SUPPLEMENTARY MATERIALS
CloSe: A 3D Clothing Segmentation Dataset and Model

Dimitrije Antić [1] Garvita Tiwari [2,3,4] Batuhan Ozcomlekci [2] Riccardo Marin [2,3] Gerard Pons-Moll [2,3,4]

[1]University of Amsterdam, Netherlands [2]University of Tübingen, Germany [3]Tübingen AI Center, Germany

[4]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

d.antic@uva.nl, gtiwari@mpi-inf.mpg.de, batuhan.oezcoemlekci@student.uni-tuebingen.de,

{riccardo.marin, gerard.pons-moll}@uni-tuebingen.de

1. Dataset

Our dataset CloSe-D comes from two sources, 1) CloSe-Di, which is dataset captured in our lab and 2) CloSe-Dc, dataset from commercial data sources. We explain the dataset capturing details in the following section, followed by the process for obtaining segmentation labels.

CloSe-Dc Data. We collect scans from different commercial dataset such as AXYZ [1], Twindom [5], Treedy [4], Renderpeople [3]. Due to licensing issues, we will not provide the scans from these datasets, but we will release the segmentation labels and detailed instructions to purchase these datasets from respective sources.

CloSe-Di Data Capture. Following data capture setup in [15, 19], we create a dataset of approximately 100 subjects in 7 diverse poses, wearing 12 garment classes. We use Treedy's scanner [4], which consists of $\sim$ 130 high-resolution camera at a fixed position. We use Metashape [2] for 3D reconstruction, which is photogrammetry-based reconstruction. Reconstructed scans are highly detailed and have high-resolution texture maps associated with them. We also register SMPL [11] to each scan, with the registration method used in [6, 10, 15].

Ground Truth Segmentation Labels of CloSe-Dc Scans. We follow the pipeline similar to the one in MGN-Seg [6]. We first register the scans to SMPL and SMPL+D [6]. We then render the registered meshes from 72 different views and apply SotA 2D Human Parsing method, PGN [8]. One of the major limitations of such a pipeline is inconsistent multiview prediction of the 2D Human Parsing method, as shown in Fig. 1. This is expected behavior from such methods as 1) they are not trained with any explicit loss to produce multi-view consistent results, and 2) they are not trained on multi-view images of the same scene. As a result, we observe many patches of undesired clothing classes in the 2D segmentation and hence in the lifted 3D segmentation as well. MGN-Seg [6] tried to solve this problem by using a pre-defined prior, but these priors are limited to 3 classes. We propose to clean such inconsistency using our hand-crafted heuristics and CloSe-T(see Fig. 5(left)). Moreover, PGN labels are inconsistent with our CloSe-Net labels, so we apply some merging and splitting in labels. We first explain heuristics for merging and segregation of labels in the following points:

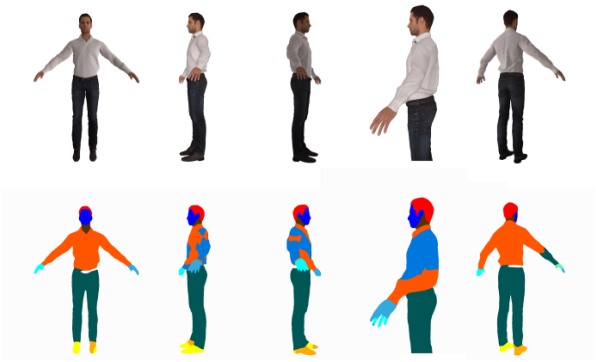

Figure 1. *Top*: Multiview rendered images of a scan.*Bottom*: Clothing segmentation obtained using 2D Parsing method [8]. 2D Parsing method generates inconsistent labels across views. Consequently, when these labels are elevated from 2D to 3D using the 2D-to-3D lifting technique, the resulting segmentation is noisy.

- *Merging body parts:* In PGN there are separate labels for left-leg, right-leg, left-arm and right-arm. We instead use a single label for all these parts, so we merge them into a single category.
- *Separate labels for Upper and Lower Garments :* PGN generates only two kinds of upper garment labels, namely 'Shirt' and 'Coat'. On the other hand, our model uses more fine-grained labels, *e.g.* 'Shirt' is further divided into 'TShirt', 'Vest', 'Hoodies' *etc*. We use the *change all* option provided in CloSe-T to correct such labels, as shown in Fig. 3. Similarly, there is only one label for lower garments: 'Pants', which we split into 'Pants' and 'Short-Pants'.

