Open-Vocabulary Semantic Part Segmentation of 3D Human

Keito Suzuki[*,1] Bang Du[*,1] Girish Krishnan[*] Kunyao Chen[†] Runfa Blark Li[*] Truong Nguyen[*]
[*]University of California, San Diego [†]Qualcomm
{k3suzuki, b7du, gikrishnan, kuc017, runfa, tqn001}@ucsd.edu

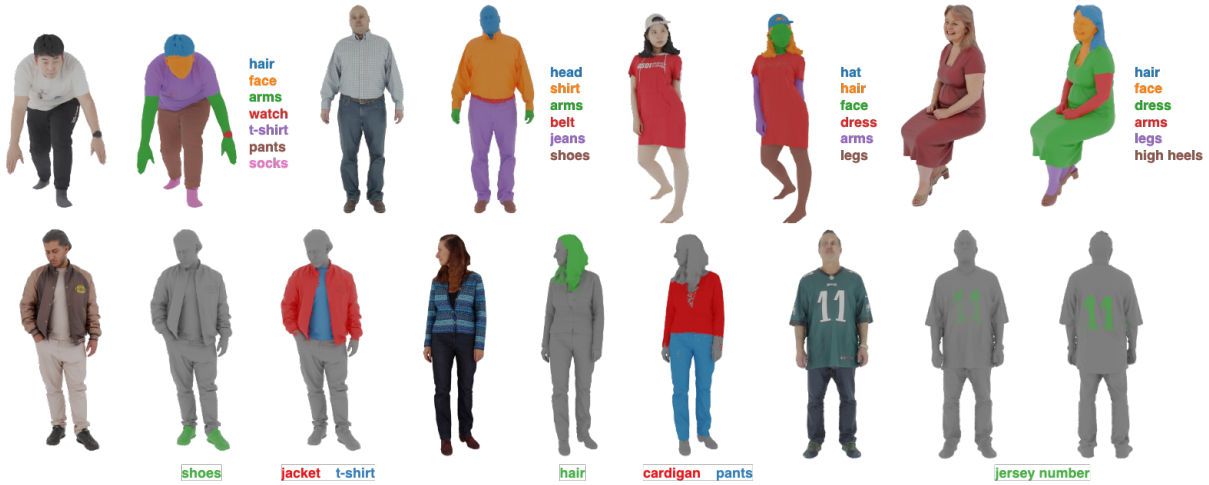


Figure 1. We propose the first open-vocabulary method for the segmentation of 3D human. It infers 3D segmentation by rendering multi-view images and leveraging pre-trained vision-language models. The figure displays the input text prompts and the corresponding segmentation results for 3D humans from various datasets. Our method supports arbitrary queries and generates non-overlapping masks in the 3D model. See Figure 7 and Figure 8 for more results.

Abstract

3D part segmentation is still an open problem in the field of 3D vision and AR/VR. Due to limited 3D labeled data, traditional supervised segmentation methods fall short in generalizing to unseen shapes and categories. Recently, the advancement in vision-language models' zero-shot abilities has brought a surge in open-world 3D segmentation methods. While these methods show promising results for 3D scenes or objects, they do not generalize well to 3D humans. In this paper, we present the first open-vocabulary segmentation method capable of handling 3D human. Our framework can segment the human category into desired fine-grained parts based on the textual prompt. We design a simple segmentation pipeline, leveraging SAM to generate multi-view proposals in 2D and proposing a novel HumanCLIP model to create unified embeddings for visual and textual inputs. Compared with existing pre-trained CLIP models, the HumanCLIP model yields more accurate embeddings for human-centric contents. We also design a simple-yet-effective MaskFusion module, which classifies and fuses multi-view features into 3D semantic masks without complex voting and grouping mechanisms. The design of decoupling mask proposals and text input also significantly boosts the efficiency of per-prompt inference. Experimental results on various 3D human datasets show that our method outperforms current state-of-the-art open-vocabulary 3D segmentation methods by a large margin. In addition, we show that our method can be directly applied to various 3D representations including meshes, point clouds, and 3D Gaussian Splatting.

1. Introduction

The advancements in 3D technologies have led to an increased demand for automated analysis of 3D shapes.

[1]Equal Contribution.

Among the related tasks, 3D part segmentation plays a pivotal role in supporting a wide spectrum of applications, including robotics and AR/VR.

