

CloSe: A 3D Clothing Segmentation Dataset and Model

Dimitrije Antić¹ Garvita Tiwari^{2,3,4} Batuhan Ozcomlekci² Riccardo Marin^{2,3} Gerard Pons-Moll^{2,3,4}

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

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

d.antic@uva.nl, gtiwari@mpi-inf.mpg.de, batuhan.oezcomlekci@student.uni-tuebingen.de,
{riccardo.marin, gerard.pons-moll}@uni-tuebingen.de



Figure 1. *Left*: We present **CloSe-D**, a large-scale dataset of people in clothing with fine-grained segmentation labels. We use this dataset to train our clothing segmentation model, **CloSe-Net**, tailored to segment clothing from 3D scans. *Right*: We show results of CloSe-Net on the diverse set of scans, where each instance represents GT, Input, and Prediction.

Abstract

3D Clothing modeling and datasets play crucial role in the entertainment, animation, and digital fashion industries. Existing work often lacks detailed semantic understanding or uses synthetic datasets, lacking realism and personalization. To address this, we first introduce **CloSe-D**: a novel large-scale dataset containing 3D clothing segmentation of 3167 scans, covering a range of 18 distinct clothing classes. Additionally, we propose **CloSe-Net**, the first learning-based 3D clothing segmentation model for fine-grained segmentation from colored point clouds. CloSe-Net uses local point features, body-clothing correlation, and a garment-class and point features-based attention module, improving performance over baselines and prior work. The proposed attention module enables our model to learn appearance and geometry-dependent clothing prior from data. We further validate the efficacy of our approach by successfully segmenting publicly available datasets of people in clothing. We also introduce **CloSe-T**, a 3D interactive tool for refining

segmentation labels. Combining the tool with CloSe-Net in a continual learning setup demonstrates improved generalization on real-world data. Dataset, model, and tool can be found at <https://virtualhumans.mpi-inf.mpg.de/close3dv24/>.

1. Introduction

Clothing plays a vital role in shaping our identity. Our dressing choices contribute to our representation, conveying cultural traits, religious beliefs, geographical origin, or mood. In light of the recent attention on the expressiveness of avatarization processes, Computer Vision scholars dedicate tremendous efforts in acquiring, modeling, and comprehending digital clothing, enabling uncountable applications: from digital fashion to AR/VR; from home entertainment to industrial-scale content creation.

Semantic understanding of humans in clothing from 2D images has seen significant progress [16, 17, 40, 57], but the lack of geometrical information is an obstacle for AR/VR

applications, where actions occur in a 3D world. Despite the development of 3D data capture techniques [1, 3, 7] and the consequent abundant geometrical data of 3D people in clothing [12, 20, 29, 59], garment analysis and its semantic understanding remains an open problem.

The primary challenge lies in representing 3D digital garments and their semantics. Certain existing approaches rely on synthetic clothing meticulously designed by experts [9, 35] or involve expensive acquisitions [23, 66], where accurate information comes at the expense of scalability. Some recent methods address human and clothing models within a unified representation [14, 42, 48] but lack a semantic understanding of the distinct parts. We posit that the absence of large-scale and high-quality segmented datasets serves as a fundamental barrier.

Consequently, the development of robust methods for understanding the 3D cloth semantics has been hampered. Prior work like MGN-Seg [10] requires expensive pipelines involving rendering, SMPL+D registration [24], 2D segmentation [16], and hand-crafted clothing priors, making it time-consuming (15-20 minutes per scan) and prohibitive for complex clothing items. In contrast, GIM3D [36] uses a SotA part segmentation method trained on synthetic data, accounting only for geometric information, namely location and normals. Both are limited by 3-class prediction: upper and lower garments and the human body, which trivialize the problem. Considering the clothes variations in style and textures, getting semantic information from 3D scans requires fast, accurate, generalizable, and scalable method.

This work aims to fill this gap with a three-fold contribution. First, we introduce *CloSe-D*, a large-scale dataset of people in clothing with fine-grained segmentation labels, for a total of 3167 scans and 18 garments categories. To our knowledge, this is the first real-world dataset with such fine-grained segmentation labels. Secondly, we use CloSe-D to train *CloSe-Net*, a 3D clothing segmentation model to predict 18 distinct types of clothing. Our model is based on two key intuitions: we correlate body parts with clothing classes, leveraging the SMPL [28] body model (Sec. 4.1.2); and we address the relationship between local geometric-appearance cues and clothing class in segmentation using an attention module (Sec. 4.1.3), learning associations between point features and clothing classes. We demonstrate a significant improvement over baselines and prior works (Sec. 5.2 and Sec. 5.2.1). Finally, to achieve the most comprehensive generalization possible, we develop *CloSe-T*, an interactive 3D tool to provide quick human feedback/annotation. This tool allows users to refine segmentation predictions and rectify segmentation labels. Our tool can also be integrated with CloSe-Net, where the feedback is backpropagated, and the network is fine-tuned in a continual learning setup (Sec. 5.3). We leverage this tool to prepare high-quality segmentation training data CloSe-D. We also use the tool in conjunction

with CloSe-Net to prepare high-quality segmentation labels on public datasets, which we release as CloSe-D++.

In summary, our contributions are:

- **CloSe-D**: A high-quality fine-grained clothing segmentation dataset containing segmentation labels for 3167 scans, covering 18 clothing classes.
- **CloSe-Net**: A human prior and clothing classes attention-based 3D clothing segmentation method that outperforms baselines and prior work.
- **CloSe-T**: A 3D interactive tool to refine the model in a continual learning framework, improving generalization to new datasets.
- **CloSe-D++**: Fine-grained semantic segmentation for a subset of publicly available real-world datasets.

We release our data, model, and tool for further research.

2. Related Work

Our work includes a new 3D clothing dataset, a 3D segmentation model, and an interactive refinement method, thus the related work covers these three areas.

2.1. 3D Clothing Datasets

The rise of learning-based digital fashion and virtual try-on led to the creation of image-based clothing datasets with semantic labels [17, 27]. However, these datasets lack pose variation, are 2D and mostly frontal, and are unsuitable for learning overall human/clothing shape, deformations, and 3D/4D models.

3D/4D Clothing datasets can be grouped into two categories: Synthetic and Captured ones. Synthetic datasets [9, 30, 35, 51] are obtained using physics-based simulation software [4], their generation requires expert intervention and mostly contain geometric information without texture. Moreover, this may not scale for complex clothing and multiple layers, often resulting in non-realistic deformations.