We show some examples of heuristics-based segmentation and manually refined segmentation in Fig. 2 and Fig. 3. We explain more details about our interactive tool in Sec. 3.

Ground Truth Segmentation Labels of CloSe-Di Scans. For CloSe-Di, we follow a similar idea, but instead of using SMPL+D registration and SMPL UV space, we use Metashape [2] to perform 2D-to-3D lifting of segmentation labels. The recovered 3D segmentation might be inaccurate because of 1) inaccurate 2D segmentation prediction, and 2) inconsistent 2D segmentation labels across different views. Similar to our processing of CloSe-Dc, we clean noise using heuristics. We define heuristics-based priors on SMPL mesh and clean the labels in for scan points directly. This alleviates the problem of obtaining SMPL+D [6] registrations. We deployed two different classes of heuristics:

- *Body Parts Heuristics:* We rely on the prior knowledge that some garments should not belong to unusual body parts (e.g., t-shirts on feet, trousers on arms *etc*.).
- *Garments Class Heuristics:* In some cases, we observed artifacts related to specific combinations of garments. In these cases, we deploy an additional set of rules to address these issues specifically.

2. Method

We explain the details of our model CloSe-Net in this section.

Point Encoder. We use the official implementation of DGCNN [17] and use 3 layers of EdgeConvoluation operation, followed by a single-layer MLP.

Clothing Encoder. We use a multi-head attention module in the encoder, where $\mathrm{n_{head}} = 4$ in our case. We also apply positional encoding to the query vector($\mathbf{p'}_{\mathrm{i}}^{2}$), before calculating the attention score.

Body Encoder. F^{b} requires the computation of nearest neighbors for each point within the batch, potentially leading to computational overhead during the training process. To mitigate this, we opt to precompute F^{b}. This is done by finding the nearest point for each scan point from the posed SMPL mesh ($M(\boldsymbol{\beta}, \boldsymbol{\beta})$). Subsequently, during inference, a preprocessing step is employed to calculate F^{b} beforehand, which is then used during inference.

3. Interactive Tool

In this section, we explain common functionalities provided by our tool and its usage in data annotation and network refinement.

Figure 2. Segmentation labels obtained using our heuristics might result in unclear boundaries(top, middle) and undesired noisy patches (bottom, middle). We clean such noise using CloSe-T and obtain high-quality labels, as shown on the right.

Interactive Tool Interface. We implement CloSe-T using Open3D [22] in C++ and introduce an easy-to-use, light-weight interactive 3D tool, which provides following functionalities:

- **I/O operations**: Loading/Saving meshes and labels,

Figure 3. Due inherent uncertainty in clothing classification, the segmentation labels acquired through [8] might be incoherent. However, such labeling discrepancies can be easily corrected using CloSe-T.

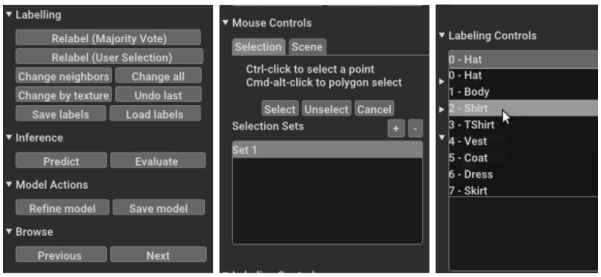

Figure 4. Functionalities provided in CloSe-T interface includes based I/O operations, mouse-controlled camera movement in the scene, region selection, relabelling, evaluation, and fine-tuning CloSe-Net.

Loading/evaluating pre-trained model, Saving/Evaluating refined network.

- **Scene**: Move in the scene with mouse control, change lights, background, *etc*.
- **User selection**: Easy polygon-based region selection by selecting the polygon edges by clicking.
- **Labeling**: Relabel region based on user selection/majority vote.