With the introduction of deep neural networks [32, 33, 40, 46, 51] and labeled 3D datasets [6, 45], 3D part segmentation has seen remarkable progress in recent years through supervised training. Nonetheless, creating 3D datasets is expensive and time-consuming. Compared with image data, current 3D part-annotated datasets are orders of magnitude smaller in scale. Within the limited 3D data, the human category represents only a tiny fraction. Existing human parsing methods have been trained to segment clothed data [2, 3, 30] or underlying body parts [4, 39], but they fall short of generalizing to unseen models and classes. Thus, enabling machines with the ability to parse objects into semantic parts and generalize to new categories still remains difficult, especially for human-related data, which usually contain more complex geometry with richer semantic attributes than general 3D objects.

Recent developments in vision-language learning gave rise to many 2D image-based models [25, 34] with exceptional zero-shot generalization capabilities. Many works seek to transfer 2D knowledge to 3D through pre-trained image-language models. [1, 28, 50, 54, 55] leverage these models through multi-view rendering and aggregate the information in 3D for the final segmentation result. Another line [43] focuses on distilling the information for a better 3D model. While these methods have shown promising improvements in object data, they have not exhibited the same quality of results on 3D human data.

In this paper, we aim to bring the open-vocabulary 3D part segmentation performance to human data. We introduce the first framework for 3D human parsing that semantically segments whatever parts you want and supports various 3D representations, including meshes, point clouds, and 3D Gaussians [21]. Inspired by [9], we formulate the segmentation task as a mask classification problem. Firstly, we generate class-agnostic instance mask proposals on rendered images through a pre-trained 2D segmentation model, SAM [22]. Secondly, we propose a novel HumanCLIP model that encodes each mask into embeddings within the CLIP feature space. Compared to the vanilla CLIP model [34], HumanCLIP produces more accurate text-aligned embeddings for human-related cases, enhancing the precision of the final segmentation. Since mask proposals are class-agnostic and independent between views, we propose a novel MaskFusion module that simultaneously classifies the semantic labels given the text prompt and fuses the multi-view inconsistent masks to generate 3D semantic segmentation for the input. It decouples the mask proposal step from reliance on text prompts, thereby enhancing inference efficiency. In summary, our contributions mainly include:

- We introduce the first open-vocabulary framework that segments fine-grained parts for 3D humans.
- We present a HumanCLIP model capable of extracting discriminative CLIP embeddings for human-centric data.
- We propose a MaskFusion module that simultaneously classifies semantic labels and converts multi-view 2D proposals into 3D segmentation, which significantly enhances inference efficiency.
- Our framework shows state-of-the-art performance on five 3D human datasets and shows compatibility with various 3D representations including 3D Gaussian Splatting.

2. Related Works

2.1. 3D Human Part Segmentation

The 3D human part segmentation field is largely driven by the advancement of 3D neural networks as well as the labeled dataset. [19, 30, 42] train point cloud [32, 33, 40] or mesh segmentation networks [13, 17, 37] through direct supervision or leveraging the unclothed human parametric templates or physical simulation of garments. [39] curates synthetic data to boost the performance on body parts. Datasets such as SIZER [41], MGN [3], and CTD [8], provide coarse clothing labels from 3D scans of clothed humans, but having only three categories and less variations of poses limit their applications. [2] presents Close-D which consists of 18 garment categories. To the best of our knowledge, it is the most comprehensive dataset up-to-date. However, due to the still limited size, none of these methods show generalizable ability and cannot segment labels outside of the predefined taxonomy.

2.2. Open-Vocabulary 3D Segmentation

In recent years, large vision language models [20, 34] have grown popular due to their ability to perform zero-shot recognition. As a result, many works have incorporated these models [12, 14, 24, 26] to conduct open-vocabulary 2D segmentation. To transfer the knowledge into 3D, some methods [18, 50, 55] apply CLIP to depth maps for zero-shot object classification and segmentation. For scene-scale data, CLIP2Scene [7] and CLIP^2 [49] train an additional 3D encoder with a contrastive loss. Although these models have shown their effectiveness on general 3D scenes or objects, we find that they do not work well for 3D humans. As one of the most important categories, a framework tailored for open-vocabulary 3D human parsing is demanded.