On the other hand, the accessibility of capturing datasets has increased recently, courtesy of advancements in 3D/4D capture systems [1, 7]. THuman1-4 [43, 45, 59, 64] propose medium to high-quality static scans of individuals in limited clothing styles/variations with limited pose variations. BUFF/CAPE [29, 37, 60], and HuMMan [12] provide dynamic scans of subjects in different clothing items. These datasets contain point clouds, occasionally texture, and SMPL parameters, but lack clothing semantic segmentation. Prior work such as MGN [10], SIZER [47], GIM3D [36] provide coarse 3D clothing segmentation labels, containing only three categories, namely *upper garment*, *lower garment*, and *body*. Deepfashion3D [66] is a dataset of high-quality and diverse 3D Clothing items, scanned on mannequins consisting of 10 clothing classes with keyline annotations and SMPL pose parameters. However, the clothing items are scanned separately and cannot be used to create a fully clothed per-

Dataset	# scans	Segmentation	Garment Class	Texture
MGN [10]	~300	✓	3	✓
SIZER [47]	~2000	✓	3	✓
DeepFashion3D [66]	2078(563)	✓	10	✓
THuman [45, 59, 63]	~1000	✗	-	✓
CloSe-D	~3000	✓	18	✓
CloSe-D++	~(+1000)	✓	18	✓

Table 1. Compared to current 3D clothing(real and static) datasets, CloSe-D is the first extensive dataset featuring fine-grained clothing segmentation labels and a variety of clothing items.

son. In contrast, our dataset is more realistic, featuring a broader range of clothing classes.

None of the existing real-world clothing datasets contains fine-grained clothing segmentation labels of clothed humans. In CloSe-D, we provide clean and fine-grained clothing segmentation labels of scans along with colored point clouds, and SMPL [28] parameters. We compare CloSe-D with existing real-world datasets in Table 1.

2.2. 3D Clothing Segmentation

2D clothing segmentation and human parsing have been extensively studied. Several prior works [16, 25, 55, 56, 62] propose learning-based human parsing in images, trained using 2D clothing datasets such as [15, 17]. These methods exploit human body parts and pose information to improve accuracy and generalization. However, they cannot be used for 3D segmentation, as they are not trained to produce multi-view consistent results, and lifting labels from 2D to 3D requires a slow optimization process.

MGN-Seg [10] is an optimization-based 3D clothing segmentation method that involves registering scans to SMPL+D [10, 11], applying PGN [16] for 2D segmentation of multi-view renderings of a mesh. These 2D segmentation images are then lifted back to 3D by solving GrabCut [41] in SMPL-UV space with a handcrafted clothing class based prior. This takes roughly 20 minutes per scan, is limited to 3 clothing classes, and requires expensive SMPL+D registration and hand-crafted clothing prior. On the other hand, CloSe-Net only needs colored point cloud, SMPL(θ , β) parameters, and clothing classes present in the scan and learns clothing prior from data.

3D clothing segmentation from a point cloud resembles a part segmentation setup. There are various existing works in learning-based 3D part segmentation, such as the seminal PointNets [13, 38] and self-attention-based PointTransformer [61]. More recent methods like KPConv [46] use kernel points defined in Euclidean space to apply convolution. SGPN [52] uses a single network to predict point grouping proposals and a corresponding semantic class for each proposal. Recent SotA methods [18] like DGCNN [53] introduces EdgeConv, a differentiable layer representing data on graphs dynamically computed in each network layer. Delta-

Conv [54] uses anisotropic convolution layers and introduces a convolution layer that combines geometric operators from vector calculus to construct anisotropic filters on point clouds. Existing works lack evaluations on clothing segmentation and do not integrate human-clothing-specific knowledge. In contrast, CloSe-Net employs a DGCNN-based point-feature module, leveraging human body and clothing priors for improved performance in clothing segmentation. We also show in our experiments that SotA DGCNN and DeltaConv, have limitations when applied to this task.

Most relevant to our work is GIM3D [36], which uses SotA part segmentation methods to segment directly from a point cloud, but similar to [10, 47] it is limited to three classes. To our knowledge, no existing method directly operates on a colored 3D point cloud/mesh to generate fine-grained 3D clothing segmentation. GIM3D+ [33] extends GIM3D, by including more diverse fabrics, sizes, and poses in the dataset, but is still limited to three classes and doesn't consider texture information.

2.3. 3D Interactive Segmentation and Refinement

Tools for interactive segmentation refinement are crucial for developing large-scale datasets and incorporating human feedback for improving segmentation. In 2D, classical methods like GrabCut [41] and a more recent interactive segmentation refinement [44] use human feedback input to improve segmentation. In 3D, 3D-GrabCut [31] and mesh cutting tools [19] are used for foreground/background segmentation. However, these methods do not leverage learned neural features and initial predictions. Recent works like [22, 39] are related to 3D interactive part segmentation annotation tools. iSeg3D [39] utilizes primitive-aware embedding and doesn't consider color information of data. InterObject3D [22] proposes a generalizable pipeline for interactive 3D segmentation, where the network is refined based on user clicks for a given target domain. Yet, none of these methods is designed for clothing segmentation and doesn't consider catastrophic forgetting. We introduce CloSe-T, a novel, fast, and easy-to-use interactive tool, tailored for the clothing domain. We propose the network refinement in a continual learning setup [50], such that the network not only performs well on the target domain but also learns from it, improving generalization. Prior works like [21, 26] introduce weighted loss term and EWC(elastic weight consolidation) to avoid catastrophic forgetting. We use [26] based weighted loss in our setup.

3. CloSe-D: 3D Clothing Segmentation Dataset

We introduce CloSe-D, a large-scale 3D clothing segmentation dataset that contains labels for 3167 scans comprising of 18 garment classes. CloSe-D consists of scans from two kinds of sources: 1) CloSe-Di: Scans captured using Treddy [7] scanner, and 2) CloSe-Dc: Scans from the

Class	T-shirt	Shirt	Vest	Coat	Jacket	Hoodies	Short-Pants	Pants	Skirts
Data									
CloSe-Di	415	556	85	191	-	209	500	897	59
CloSe-Dc	775	739	107	306	42	42	252	1404	34

Class	Dress	JumpS.	SwimS.	UnderG.	Scarf	Hat	Shoes	Body	Hair
Data									
CloSe-Di	-	-	-	-	-	114	1437	1455	1382
CloSe-Dc	50	6	23	10	25	64	1686	1732	1717

Table 2. Number of scans per clothing class in CloSe-D (i.e., the union of CloSe-Di and CloSe-Dc).

commercial datasets, such as twindom, renderpeople [2, 6–8]. For CloSe-Di we will release scans, SMPL parameters, and segmentation labels, while for CloSe-Dc we will provide SMPL parameters and segmentation labels only, due to license concerns. We show examples from our dataset in Fig. 1(left), and the details in Table 2.

Ground Truth Segmentation Labels. To obtain the ground truth segmentation labels, we adopt the pipeline utilized in [10, 47], as mentioned in Sec. 2.2. However, unlike MGN-Seg [10], our pipeline does not require SMPL+D registration, as we directly apply all the steps on scan and use [5] for lifting labels to 3D. Due to inconsistent predictions across views and limited generalization of 2D segmentation methods, the 3D segmentation labels may contain noise. To address this, we manually refine the segmentation using CloSe-T, which is explained in Sec. 4.2.

