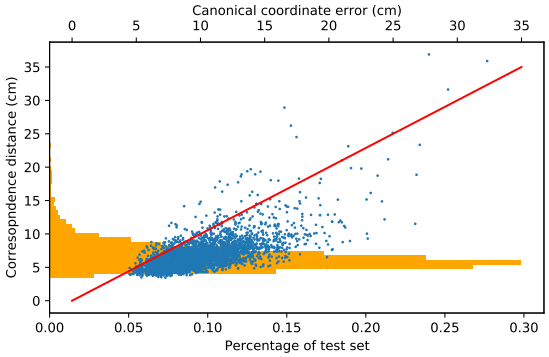

Figure 1. **Error distribution and correlation.** The correspondence distance vs canonical coordinate prediction error for Dress category is shown in blue dots. The line of slope 1 is shown in red. The correspondence distance error histogram is shown in orange.

	Dress	Jumpsuit	Skirt	Top	Trousers	Tshirt
training	1631	1825	376	840	1353	889
validation	203	227	46	104	169	111
testing	203	227	46	104	169	111

Table 3. **Training, Validation and Testing Split.** The number of garment instances used for each category. Each garment instance is simulated 21 times using randomly selected gripping point.

3. Limitations and Future Work

GarmentNets demonstrates promising result on real-world data while being trained only on synthetic data. However, the inability to propagate gradients from shape completion and warp field prediction modules to the canonical coordinate prediction module prevents us form training end-to-end. More specifically, the correspondence error is highly correlated to the canonical coordinate error, which suggests that jointly optimizing for both metrics might yield performance improvement. This limitation also requires us to manually define a dense correspondence from input to the canonical space, which is expensive to obtain on real-world data.

References

[1] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: Clothed 3d humans. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 344–359, Cham, 2020. Springer International Publishing. 1

[2] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432, Cham, 2016. Springer International Publishing. 1

[3] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[5] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5105–5114, Red Hook, NY, USA, 2017. Curran Associates Inc. 1