CLIP Embeddings. CLIP [34] is one of the most widely used vision-language models in both 2D and 3D open-world segmentation approaches. Due to being trained with natural images, the vanilla CLIP does not perform well on special input subsets, such as masked images or fashion images. Many approaches try to adapt CLIP models to new tasks. [5, 11] presents fine-tuned CLIP on fashion data. [26] propose the Mask-adapted CLIP. AlphaCLIP [38] adds an ad-

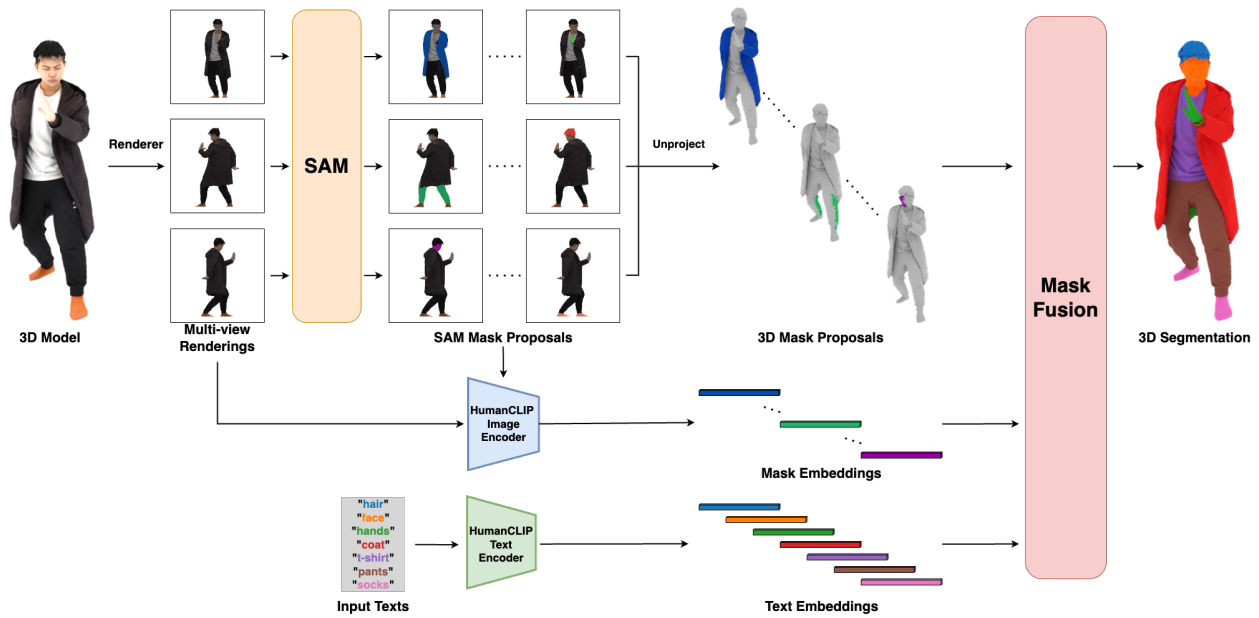


Figure 2. Overview of the proposed framework. Given a 3D human model, it is first rendered to obtain multi-view 2D images. The images are then fed to SAM to generate class-agnostic 2D masks and unprojected to obtain binary 3D masks. Additionally, each pair of image and 2D masks are fed to the human-centric mask-based text-aligned image encoder to obtain CLIP embeddings for each mask. Simultaneously, the input class texts are fed to the text encoder to obtain corresponding text embeddings. The 3D mask proposals, mask embeddings, and text embeddings are fed to the mask fusion module to obtain the final segmentation result.

ditional alpha channel as input so that it composes both regional and global information for better understanding. While AlphaCLIP provides better embeddings than CLIP for a region of interest, we observe that it is still inadequate to distinguish human body parts and garments.

3. Proposed Method

3.1. Overview

The overview of the proposed framework is shown in Figure 2. We assume a point-based 3D shape with size P as the input. Given the 3D human model and K semantic text prompts, the goal is to parse the 3D human into segments that semantically correspond to the input prompts.

Inspired by recent achievements in 2D and 3D segmentation methods [10, 12, 36, 39, 52], we formulate the semantic segmentation task as mask classification, originated from MaskFormer [9]. To bridge 3D data with 2D pre-trained models, we render the input from V predefined camera views. Segment-Anything-Model (SAM) [22] is leveraged to generate mask proposals for each view (Section 3.2). We introduce the novel HumanCLIP model, which encodes these proposals into embeddings within the unified CLIP feature space (Section 3.3). To lift 2D labels into 3D, it is usually required to assign "super points" and have carefully designed voting and grouping. In this work, we present a simple MaskFusion module. It takes HumanCLIP encoded text prompts and simultaneously performs classification and multi-view aggregation without the need for complex operations (Section 3.5). Note that the generation of mask proposals and embeddings is performed just once per model. Subsequently, segmentation can be executed in just a few milliseconds per prompt, significantly enhancing efficiency compared to previous methods.