4. Method

In this section, we introduce CloSe-Net (Sec. 4.1), a 3D clothing segmentation model that predicts fine-grained clothing labels from a colored point cloud. Additionally, we present CloSe-T (Sec. 4.2), a 3D interactive tool used for creating high-quality segmentation labels of CloSe-D. We demonstrate the utility of CloSe-T in enhancing the generalization of our model on real-world datasets.

4.1. CloSe-Net: 3D Clothing Segmentation Network

Overview. CloSe-Net predicts fine-grained clothing segmentation labels directly from colored point clouds, given SMPL parameters and clothing classes as input. As shown in Fig. 2, CloSe-Net consists of four modules: *Point Encoder*, *Body Encoder*, *Clothing Encoder*, and *Segmentation Decoder*. Previous methods (e.g., MGN-Seg [10]) manually define clothing priors, leading to poor generalization across garment styles. In contrast, our model learns clothing priors by establishing the correlation between body parts and local clothing, utilizing the *Body Encoder*, and understanding the connection between point features and clothing class through the *Clothing Encoder*. As a result, our approach learns a prior that incorporates the garment’s style, body information, and the combined local geometric and appearance features.

Input/Output. CloSe-Net takes a point cloud $\mathbf{P} \in \mathbb{R}^{n \times 9}$ as input, consisting of n points, denoted as $\mathbf{p}_i = \{\mathbf{x}_i | \mathbf{c}_i | \mathbf{n}_i\}$, where $\mathbf{x}_i \in \mathbb{R}^3$ represent Euclidean coordinates, $\mathbf{c}_i \in \mathbb{R}^3$

represent per-point colors and $\mathbf{n}_i \in \mathbb{R}^3$ represent normals. CloSe-Net predicts per-point segmentation labels, $y_i \in \{1 \dots K\}$, where K is the number of classes.

As CloSe-Net incorporates human body prior and clothing class-based attention module, our method also needs SMPL (θ, β) parameters and clothing classes (g) of the scan. We use SMPL registration library [11, 24] to obtain SMPL parameters. For clothing class labels, we render a single viewpoint of the scan and infer the clothing classes using a SotA human parsing network [40].

4.1.1 Point Encoder

Semantic/part segmentation of a point cloud needs meaningful local and global geometric features [18]. Following this, we implement our Point Encoder, f_{point} using SotA *EdgeConv* based architecture, called DGCNN [53]. DGCNN operates on a point cloud by constructing a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} \in \{1, \dots, n\}$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. The edges \mathcal{E} are obtained using k -nearest neighbors of $\mathbf{p}_i \in \mathbb{R}^F$, where $F = 9$ for first layer, and 64 for subsequent layers. We then calculate per-point features using Edge Convolution given in Eq. (1), where h_θ is the learnable edge feature layer.

$$\mathbf{p}'_i = \max_{j:(i,j) \in \mathcal{E}} h_\theta(\mathbf{p}_i, \mathbf{p}_j) \quad (1)$$

Similar to [53], we use 3 EdgeConv layers followed by an MLP (f_{MLP}) to learn a global encoding. This results in multi-scale per-point features given by $F_i^{\mathbf{p}} = \{\mathbf{p}'_i^s | \mathbf{p}'_g\}$, where $s = \{0, 1, 2\}$, $\mathbf{p}'_i^s \in \mathbb{R}^l$ is per-point feature learned by s^{th} EdgeConv layer and $\mathbf{p}'_g \in \mathbb{R}^{1024}$ is a global encoding of the point cloud. \mathbf{p}'_g is obtained using an MLP, $\mathbf{p}'_g = f_{\text{MLP}}(\mathbf{P}')$, where $\mathbf{P}' = (\mathbf{p}'_0, \dots, \mathbf{p}'_n) \in \mathbb{R}^{n \times (l+l+l)}$, $\mathbf{p}'_i = \{\mathbf{p}'_i^0 | \mathbf{p}'_i^1 | \mathbf{p}'_i^2\}$ is the concatenation of intermediate per-point features.

4.1.2 Body Encoder

We incorporate the correlation between body parts and clothing class using SMPL mesh, given by $M(\beta, \theta)$ and SMPL template mesh \mathbf{T} . In particular, for every point \mathbf{x}_i in the input point cloud, we find the index (j) of the nearest vertex on SMPL mesh, given by $M_j(\beta, \theta)$. We then find the corresponding vertex location in SMPL template \mathbf{T} . In this way, we associate each point in the point cloud with fine-grained semantic information about the human body. This module is not learnable and only requires a nearest-neighbor search in $\mathbb{R}^{n \times 3}$. We represent the encoded body feature as $F_i^{\mathbf{b}} = \mathbf{T}_j$ and call this a Canonical Body Encoder.

4.1.3 Clothing Encoding and Class-based Attention

Clothing classification is ambiguous due to its subjective nature (e.g., think about the difference between jackets and coats). To address this, we use a learnable codebook

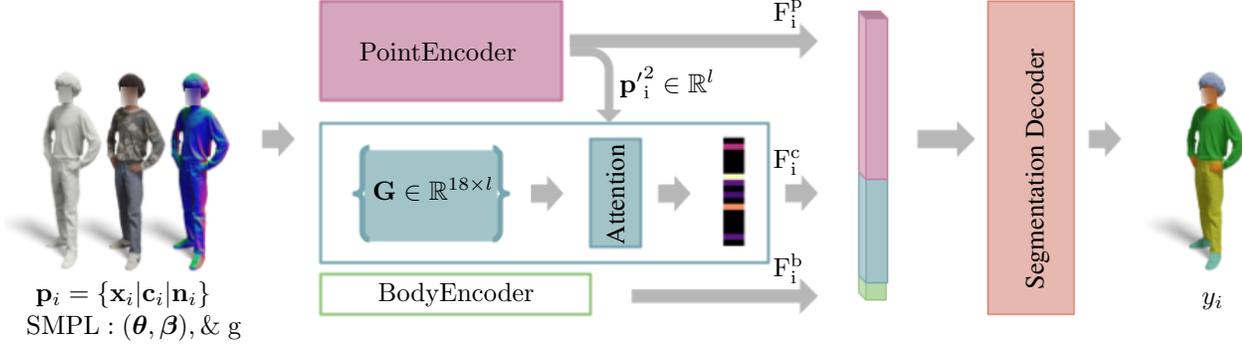


Figure 2. **CloSe-Net**: Given a colored point cloud $\mathbf{P} = \{\mathbf{p}_1 \dots \mathbf{p}_n\}$ with SMPL parameters (θ, β) , and clothing classes (g) detected in the scan, where $\mathbf{p}_i = \{\mathbf{x}_i | \mathbf{c}_i | \mathbf{n}_i\}$ represent point location, color and normal of a point, CloSe-Net predicts fine-grained per-point segmentation labels. (a) **Point Encoder** (Sec. 4.1.1) takes \mathbf{P} , as input and predicts per-point features F^P . (b) **Clothing Encoder** (Sec. 4.1.3) consists of a learnable codebook G and an attention module, which predicts F^C , based on per-point feature $\mathbf{p}_i^{/2}$ and G . This $\mathbf{p}_i^{/2}$ is intermediate feature of Point Encoder. (c) **Body Encoder** (Sec. 4.1.2), finds per-point canonical vertex in SMPL template, given SMPL θ, β parameters. (d) Finally, the **Segmentation Decoder** (Sec. 4.1.4) takes F^P, F^C, F^B and predicts segmentation labels, y_i for i^{th} point. Solid boxes in model are learnable, while others are fixed.