Label Correction. There are multiple options to label the selected regions

- *User selected class* : Manually set the class assigned to the selected areas. The predefined list of classes is shown in the dropdown menu; see Fig. 4(right).
- *Majority Vote*: If a partial/inaccurate initial segmentation of the scan already exists, the selected region can be labeled more efficiently using the "majority vote" procedure. More precisely, for example, if there is a patch of mislabeled points, the user can select the wider region around it, and label the whole region by the class that is the most commonly present in the region. This makes the labeling procedure much faster.

CloSe-T for Data Annotation .

We use CloSe-T to manually clean segmentation and generate high quality segmentation data, as shown in Fig. 2 and Fig. 3. We provide a demo of labeling process in the supplementary video and visualize key-stage of pipeline in Fig. 5.

Due to the inherently error-prone nature of the segmentation label generation pipeline, numerous scans displayed noisy boundaries and improperly labeled clothing classes. To address this issue, approximately 1000 scans were annotated using CloSe-T within the CloSe-D dataset, while the remaining were carefully verified. Consequently, CloSe-D comprises a curated segmentation label dataset that has been meticulously verified.

CloSe-T for Network Refinement. We also use CloSe-T to improve the generalization of our model for real-world datasets. We first predict the segmentation label for a given scan using the pre-trained CloSe-Net. Since the given scan is out-of-distribution, network results might be incorrect and noisy. We then refine the network prediction for the given scan using the steps mentioned in *data annotation*. We explain training details and experiments in Sec. 4. The new network is used to infer the given scan again and also evaluated on the test-set of CloSe-D. All these functions are implemented as a simple button click in the tool, see Fig. 4. The newly trained model can be saved and used of this new out-of-distribution dataset for better generalization.

4. Results

In this section, we provide more results of our model. In Sec. 4.1, we provide more comparison with baseline methods, followed by comparison on BUFF [21] dataset in Sec. 4.2. Finally, we provide ablation studies for continual learning setup of our model and show more results on real-world datasets in Sec. 4.3.

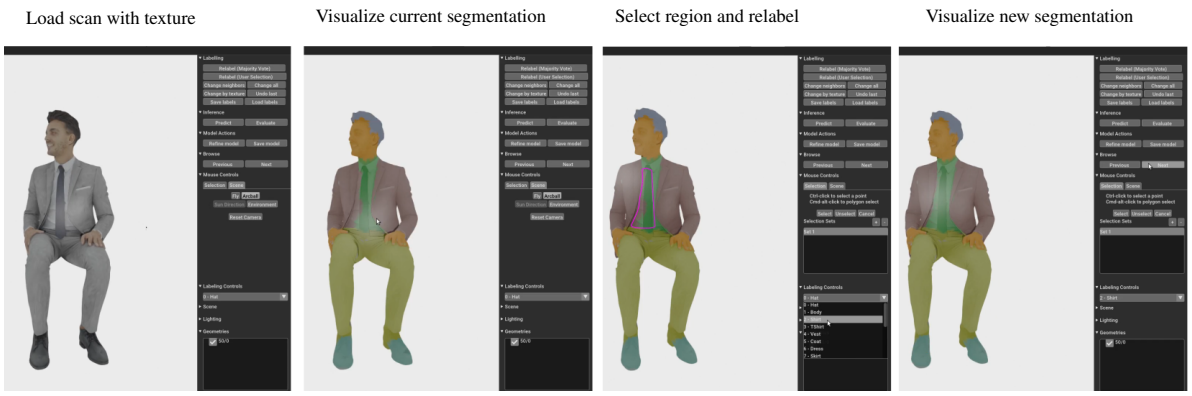

Figure 5. **Annotation using CloSe-T.**Using CloSe-T, we first *load the scan with texture* to understand the scan. We then *visualize current segmentation* as an overlay on the textured scan. After inspection, we identify and *select mislabeled regions* and assign them the correct label from a predefined set. Finally, we *visualize the new segmentation* and inspect by moving the camera around the scene.

4.1. Comparison with baseline

In this section, we analyze more comparisons with part segmentation methods to understand the cause of superior performance of CloSe-Net. We broadly classify them into 5 factors, as discussed below. These factors act mutually in many cases, widening the disparity between the performance levels of baseline techniques and our proposed approach. In the table Table 1 we provide a quantitative comparison on the test split of CloSe-Di.