3.2. Multi-view Mask Proposals

We choose SAM [22] to generate mask proposals on multi-view rendered images. SAM demonstrates remarkable zero-shot capabilities in image segmentation. From the predefined camera poses, we render the 3D human into V RGB images $I_i \in \mathbb{R}^{H \times W \times 3}$ where $i \in [1, V]$ is the index of the view. Each image I_i is then independently fed into SAM in the "segment everything" mode to generate N_i class-agnostic overlapping masks:

$$[m_{i,1}^{2D}, ..., m_{i,N_i}^{2D}] = SAM(I_i) \tag{1}$$

where $m_{i,j}^{2D}$ is the j-th 2D mask generated by SAM from the i-th view. This results in a total of $N = \sum_{i=1}^{V} N_i$ binary 2D masks at a varying granularity of whole, part, and subpart.

Each 2D mask $m_{i,j}^{2D}$ is then unprojected to 3D using the camera parameters of view i to construct the corresponding 3D proposal $m_{i,j}^{3D}$.

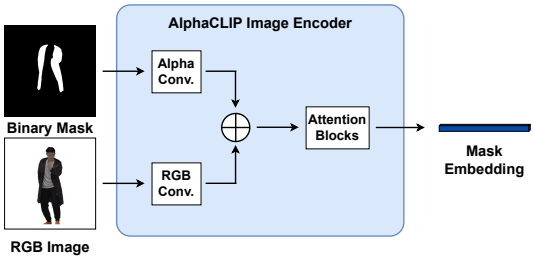


Figure 3. AlphaCLIP Image Encoder.

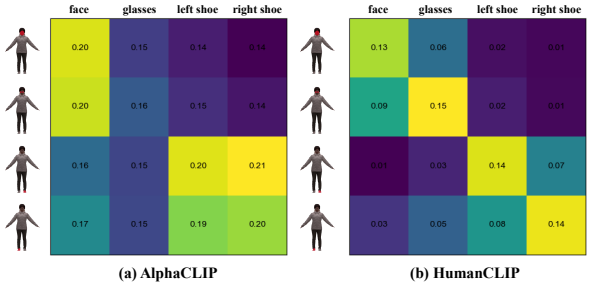


Figure 4. Comparison between (a) pre-trained AlphaCLIP and (b) the proposed HumanCLIP. The plots show the cosine similarity between the embedding of the masked region corresponding to *face*, *glasses*, *left shoe*, and *right shoe* and their text embeddings.

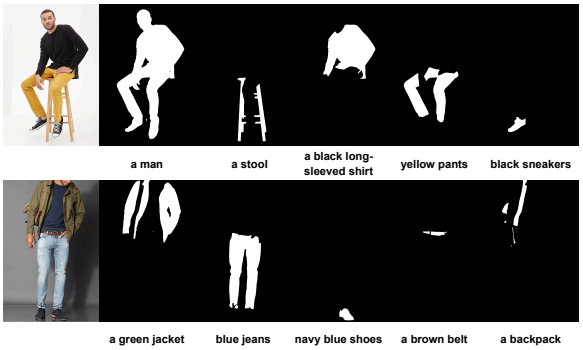


Figure 5. Example of mask-caption pairs generated by utilizing KOSMOS-2 and SAM.

3.3. HumanCLIP Encoding

We propose HumanCLIP to generate proposal embeddings with size C, where C is the embedding dimension of a CLIP model. The unified image-text feature space of CLIP allows the framework to perform open-vocabulary mask classification. Each 2D mask $m_{i,j}^{2D}$ with its corresponding image I_i is fed to the image encoder to get the proposal embedding $q_{i,j} \in \mathbb{R}^C$.