$G \in \mathbb{R}^{18 \times l}$. Our model learns distinct latent vectors [49] for each clothing class in an auto-decoding manner [34]. This enables the model to acquire per-clothing attributes and relevant characteristics for segmentation. The learned codebook is fixed during inference. Compared to a non-learnable binary/one-hot encoding-based representation, the learnable codebook yields better performance in challenging regions, like clothing boundaries and uncommon clothing items like jumpsuits (see Sec. 5.2.3). The key idea of our model is learning an explicit association between per-point features and clothing class. We implement this using an attention module, where we compute how much each point feature attends to a clothing feature. The attention between a point feature and a specific clothing latent code increases when that particular clothing is present at the query point. We define query vector using per-point EdgeConv features $\mathbf{p}_i^{/s}$, $s = 2$, and key-value pair using the learnable codebook G and define the attention mechanism in Eq. (2), where \circ is masking operator. Simply using $\mathbf{p}_i^{/s} \times G$, would also yield features for clothing items not present in the scan. This might result in learning spurious correlations between features and labels. To avoid this, we mask the key matrix G , using a binary encoding g of length $K = 18$, where $g_j = 1$, if j^{th} class is present in the scan, and $g_j = 0$ otherwise. Our model is steered towards learning the clothing-specific prior, leveraging fine-grained local point features and clothing latent codes.

$$F_i^C = \text{softmax}(\mathbf{p}_i^{/s} \times (g \circ G)^T) G. \quad (2)$$

4.1.4 Segmentation Decoder

Finally, we concatenate all the features $F_i^{\text{all}} = \{F_i^P | F_i^B | F_i^C\}$, and pass it through a segmentation decoder f_{dec} , which predicts per-point segmentation labels. The network f_{dec} is

implemented as an MLP.

Loss. We train CloSe-Net with the cross-entropy loss in Eq. (3), where \hat{y}_i^k is the true label, y_i^k is the predicted probability of the k^{th} class, and K is the number of classes.

$$\mathcal{L}_{\text{CE}} = - \sum_{k=1}^K \hat{y}_i^k \log(y_i^k) \quad (3)$$

4.2. CloSe-T

In the previous section, we presented our method which learns clothing prior from data for improved generalization over garment styles, appearance and categories. However, we foresee that a variety of clothes often shows unique characteristics that are difficult to catch statistically, especially with the datasets available at the present date. For this reason, we introduce CloSe-T, a fast-interactive tool to streamline the label refinement process. It provides a graphical interface built explicitly for clothing segmentation and relies on Open3D library [65], offering a broad set of functionalities (*e.g.*, points selection, labels updating, segmentation prediction, model refinement). We use CloSe-T to refine the training dataset, and to improve our network generalization on publicly available dataset by backpropagating user feedback in a continual learning setup [50].

Continual Learning CloSe-Net Refinement. Let y_i^k be the predicted probability of the k^{th} class label of i^{th} point in given pointcloud, and \hat{y}_i^k be the correct segmentation label provided by user using CloSe-T. We define the loss as:

$$\mathcal{L}_{\text{refine}} = \lambda_c \mathcal{L}_{\text{CE}}(y_i^k, \hat{y}_i^k) + \lambda_f \mathcal{L}_{\text{CE}}(y_i^k, \hat{y}_i^k) + \lambda_w \mathcal{L}_W(\theta, \theta'), \quad (4)$$

where \mathcal{C} is the set of indices of point cloud corrected by the user and \mathcal{F} is the set of remaining points. \mathcal{L}_{CE} is cross-entropy loss defined in Eq. (3), \mathcal{L}_W is weight regularization,

penalizing weights difference between refined model (θ') and original pre-trained model (θ). $\lambda_c, \lambda_f, \lambda_w$ are the weights associated with each loss term. We only fine-tune the last layer of the segmentation decoder and MLP of the Point Encoder. Following [26], we use $\lambda_c \ll \lambda_f$, in order to avoid catastrophic forgetting.

5. Experiments and Results

We describe the experiment setup in Sec. 5.1, evaluate our proposed CloSe-Net and compare it with SotA part segmentation in Sec. 5.2, and with prior methods in Sec. 5.2.1. We analyze each module of CloSe-Net in Sec. 5.2.3 and discuss the attention-based clothing prior in Sec. 5.2.2. Additionally, we present results on publicly available clothing datasets in Sec. 5.3 and showcase improvements in generalization through CloSe-T-based refinement.

5.1. Experimental Setup

Implementation Details: We use the official DGCNN implementation [53] with 3 EdgeConv layers (feature-length $l = 64$ and $|\mathbf{p}'_{\text{global}}| = 1024$). For the clothing codebook(\mathbf{G}), we use $l = 64$. The clothing-class-based attention module is based on multi-head attention. The train-val-test splits of CloSe-D are 2652/265/270.

Error Metric. Intersection over Union (IoU) is a popular metric for segmentation, quantifying the overlap between predicted and ground truth labels. We consider both per-class IoU and mean IoU (IoU_{mean}) over all classes.

5.2. 3D Clothing Segmentation

We evaluate CloSe-Net on test split of CloSe-D, both qualitatively (Fig. 14) and quantitatively (Table 3). Given the similarity between clothing segmentation and part segmentation tasks [36], we compare our method with SotA 3D part segmentation models trained on CloSe-D. For this, we employ methods [18] like DGCNN [53] and DeltaConv [54]. As observed in Fig. 14-middle, DGCNN and DeltaConv struggle with multi-layer clothing. This arises from the absence of clothing-related information in the model and its inclination to overfit to single-layer scans, which constitute the majority of the dataset. On the other hand, our model incorporates clothing information via the learned distinct codebook and hence predicts correct labels with crisp boundaries. Moreover, as seen in Fig. 14-top both baseline methods inadequately segment less common classes, such as hats. We also observe texture bias in both baseline methods (Fig. 14-bottom); both networks fail to predict the same class for nearby points with different texture colors. We attribute this to the codebook and masked attention module in our model, which enables the network to learn distinct features for each clothing and avoid learning spurious correlations between point features and absent classes.



Figure 3. Comparison with SotA part segmentation models: **DGCNN [53]** and **DeltaConv [54]**. Our model predicts accurate clothing classes and finer boundaries in complex scans. This can be attributed to our model’s utilization of local point features, body priors, and clothing class-based attention features.