Figure 2. **Qualitative Results on Unseen Garment Instances (Real World).**

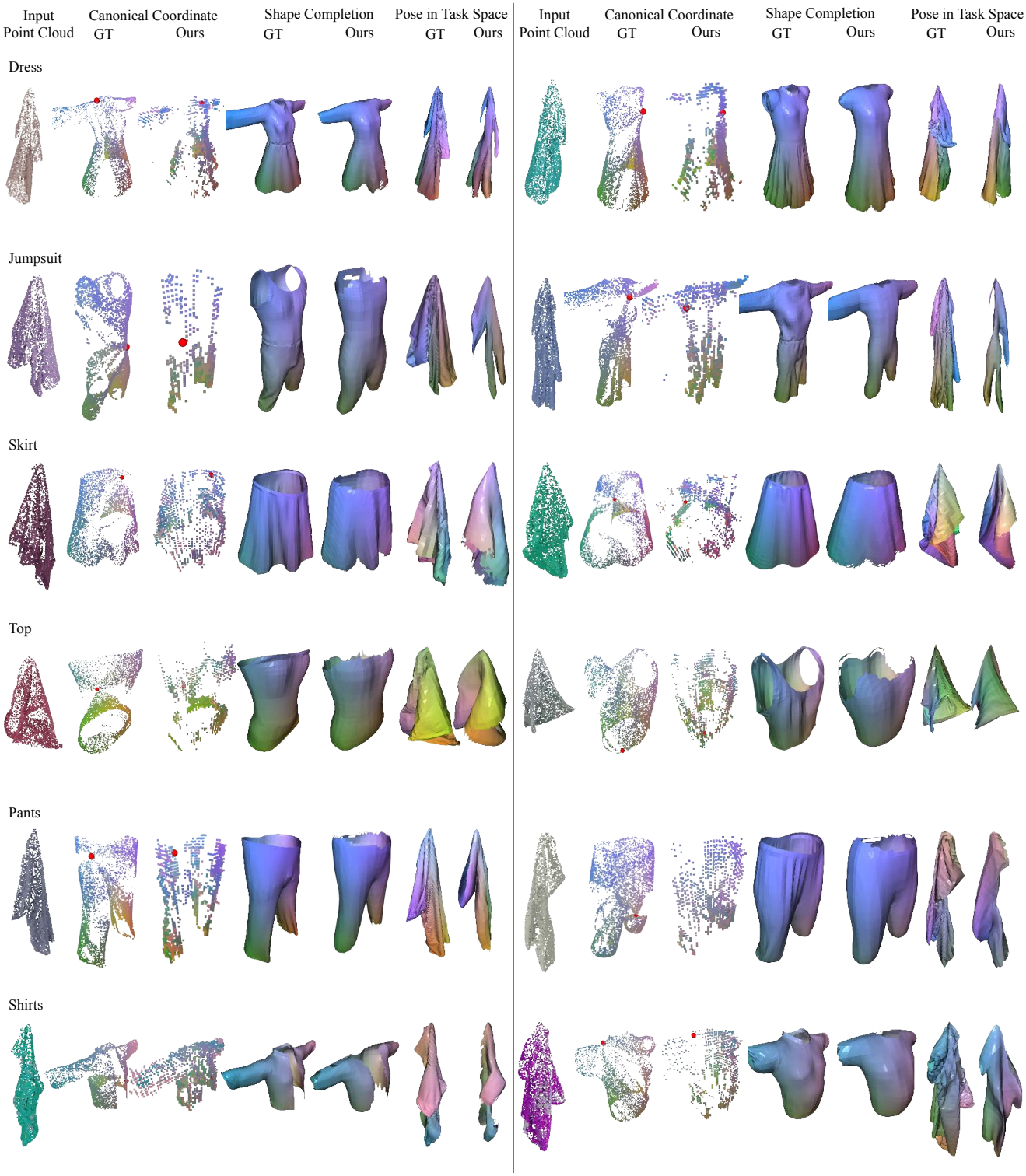

Figure 3. **Qualitative Results on Unseen Garment Instances (Simulation).**

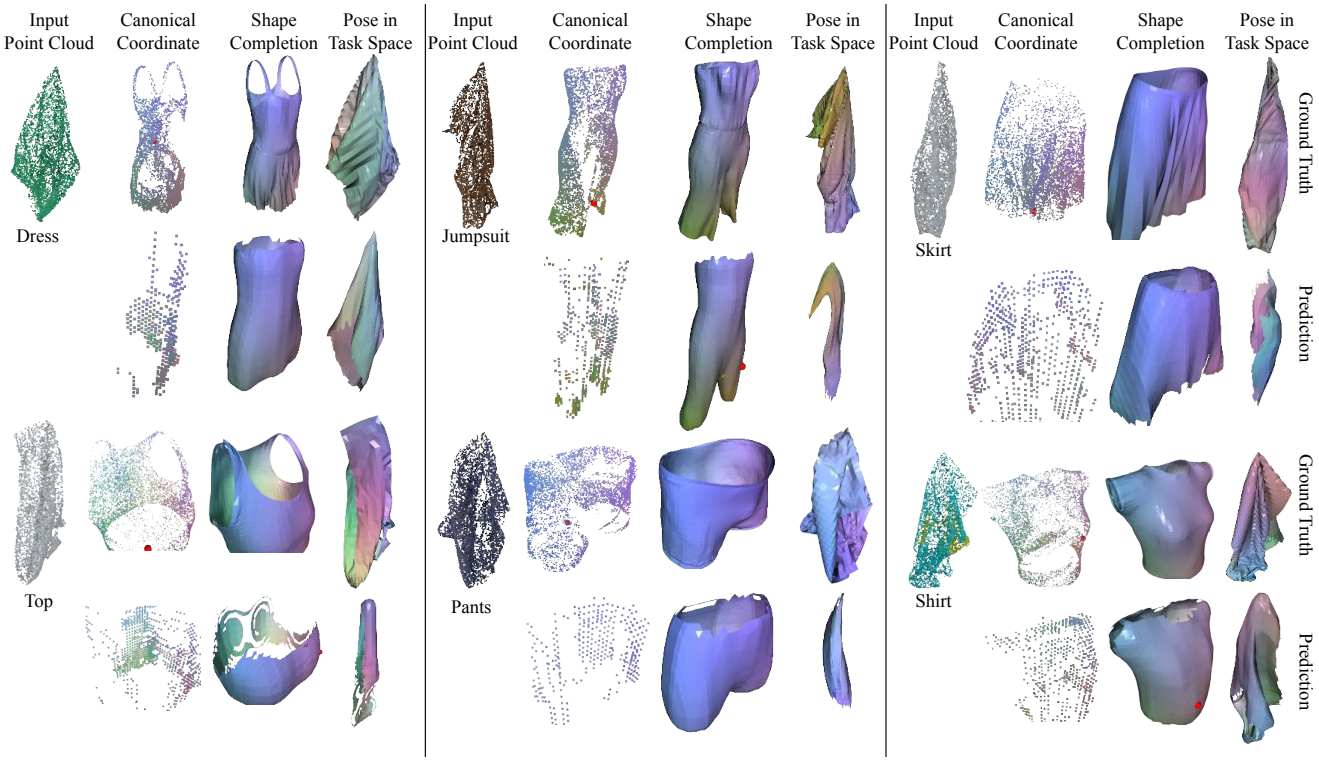

Figure 4. **Failure cases on Unseen Garment Instances (Simulation).**