It is well-discussed that the vanilla pre-trained CLIP encoder does not perform well on specialty-formed inputs [26, 28], including masked and cropped images. Moreover, masking or cropping an image results in the loss of crucial contextual information, which is essential to the understanding of the specific area in an image. Therefore, we adopt the design of AlphaCLIP [38] to build our image encoder. As shown in Figure 3, the encoder accepts an additional alpha channel as input, which highlights the region of interest on the original rendered images. The input mask is processed with a parallel convolution layer to the RGB image and combined to go through a series of attention blocks to produce the final mask embedding in CLIP feature space. To further mitigate the domain gap, we finetune the encoder on a dataset of over 1.3 million RGBA region-text pairs with human-centric contents. We visualize the image-text alignment before and after fine-tuning in Figure 4. The pre-trained AlphaCLIP model fails to provide well-aligned embeddings for small parts such as the glasses as well as to distinguish left and right parts. The proposed HumanCLIP model generates more discriminative mask embeddings, facilitating the downstream classification tasks.

3.4. Region-Text Pair Generation

To tailor the image encoder for processing human-centric data, we finetune the model with region-text pair data. A straightforward method to acquire this data is utilizing 2D human segmentation datasets, where segmentation maps and category names directly form region-text pairs. Although efficient, this method yields less diverse masks and less informative captions. Therefore, we devise a pipeline to augment the training data. We source images from LIP [15], ATR [27], DeepFashion [29], and CIHP [16] datasets and employ KOSMOS-2 [31] and SAM [22] to automatically generate masks and corresponding captions for these images. An example of the generated pairs is depicted in Figure 5. Compared with the original labels, it provides more descriptive captions and introduces novel masks for objects that humans typically interact with, such as as 'a stool'. Further details of the data generation process are presented in the supplementary.

3.5. 3D Semantic Segmentation

To obtain the segmentation result with the desired semantic labels, our pipeline accepts K text prompts corresponding to the labels per inference. These texts are fed to the HumanCLIP text encoder to obtain CLIP text embeddings $\mathbf{W} \in \mathbb{R}^{K \times C}$. Then, the proposed MaskFusion module semantically classifies and synthesizes multi-view embeddings into 3D segmentation masks. Specifically, we utilize the correlations between the text embeddings and the mask embeddings to build correspondences and fuse the independent 3D masks.

Recap the N 3D mask proposals generated in Section. 3.2. The proposals and their embeddings are stacked to get $\mathbf{M} \in \mathbb{R}^{P \times N}$ and $\mathbf{Q} \in \mathbb{R}^{N \times C}$ respectively. We compute

Table 1. Statistics of region-text pairs used for HumanCLIP training. Each mask is accompanied by a descriptive caption.

Dataset	Images	Original Masks	Generated Masks	Total Masks
LIP	30462	173578	91505	265083
ATR	17706	175604	87698	263302
DeepFashion	12701	100632	43404	144036
CIHP	28280	647072	63616	710688
HumanCLIP	**89149**	**1096886**	**286223**	**1383109**

the classification logits $\mathbf{P} \in \mathbb{R}^{N \times K}$ by taking the cosine similarity between each mask embedding and each text embedding:

$$\mathbf{P}_{n,k} = \frac{\mathbf{Q}_n \cdot \mathbf{W}_k}{||\mathbf{Q}_n|| ||\mathbf{W}_k||} \tag{2}$$

It is used to guide the grouping of raw masks, which are class-agnostic and inconsistent across views.

In the final step, for each 3D point, we aggregate the class scores from the associated masks to get the final 3D segmentation result $\mathbf{Y} \in \mathbb{R}^{P \times K}$. Y is computed as the simple weighted average of 3D masks $\mathbf{M}$ based on the classification logits $\mathbf{P}$:

$$\mathbf{Y} = \mathbf{M} \times \mathbf{P} \tag{3}$$

We decouple the procedures for mask proposal and for text classification. Therefore, it is not guaranteed that each text input is valid for the 3D model. To ensure that only existed classes are segmented in the final result, we set a threshold τ on the final segmentation logits. If the maximum logits of a point fall below τ, the point is attributed to an 'other' class.

4. Experiments

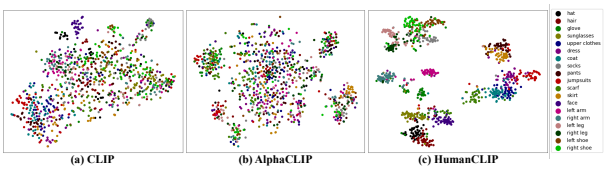
4.1. Implementation Details

Segment Anything Model. When applying SAM in our framework for both 3D segmentation and training data generation, we adopt the ViT-H model checkpoint. To create the mask proposals, we feed a rendered image of resolution 512×512 in the "segment everything" mode where we sample 64 points along each side of the image.