5.2.1 Comparison with Prior Work

We compare CloSe-Net with prior 3D clothing segmentation methods like MGN-Seg [10] and GIM3D [36]. Since both GIM3D and MGN-Seg [10] predict 3 classes, we merge the predictions of our model into the respective classes. For GIM3D, we only compare with PointNet++ [38], as it is the only one authors provided to us. We observe from Table 4 that CloSe-Net largely improves over prior work. Moreover, CloSe-Net takes approximately 5 – 6 seconds to infer the segmentation labels for the whole scan(270k vertices) on 12 GiB GPU (3080Ti). Whereas for MGN-Seg [10] it takes roughly ~ 20 minutes to get the segmentation labels for SMPL+D mesh (27k vertices).

We further provide qualitative results in Fig. 4. We observe that MGN-Seg [10] is not able to generate precise labels at boundaries, *e.g.* at the leg. This is because the prior designed for lower garments comes from a fixed template of “long pants”, whereas the pant in this scan is smaller than the pre-defined template. We also observe that sometimes the handcrafted features are not able to correct segmentation errors due to texture bias and inconsistent multiview prediction of 2D human parsing [16, 40]. This is visible in the right hand of the scan. We notice that GIM3D fails near boundaries. This stems from the fact that normal information is not sufficient to distinguish between different clothing, especially in the case of real-world scans, where normals can be noisy. On the other hand, our model results in accurate boundaries as it takes texture information into

Method	Mean	T-shirt	Shirt	Vest	Coat	Jacket	Hoodies	Short-Pants	Pants	Skirts	Dress	JumpS.	SwimS.	UnderG.	Scarf	Hat	Shoes	Body	Hair
DGCNN [53]	87.11	87.88	81.02	90.58	81.60	96.16	97.46	94.60	82.50	96.44	73.61	76.67	98.89	99.26	73.33	95.19	79.47	80.45	82.83
DeltaConv [54]	84.78	87.22	73.56	84.68	80.06	98.52	96.99	89.58	78.37	94.11	67.08	77.04	99.26	99.26	73.33	95.19	72.98	76.69	82.13
Ours	91.23	95.47	92.94	98.86	90.12	99.23	99.43	98.32	85.96	98.12	79.11	77.73	99.78	99.96	73.23	97.72	82.96	85.76	87.49

Table 3. We quantitatively compare the results of our method SotA part segmentation methods, DGCNN [53] and DeltaConv [54]. We report IoU for every class and mean over all the classes(IoU_{mean}).

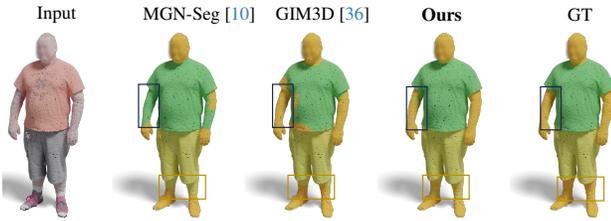


Figure 4. Comparison with MGN-Seg [10] and GIM3D [36].

Dataset	MGN [10]	GIM3D [36]	Ours
CloSe-D-Test	88.88	72.04	92.47

Table 4. Comparison with MGN [10] and GIM3D [36] on three class(upper, lower and body) segmentation.

account as well and doesn't rely on any pre-defined prior.

5.2.2 Learned Clothing Prior

A good clothing prior is crucial for segmentation. In MGN-Seg [10], a geodesic distance-based prior was manually crafted. This approach lacks scalability for new clothing and struggles with varying shapes within the same class (e.g., different jacket lengths or shirt sleeve styles *etc.*). Interestingly, our model's attention-based clothing encoder learns clothing prior from point features (F^p) and a garment codebook (F^c). This is visualized in Fig. 5, where attention for the j^{th} garment class at point i is calculated as $\mathbf{p}_i^k \times \mathbf{G}_j$. As compared to MGN-Seg [10] prior, this is not manually defined and also incorporates the local properties of the clothing.

5.2.3 Ablation

We experimentally validate and discuss the design choices of each module of CloSe-Net in Table 5 and Fig. 6.

Point Encoder. We experiment with two part segmentation models: DGCNN [53] and DeltaConv [54] as Point Encoder. Table 5 shows the DGCNN-based model significantly outperforms the DeltaConv-based one. This behavior is similar to the standalone DGCNN and DeltaConv, as shown in Table 3. DGCNN surpasses DeltaConv by learning feature spaces where semantically similar features are closely clustered. Hence, we use DGCNN as Point Encoder in our model.

Body Encoder. We explore different body encodings: *Canonical Body Encoder*, F^b (Sec. 4.1.2), and a fusion of F^b with a coarse feature encoder based on COAP [32], resulting in the *Hybrid Body Encoder*. The lack of Body Encoder in the model leads to mislabeled regions in improbable

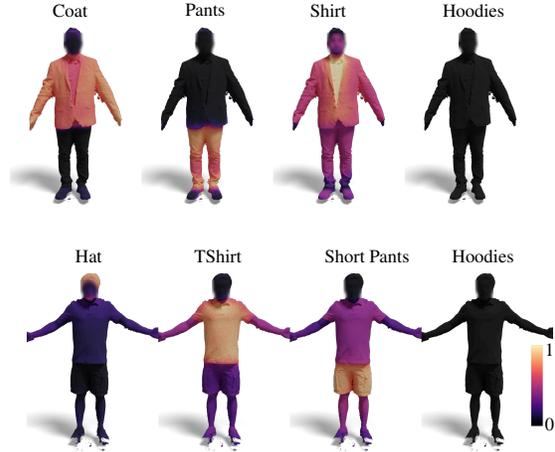


Figure 5. **Clothing prior learned using attention module:** Attention module in Clothing Encoder, learns a robust clothing prior based on point features. Here we visualise the attention of point feature on different clothing class.

locations, such as a t-shirt label appearing in the skirt region. These mislabels are evident as patches in Fig. 6-top. The *Hybrid Body Encoder* learns vertex and body part associations with the clothing, but results in smudged boundaries(see skirt and legs in Fig. 6-top). This is because the hybrid model contains body-part features, so it tends to associate the same labels to all the points in a body part if there is no significant difference in geometry or appearance. Our proposed Body Encoder uses fine-grained correspondence and establishes accurate correlations between clothing labels and body locations.

Clothing Encoder. We compare our attention-based clothing encoder, with a binary encoding-based one. In binary encoding, we use \mathbf{g} , instead of F^c . The attention-based model boosts performance and also learns garments prior from data. Binary encoding is not consistently effective in predicting the correct garment, particularly when confronted with uncommon clothing styles, as exemplified by the jumpsuit case in Fig. 6-bottom, similar to models without Clothing Encoder. Moreover, without any Clothing Encoder, models exhibit texture bias. Leveraging the learned clothing prior(Sec. 5.2.2) significantly improves performance, and alleviates the mentioned problems.