Clothing Information. Baseline methods DGCNN [17] and DeltaConv [18] have no prior about clothing present in the scan. As a result, these methods rely on local/global geometric and appearance features. Given the diversity and complexity of clothing items, it is challenging to learn about robust semantics from limited information. As a result, baseline methods seem to generate multiple clothing classes in a vicinity, mislabel clothing classes, and are not able to learn the shape/structure of clothing items. This is evident from all the examples shown in Fig. 6. CloSe-Net not only takes advantage of clothing information but also learns a more distinctive feature for each clothing class and consequently learns clothing prior based on local features and these clothing features(via attention module).

Texture Bias. As observed in Fig. 6(first and second row), baseline methods are highly sensitive to changes in texture. As a result any steep change in texture results in a new clothing class. However CloSe-Net produces accurate results. For baseline methods color, normal and location are the only guiding signal without any prior. Given limited training data, they tend to overfit to textures scene during training.

Multi-layer Clothing. We also observe that baseline methods are not able to recover multi-layer clothing labels see Fig. 6(third row). As there is no prior knowledge about clothing present in the scan, baselines rely on texture and geometry information. In such cases, baselines seem to predict the most commonly seen example with texture during training such as hoodies or shirts. On the other hand, the clothing information used in CloSe-Net helps with better comprehension, even if local features are very similar.

Shape/geometry Bias. Similar to texture bias, the baseline method also has geometry bias to some extent. As shown in Fig. 6(fourth row) loose upper clothing with larger shapes are classified as hoodies, although the labels are not noisy.

Sparse Clothing Classes. We also observe that CloSe-Net performs well for rare clothing classes such as dresses, hats, etc. On the other hand baseline methods fail to generate consistent labels.

4.2. Comparison with Prior Work

We compare or model with prior work GIM3D [12] on Buff dataset [21]. We use 15 scans from BUFF, as in [12] for evaluation on the 3-class segmentation problem. We use PointNet++ [13] based model from GIM3D and report the number in Table 2. We observe that for both CloSe-D-test and BUFF dataset, our model significantly outperforms GIM3D [12].

4.3. CloSe-Net on Real-world Datasets

We qualitatively evaluate CloSe-Net on publicly avaialble real-world datasets such as THuman2.0 [20], THuman3.0 [14], HuMMan [7], 3DHumans [9]. We have added more results in Fig. 8. We observe that for all datasets, CloSe-Net generates good results and generalizes well. However, in some cases, it results in blurry boundaries and noisy patches of labels, as shown in Fig. 7.

Method	Mean	T-shirt	Shirt	Vest	Coat	Hoodies	Short-Pants	Pants	Skirts	Hat	Shoes	Body	Hair
DGCNN [17]	92.65	97.50	93.23	95.78	86.89	99.54	96.89	87.27	98.90	97.26	86.17	84.19	88.14
DeltaConv [18]	91.30	97.19	88.12	96.57	86.98	98.55	94.39	86.87	98.69	97.26	83.42	80.33	87.29
Ours	**95.19**	**99.12**	**96.18**	**99.48**	**87.93**	**99.69**	**97.98**	**89.39**	**99.05**	**99.06**	**89.97**	**89.78**	**94.66**

Table 1. We quantitatively compare the results of our method SotA part-segmentation methods, DGCNN [17] and DeltaConv [18]. We report IoU for every class and mean over all the classes(IoU_{mean}).

Figure 6. Baseline method like DGCNN [17] and DeltaConv [18] have **Texture bias** (a, b), are unable to distinguish between **multi-layer clothing** (c), produces incorrect labels if **geometry deviates significantly from average body and clothing shapes** (d) and underperform for **unbalanced classes** such as dress and hats(e, f).

Dataset	MGN [6]	GIM3D [12]	Ours
CloSe-D-Test	88.88	72.04	**92.47**
Buff [21]	-	75.41	**90.13**

Table 2. Comparison with MGN [6] and GIM3D [12] on CloSe-D and BUFF dataset.

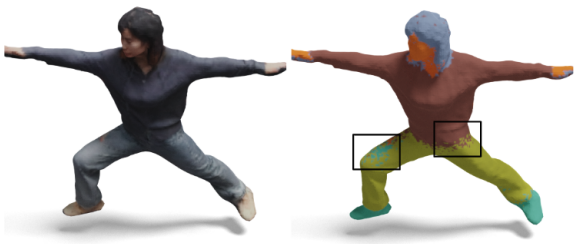

Figure 7. CloSe-Net predicts blurry boundaries for out-of-distributions scans.