HumanCLIP. We initialize the HumanCLIP image encoder with the AlphaCLIP ViT-L/14 checkpoint, which is pre-trained on GRIT-20m dataset [31] with an image resolution of 224×224. The model is then finetuned on the curated HumanCLIP dataset, in which the images combines four 2D human parsing datasets: LIP [15], ATR [27], DeepFashion [29], and CIHP [16]. We utilize both the ground truth segmentation maps and the augmented region-text pairs described in Section 3.4, resulting in approximately 1.38 million RGBA-caption pairs for training. The distribution of images and masks across these datasets is detailed in Table 1. We keep the text encoder frozen and finetune the image encoder for a total of 3 epochs with a batch size of 18.

Table 2. Comparison of CLIP, AlphaCLIP, and HumanCLIP on mask classification accuracy of the LIP [15] and CCP [47] dataset.

Model	LIP	CCP
CLIP	22.12	21.75
AlphaCLIP	27.86	22.51
HumanCLIP	**79.98**	**52.96**



(a) CLIP (b) AlphaCLIP (c) HumanCLIP

Figure 6. Comparison of (a) CLIP, (b) AlphaCLIP and (c) HumanCLIP's t-SNE [44] visualizations of the mask embeddings for categories in the LIP dataset.

4.2. Effectiveness of HumanCLIP

Embedding Space. In Figure 6 we draw the t-SNE [44] projection of the mask embeddings extracted by CLIP, AlphaCLIP, and HumanCLIP on the LIP dataset. For CLIP and AlphaCLIP, significant overlap among embeddings of different categories is observed, complicating accurate class distinction based on text. In contrast, our proposed HumanCLIP forms more well-defined clusters for each class, enhancing the discriminativeness of the features.

Mask Classification. To further assess the effectiveness of the proposed HumanCLIP model in accurately embedding human parts, we conducted a mask classification task comparing it with the vanilla CLIP and pre-trained AlphaCLIP models. In this task, we first feed each ground truth image and mask to the image encoder to extract the mask embedding and then compute the cosine similarity with the text embedding for each class. For the CLIP models, the image segment cropped from the ground truth mask is input to the encoder to generate image embeddings. Classification is determined by the highest similarity score. The results, displayed in Table 2, show classification accuracy on the LIP [15] and CCP [47] datasets, which feature 19 and 54 classes respectively. The results indicate that both CLIP and AlphaCLIP models underperform, attributed to their lack of training on specific human parts data. AlphaCLIP shows slight improvements over CLIP as it provides better mask-wise embeddings. It is evident that the proposed HumanCLIP significantly outperforms both models in correctly classifying each mask based on the extracted embeddings.

4.3. Comparison with Open-Vocabulary 3D Segmentation Methods

Methods. To the best of our knowledge, there is no method dedicated to open-vocabulary 3D human parsing. Hence, we conduct comparisons with four general 3D segmenta-

Table 3. Comparison with open-set 3D segmentation methods. OA, mAcc, and mIoU are the overall accuracy, mean class accuracy, and mean Intersection over Union respectively. For each metric, a higher value is better. The best results are shown in **bold**.

Model	MGN			SIZER			CTD			THuman2.0			PosedPro			Average		
	OA	mAcc	mIoU	OA	mAcc	mIoU	OA	mAcc	mIoU	OA	mAcc	mIoU	OA	mAcc	mIoU	OA	mAcc	mIoU
PointCLIP V2	21.41	24.15	13.31	44.80	34.42	20.92	11.06	13.68	5.94	3.46	14.51	1.77	6.67	13.53	2.33	17.48	20.06	8.85
SATR	84.72	77.30	67.17	82.00	81.97	67.38	78.55	86.20	64.98	56.05	34.31	20.60	51.61	43.28	22.43	70.59	64.61	48.51
PartSLIP	90.03	86.53	78.63	84.94	82.18	70.79	75.11	71.20	55.46	77.46	44.94	33.70	70.38	34.61	24.80	74.58	63.89	52.68
PartSLIP++	91.41	87.98	81.14	86.63	83.49	72.93	80.73	76.06	62.36	82.00	49.82	38.96	70.80	35.07	24.99	82.31	66.48	56.08
Ours	**94.61**	**95.03**	**88.78**	**91.24**	**90.73**	**82.55**	**93.29**	**93.21**	**83.36**	**89.88**	**69.40**	**54.50**	**80.23**	**46.37**	**37.27**	**89.85**	**78.95**	**69.29**