5.3. CloSe-D++

Our model generalizes well on real-world public datasets, showing good results in Fig. 7. However, it exhibits blurry boundaries and texture bias for some scans, as seen in Fig. 8

Points Encoder	Body Encoder	Clothing Encoder	IoU _{mean} ↑
DGCNN	Canonical	Attention	91.23
	Canonical	Binary	90.41
	Hybrid	Attention	89.70
	Hybrid	Binary	89.68
	×	Attention	89.90
	Canonical	×	87.18
	×	×	87.10
DeltaConv	Canonical	Attention	86.84

Table 5. Quantitative evaluation of the ablation study on different modules of the proposed CloSe-Net model. The table shows the performance of the model with different combinations of the Point Encoder, Body Encoder, and Clothing Encoder.

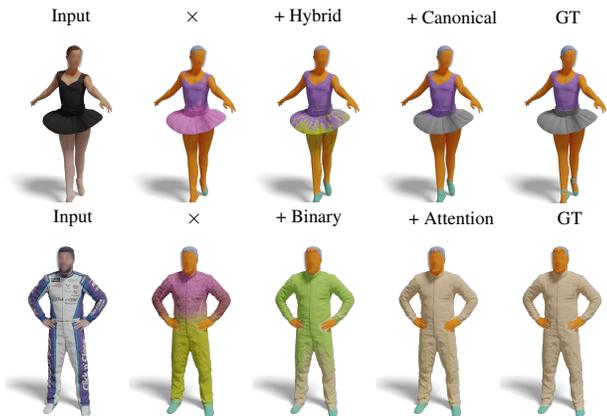


Figure 6. **Body Encoder**(Top): As opposed to others, the proposed Body Encoder(Canonical) is simple, generalizes to difficult poses, and produces fine boundaries. **Clothing Encoder**(Bottom): Attention-based encoder and codebook learn distinct garment features and prior, achieving accurate segmentation prediction.



Figure 7. Results of CloSe-Net on publicly available datasets [20, 59], showing generalization capability of the model.

(middle). Generalizing to out-of-distribution real-world datasets is challenging due to the vast variability of clothing styles and few differences between classes.

To address this, we use the proposed CloSe-T in a continual learning approach(Sec. 4.2). The goal is to improve performance over new datasets, without catastrophic forgetting. We show the results of the original model and fine-tuned model on a scan from THuman2 [59] in Fig. 8. After fine-tuning, results on CloSe-D-test is a mean IoU of 90.35, which is a small decrease from the original model(91.23) and still better than prior work and baselines. We will provide the labels of publicly available datasets such as THU-

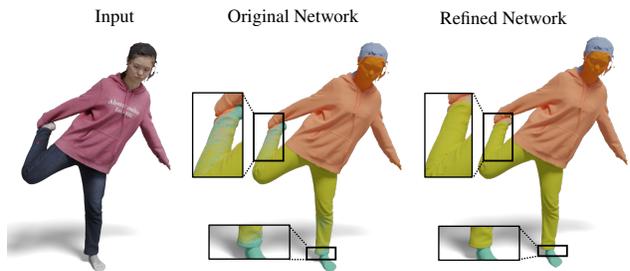


Figure 8. Improving CloSe-Net performance on THuman2.0 [59] by fine-tuning the model on few samples from THuman2.0 using CloSe-T.

man2,3 [45, 59], CAPE [29](textured-scans), 3DHumans-IIITH [20], a subset of HuMMan [12] and we call this dataset CloSe-D++.

6. Conclusions

We present a novel fine-grained 3D clothing segmentation model that works directly on point clouds, and to train it we introduce a large-scale dataset of people in diverse clothing items, poses, and with high-quality segmentation labels. We incorporate human body information to improve the pose-generalization of our model and introduce a novel garment class attention module, which learns clothing prior from data, as opposed to hand-crafted priors [10]. Our model outperforms prior work and baselines and generalizes to public out-of-distribution datasets. We further introduce a continual learning-based refinement strategy to improve the generalization of the model, without catastrophic forgetting. **Limitations and Future Work** To the best of our knowledge, this is the first dataset and model for 3D clothing segmentation from colored point clouds, which contains diverse and fine-grained segmentation labels. Future work may broaden our work by adding more clothing items through CloSe-T, *e.g.*, to include a variety of cultural styles. Our approach necessitates the garment class worn by the subject as network input, requiring a preprocessing step. Potentially, this can be obviated by incorporating clothing prediction directly within the network. Lastly, to enhance network generalization, we have integrated the continual learning [26] framework, paving the way for future exploration of recent strategies such as EWC [21].

Acknowledgments: Thanks to RVH, CVLab team, and reviewers for valuable feedback. The project was made possible by funding from the Carl Zeiss Foundation. This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans), German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. Gerard Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 - Project number 390727645. Riccardo Marin has been supported by the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101109330.

References

- [1] 3dMD. [2](#)
- [2] XYZ design. [4](#), [12](#)
- [3] LumaAI Labs. [2](#)
- [4] Marvelous Designer. [2](#)
- [5] Agisoft Metashape: Reconstruction from Images. [4](#), [12](#)
- [6] 3D People from Renderpeople. [4](#), [12](#)
- [7] Treedy static scanner. [2](#), [3](#), [12](#)
- [8] Twindom 3D Scans. [4](#), [12](#)
- [9] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. CLOTH3D: clothed 3d humans. *CoRR*, abs/1912.02792, 2019. [2](#)
- [10] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3D people from images. In *ICCV*, 2019. [2](#), [3](#), [4](#), [6](#), [7](#), [8](#), [12](#), [13](#), [17](#)
- [11] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. [3](#), [4](#)
- [12] 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. [2](#), [8](#), [15](#), [17](#), [18](#), [19](#)
- [13] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2017. [3](#)
- [14] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gdn: Towards generative detailed neural avatars. *arXiv*, 2022. [2](#)
- [15] Ke Gong, Xiaodan Liang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and A new benchmark for human parsing. *CoRR*, abs/1703.05446, 2017. [3](#)
- [16] 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](#), [2](#), [3](#), [6](#), [12](#), [14](#)
- [17] Sheng Guo, Weilin Huang, Xiao Zhang, Prasanna Srikhanta, Yin Cui, Yuan Li, Matthew R.Scott, Hartwig Adam, and Serge Belongie. The imaterialist fashion attribute dataset. *arXiv preprint arXiv:1906.05750*, 2019. [1](#), [2](#), [3](#)
- [18] Yong He, Hongshan Yu, Xiaoyan Liu, Zhengeng Yang, Wei Sun, Yaonan Wang, Qiang Fu, Yanmei Zou, and Ajmal Mian. Deep learning based 3d segmentation: A survey. *CoRR*, abs/2103.05423, 2021. [3](#), [4](#), [6](#)
- [19] Zhongping Ji, Ligang Liu, Zhonggui Chen, and Guojin Wang. Easy mesh cutting. *Computer Graphics Forum*, 25(3):283–291, 2006. [3](#)
- [20] 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. [2](#), [8](#), [15](#), [17](#), [18](#)
- [21] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. [3](#), [8](#)
- [22] Theodora Kontogianni, Ekin Celikkan, Siyu Tang, and Konrad Schindler. Interactive object segmentation in 3d point clouds, 2022. [3](#)
- [23] Zorah Löhner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part IV*, page 698–715, Berlin, Heidelberg, 2018. Springer-Verlag. [2](#)
- [24] 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. [2](#), [4](#), [12](#)
- [25] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, and Jiashi Feng. Towards real world human parsing: Multiple-human parsing in the wild. *CoRR*, abs/1705.07206, 2017. [3](#)
- [26] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. [3](#), [6](#), [8](#)
- [27] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [2](#)
- [28] 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. [2](#), [3](#), [12](#)
- [29] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to Dress 3D People in Generative Clothing. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. [2](#), [8](#)
- [30] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. [2](#)
- [31] Gregory P. Meyer and Minh N. Do. 3d grabcut: Interactive foreground extraction for reconstructed 3d scenes. In *3DOR@Eurographics*, 2015. [3](#)
- [32] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. COAP: Compositional articulated occupancy of people. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. [7](#)
- [33] Pietro Musoni, Simone Melzi, and Umberto Castellani. Gim3d plus: A labeled 3d dataset to design data-driven solutions for dressed humans. *Graphical Models*, 129:101187, 2023. [3](#)