We propose to improve the performance of our model for such out-of-distribution scans, by fine-tuning the model in a continual learning framework. We follow [16] and experiment with various loss combinations and training configurations to find an optimal setup, such that network performance improves on new out-of-distribution scans without catastrophic forgetting. We show the ablation in Table 3. We compare the mean IoU on test split of CloSe-D, after iteratively fine-tuning on 2 sets of scans from this new distribution. Based on experiments, we pick the full loss(eq. 5, main paper) as training loss and only train the last layer of the segmentation decoder and MLP of the Point Encoder. We fine-tune the model for 2 epochs only.

Table 3. Performance(IoU_{mean}) on CloSe-D-test after network refinement.

Layers trained	Naive loss	Weighted cross-entropy	Full
f_{dec}-last	90.33	90.37	90.25
f_{dec}-full	89.14	88.53	88.50
f_{dec}-last+ f_{MLP}	90.62	90.53	90.33
f_{dec}-full+ f_{MLP}	89.00	88.53	88.95
f_{dec}-last+ f_{MLP} + f^3	90.53	90.18	90.35
f_{dec}-full+ f_{MLP} + f^3	89.00	88.53	88.62

Segmenting 4D Scans using CloSe-Net and CloSe-T. We use the aforementioned setup to improve segmentation accuracy for a given 4d sequence. We randomly pick one frame of a 4D sequence and refine the model as per this scan. This is similar to one-shot fine-tuning. Then we generate the segmentation labels for the whole sequence. Since the model has now learned appearance and geometry features of one frame, this results in improved accuracy for remaining frames. We show results on a set of poses from THuman3.0 and HuMMan inFig. 9.

Finally, we have generated high quality segmentation labels of approximately 1000 scans(from diverse sources [7, 9, 14, 20]) using CloSe-Net and CloSe-T. We will release this as CloSe-D++.

References

[1] AXYZ design. 1

[2] Agisoft Metashape: Reconstruction from Images. 1, 2

[3] 3D People from Renderpeople. 1

[4] Treedy static scanner. 1

[5] Twindom 3D Scans. 1

[6] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3D people from images. In *ICCV*, 2019. 1, 2, 6

[7] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In *17th European Conference on Computer Vision, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 557–577. Springer, 2022. 4, 6, 7, 8

[8] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 3

[9] Sai Sagar Jinka, Astitva Srivastava, Chandradeep Pokhariya, Avinash Sharma, and P. J. Narayanan. Sharp: Shape-aware reconstruction of people in loose clothing. *International Journal of Computer Vision*, 2022. 4, 6, 7

[10] Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 360-degree textures of people in clothing from a single image. In *2019 International Conference on 3D Vision (3DV)*, pages 643–653. IEEE, 2019. 1

[11] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 1

[12] Umberto Castellani Pietro Musoni, Simone Melzi. GIM3D: A 3d dataset for garment segmentation. *placeholder*, 2022. 4, 6

[13] 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*, page 5105–5114, Red Hook, NY, USA, 2017. Curran Associates Inc. 4

[14] Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1581–1593, 2023. 4, 6, 7, 8

[15] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing. In *ECCV*, 2020. 1

Figure 8. CloSe-Net results on real-world public datasets [7, 9, 14, 20].

[16] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application, 2023. 6

[17] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 2019. 2, 4, 5

[18] R Wiersma, A. Nasikun, E Eisemann, and K Hildebrandt. Deltaconv: Anisotropic operators for geometric deep learning on point clouds. *Transactions on Graphics*, 41(4), 2022. 4, 5

[19] T Yenamandra, A Tewari, F Bernard, HP Seidel, M Elgharib, D Cremers, and C Theobalt. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[20] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human vol-

Figure 9. CloSe-Net is fine-tuned using CloSe-T on a single frame of a sequence to improve generalization on the remaining frames. We show results of fine-tuned CloSe-Net on THuman3.0 [14](top) and HuMMan [7](bottom). Fine-tuned networks result in consistent predictions.

umetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, 2021. 4, 6, 7

[21] Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 4, 6

[22] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing, 2018. cite arxiv:1801.09847Comment: http://www.open3d.org. 2