Figure 7. Examples of promptable segmentation.

tion approaches: PointCLIP v2 [55], SATR [1], PartSLIP [28], and PartSLIP++ [54]. PointCLIP v2 applies CLIP to multi-view depth maps for zero-shot 3D classification, part segmentation, and object detection. SATR applies the GLIP model [25] to rendered images and aggregates the multi-view bounding predictions for each prompt to yield a segmented mesh. It shows capabilities under unclothed human setting. PartSLIP also applies GLIP but for low-shot point cloud segmentation. This is then enhanced by PartSLIP++ which incorporates SAM and an EM algorithm.

Datasets. For quantitative evaluation, we benchmark the models on five labeled 3D human datasets: MGN [3], SIZER [41], CTD [8], THuman2.0 [48], and Posed Pro [35]. These dataset contain a total of 3, 9, 12, 12, and 19 classes respectively.

Quantitative Comparison. In Table 3, we show the quantitative comparison of our model with the four open-set 3D segmentation methods. We first notice that PointCLIP v2 performs poorly across all of the datasets. Since they apply CLIP to rendered depth maps, which differs greatly from real-world images that it was originally trained on, PointCLIP v2 is unable to effectively transfer the zero-shot capabilities to 3D. Among the other three methods, performance is generally best on the MGN dataset, which has the fewest classes, and declines as the number of classes increases. Across all datasets, it is evident that our proposed framework outperforms by a large margin in all metrics.

Visual Comparison. In Figure 8, we show the visual comparison of various methods on the same five 3D human datasets. Akin to the numerical results, PointCLIPv2 fails to generate reasonable segmentation results. SATR and PartSLIP are able to get a coarse segmentation: the boundaries between various segments are unclear. PartSLIP++ shows improved boundaries but still struggles with specific areas like 'hair' and 'face'. In contrast, our method delivers the most precise segmentation results.

4.4. Promptable Segmentation

The design of our MaskFusion module allows users to segment whatever they want, as highlighted in Figure 7. Beyond conducting a full segmentation of the entire body, our framework can precisely segment only the user-specified items. It also effectively recognizes and accurately segments unseen categories, such as "uniform number" and "smartphone holder".

4.5. Run-time Efficiency

Rendering 3D data into multi-view images for processing has raised efficiency concerns in applications. Our approach, which decouples mask proposal generation from textual prompt processing, offers a significant efficiency advantage over previous methods. The mask embedding serves as an attribute of the 3D model, which can be generated beforehand. During the segmentation phase, only the text encoder and MaskFusion module are active. In contrast, for [1, 28, 54], the GLIP model relies on text prompts to generate bounding boxes and masks, requiring the entire pipeline to be executed for each segmentation attempt.

Table 4 displays the inference times for various methods when executed on a system equipped with a single 24GB RTX 4090 graphics card. We compare the inference time for a one-time only run, which executes all modules from start to finish, and the average time for 100 inferences of the same model, during which we reuse any pre-generated information where possible. For instance, we do not rerender multi-view images for the subsequent inferences. The results of our approach show a significant decrease in average cost as the number of inferences increases. The most time-consuming step in our framework is the "segment everything" mode of SAM. However, this is only necessary once for multiple text inputs, making our model efficient for subsequent inferences. In scenarios common to the open-vocabulary setting, where there is a fixed amount of 3D assets but varying information is required based on user