- [34] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 5
- [35] Chaitanya Patel, Zhouyincheng Liao, and Gerard Pons-Moll. The virtual tailor: Predicting clothing in 3D as a function of human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 2
- [36] Umberto Castellani Pietro Musoni, Simone Melzi. GIM3D: A 3d dataset for garment segmentation. *placeholder*, 2022. 2, 3, 6, 7, 15, 17
- [37] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017. Two first authors contributed equally. 2
- [38] 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. 3, 6, 15
- [39] Sucheng Qian, Liu Liu, Wenqiang Xu, and Cewu Lu. iseg3d: An interactive 3d shape segmentation tool. 2021. 3
- [40] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. page 107404, 2020. 1, 4, 6
- [41] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. 2004. 3
- [42] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [43] Ruizhi Shao, Zerong Zheng, Hongwen Zhang, Jingxiang Sun, and Yebin Liu. Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *ECCV*, 2022. 2
- [44] Konstantin Sofiiuk, Ilya Petrov, Olga Barinova, and Anton Konushin. F-brs: Rethinking backpropagating refinement for interactive segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 3
- [45] 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. 2, 3, 8, 15, 17, 18, 19
- [46] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 3
- [47] 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. 2, 3, 4, 12
- [48] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-GIF: Neural generalized implicit functions for animating people in clothing. In *ICCV*, 2021. 2
- [49] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6309–6318, Red Hook, NY, USA, 2017. Curran Associates Inc. 5
- [50] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application, 2023. 3, 5, 16
- [51] Tuanfeng Y. Wang, Duygu Ceylan, Jovan Popovic, and Niloy J. Mitra. Learning a shared shape space for multi-modal garment design. *ACM Trans. Graph.*, 37(6):1:1–1:14, 2018. 2
- [52] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *CVPR*, 2018. 3
- [53] 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. 3, 4, 6, 7, 13, 15, 16
- [54] 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. 3, 6, 7, 15, 16
- [55] Kota Yamaguchi, Mohammad Hadi Kiapour, Luis E. Ortiz, and Tamara L. Berg. Parsing clothing in fashion photographs. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3570–3577, 2012. 3
- [56] Kota Yamaguchi, M. Hadi Kiapour, and Tamara L. Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, pages 3519–3526. IEEE Computer Society, 2013. 3
- [57] W. Yang and L. Luo, P. and Lin. Clothing co-parsing by joint image segmentation and labeling. 2014. 1
- [58] 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. 12
- [59] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, 2021. 2, 3, 8, 15, 17, 18
- [60] 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. 2, 14, 15, 17
- [61] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer, 2020. 3
- [62] Jian Zhao, Jianshu Li, Yu Cheng, Li Zhou, Terence Sim, Shuicheng Yan, and Jiashi Feng. Understanding humans in crowded scenes: Deep nested adversarial learning and A new

- benchmark for multi-human parsing. *CoRR*, abs/1804.03287, 2018. 3
- [63] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *IEEE Conference on Computer Vision (ICCV 2021)*, 2021. 3
- [64] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [65] 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>. 5, 13
- [66] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images, 2020. 2, 3

SUPPLEMENTARY MATERIALS

CloSe: A 3D Clothing Segmentation Dataset and Model

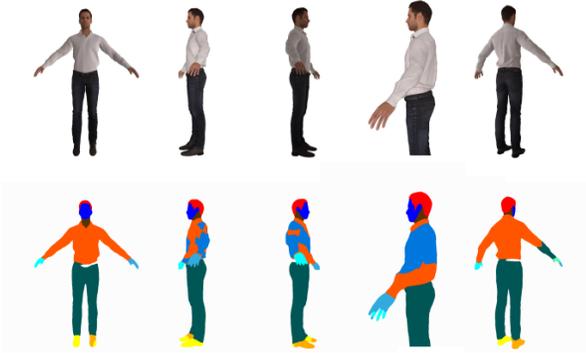


Figure 9. *Top*: Multiview rendered images of a scan. *Bottom*: Clothing segmentation obtained using 2D Parsing method [16]. 2D Parsing method generates inconsistent labels across views. Consequently, when these labels are elevated from 2D to 3D using the 2D-to-3D lifting technique, the resulting segmentation is noisy.

7. 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 XYZ [2], Twindom [8], Treedy [7], Renderpeople [6]. 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 [47, 58], we create a dataset of approximately 100 subjects in 7 diverse poses, wearing 12 garment classes. We use Treedy’s scanner [7], which consists of ~ 130 high-resolution camera at a fixed position. We use Metashape [5] 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 [28] to each scan, with the registration method used in [10, 24, 47].

Ground Truth Segmentation Labels of CloSe-Dc Scans.

We follow the pipeline similar to the one in MGN-Seg [10]. We first register the scans to SMPL and SMPL+D [10]. We then render the registered meshes from 72 different views and apply SotA 2D Human Parsing method, PGN [16]. One of the major limitations of such a pipeline is inconsistent multiview prediction of the 2D Human Parsing method, as shown in Fig. 9. 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 [10] 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. 13(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:

- *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. 11. 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. 10 and Fig. 11. We explain more details about our interactive tool in Sec. 9.

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 [5] 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 [10] 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.

8. Method

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

Point Encoder. We use the official implementation of DGCNN [53] and use 3 layers of EdgeConvolution operation, followed by a single-layer MLP.

Clothing Encoder. We use a multi-head attention module in the encoder, where $n_{\text{head}} = 4$ in our case. We also apply positional encoding to the query vector (\mathbf{p}'_i), 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 ($M(\beta, \beta)$). Subsequently, during inference, a preprocessing step is employed to calculate F^b beforehand, which is then used during inference.

9. Interactive Tool

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

Interactive Tool Interface. We implement CloSe-T using Open3D [65] 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, Loading/evaluating pre-trained model, Saving/Evaluating refined network.
- **Scene:** Move in the scene with mouse control, change lights, background, *etc.*



Figure 10. 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.

- **User selection:** Easy polygon-based region selection by selecting the polygon edges by clicking.
- **Labeling:** Relabel region based on user selection/majority vote.



Figure 11. Due inherent uncertainty in clothing classification, the segmentation labels acquired through [16] might be incoherent. However, such labeling discrepancies can be easily corrected using CloSe-T.

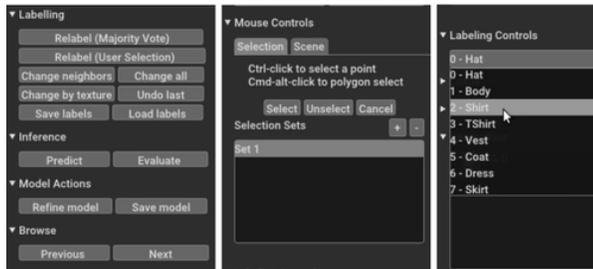


Figure 12. Functionalities provided in CloSe-T interface includes based I/O operations, mouse-controlled camera movement in the scene, region selection, relabelling, evaluation, and fine-tuning CloSe-Net.

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. 12(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. 10 and Fig. 11. We provide a demo of labeling process in the supplementary video and visualize key-stage of pipeline in Fig. 13.

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. 10. 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. 12. The newly trained model can be saved and used of this new out-of-distribution dataset for better generalization.

10. Results

In this section, we provide more results of our model. In Sec. 10.1, we provide more comparison with baseline methods, followed by comparison on BUFF [60] dataset in Sec. 10.2. Finally, we provide ablation studies for continual learning setup of our model and show more results on real-world datasets in Sec. 10.3.

10.1. 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 6 we provide a quantitative comparison on the test split of CloSe-Di.



Figure 13. **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.

Method	Mean	T-shirt	Shirt	Vest	Coat	Hoodies	Short-Pants	Pants	Skirts	Hat	Shoes	Body	Hair
DGCNN [53]	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 [54]	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 6. We quantitatively compare the results of our method SotA part-segmentation methods, DGCNN [53] and DeltaConv [54]. We report IoU for every class and mean over all the classes (IoU_{mean}).

Clothing Information. Baseline methods DGCNN [53] and DeltaConv [54] 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. 14. 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. 14 (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. 14 (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 pre-

dict 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. 14 (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.

10.2. Comparison with Prior Work

We compare our model with prior work GIM3D [36] on BUFF dataset [60]. We use 15 scans from BUFF, as in [36] for evaluation on the 3-class segmentation problem. We use PointNet++ [38] based model from GIM3D and report the number in Table 7. We observe that for both CloSe-D-test and BUFF dataset, our model significantly outperforms GIM3D [36].

10.3. CloSe-Net on Real-world Datasets

We qualitatively evaluate CloSe-Net on publicly available real-world datasets such as THuman2.0 [59], THuman3.0 [45], HuMMan [12], 3DHumans [20]. We have added more results in Fig. 16. We observe that for all datasets, CloSe-Net generates good results and generalizes



Figure 14. Baseline method like DGCNN [53] and DeltaConv [54] 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).

well. However, in some cases, it results in blurry boundaries and noisy patches of labels, as shown in Fig. 15.

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 [50] and experiment with various loss combinations and training configurations to find an optimal setup, such that network per-

formance improves on new out-of-distribution scans without catastrophic forgetting. We show the ablation in Table 8. 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.

Dataset	MGN [10]	GIM3D [36]	Ours
CloSe-D-Test	88.88	72.04	92.47
Buff [60]	-	75.41	90.13

Table 7. Comparison with MGN [10] and GIM3D [36] on CloSe-D and BUFF dataset.

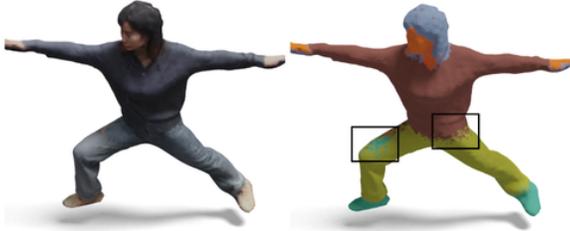


Figure 15. CloSe-Net predicts blurry boundaries for out-of-distributions scans.

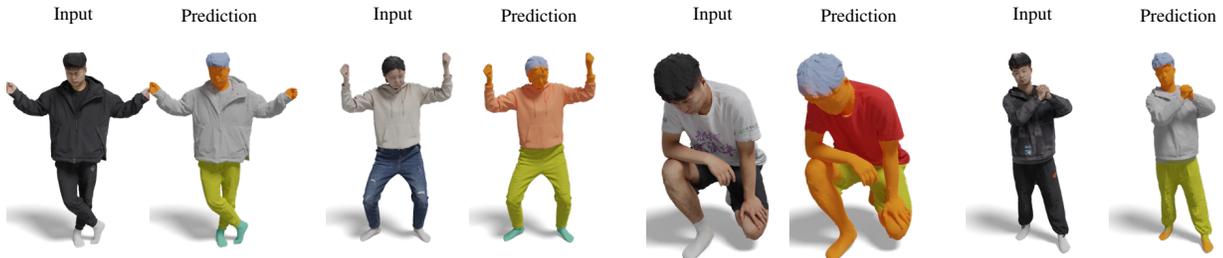
We fine-tune the model for 2 epochs only.

Table 8. Performance(IoU_{mean}) on CloSe-D-test after network refinement.

Layers trained	Naive loss	Weighted cross-entropy	Full
$f_{\text{dec}}\text{-last}$	90.33	90.37	90.25
$f_{\text{dec}}\text{-full}$	89.14	88.53	88.50
$f_{\text{dec}}\text{-last} + f_{\text{MLP}}$	90.62	90.53	90.33
$f_{\text{dec}}\text{-full} + f_{\text{MLP}}$	89.00	88.53	88.95
$f_{\text{dec}}\text{-last} + f_{\text{MLP}} + f^3$	90.53	90.18	90.35
$f_{\text{dec}}\text{-full} + f_{\text{MLP}} + f^3$	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 in Fig. 17.

Finally, we have generated high quality segmentation labels of approximately 1000 scans (from diverse sources [12, 20, 45, 59]) using CloSe-Net and CloSe-T. We will release this as CloSe-D++.



THuman3.0 Scans [45]



3DHumans Scans [20]



HuMMan Scans [12]



THuman2.0 Scans [59]

Figure 16. CloSe-Net results on real-world public datasets [12, 20, 45, 59].



Figure 17. 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 [45](top) and HuMMan [12](bottom). Fine-tuned networks result in consistent predictions.