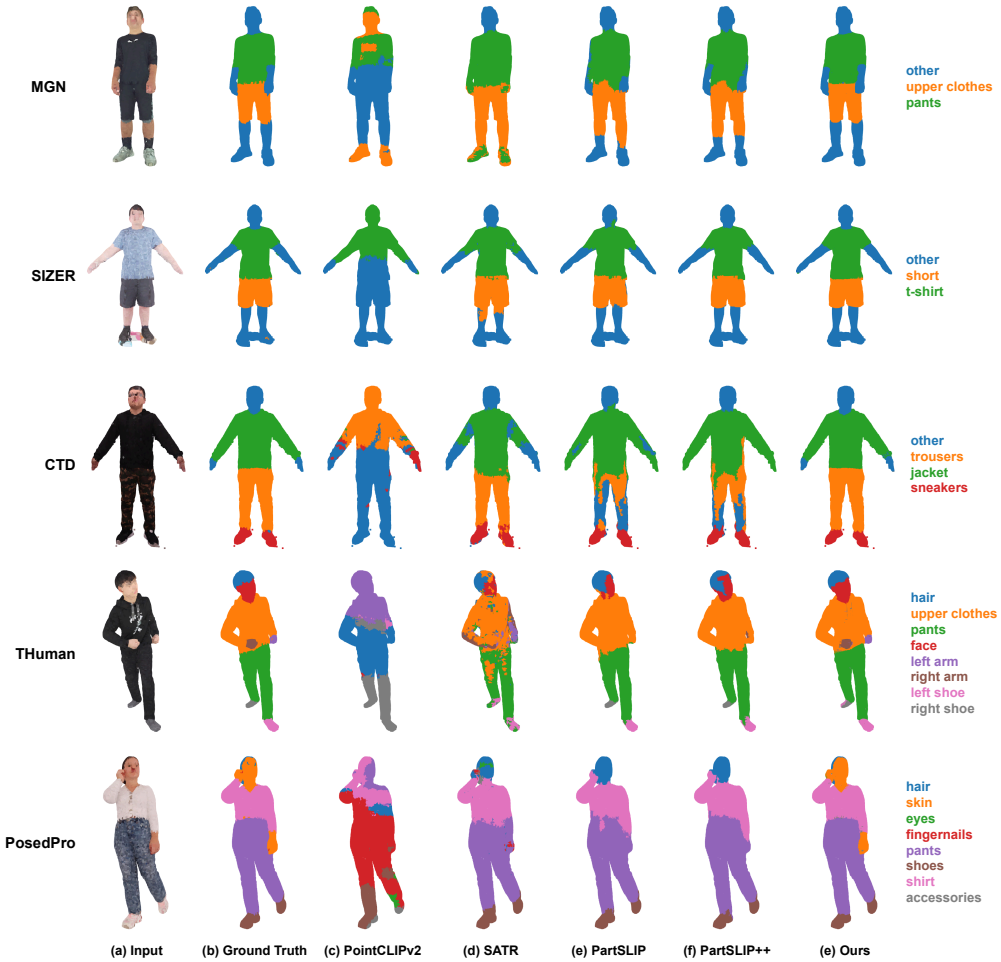


Figure 8. Qualitative analysis of segmentation results with (c) PointCLIPv2, (d) SATR, (e) PartSLIP, (f), PartSLIP++ on the 3D scans from MGN, SIZER, CTD, THuman, and PosedPro datasets.

Table 4. Comparison of inference times. We compare the average time cost in seconds assuming one-time inference only and 100 inference calls.

	PointCLIPv2	SATR	PartSLIP	PartSLIP++	Ours
One-time Inference	7.62	54.18	32.46	88.41	105.72
Average Inference	1.27	27.69	26.08	74.55	1.06

queries, our method offers considerable advantages.

4.6. 3D Gaussian Splatting Segmentation

Our framework design is compatible with various point-based 3D representations, including the popular 3D Gaussian Splatting [21] format. We follow the standard protocol to generate 3DGS for each model in the THuman2.0 dataset. As demonstrated in Figure 9 our method can generate reasonable segmentation results, making it a generalizable solution for segmenting 3DGS. This approach eliminates the need to optimize per-Gaussian semantic labels [23] or high-dimensional features [53] during the resource-intensive training stage.

Figure 9. Segmentation of 3D Gaussian Splatting.

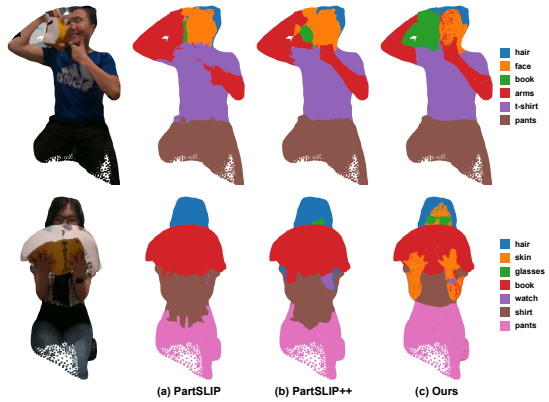


Figure 10. Visual comparison of (a) PartSLIP, (b) PartSLIP++, and (c) Ours on our in-the-wild point cloud dataset.

4.7. In-the-wild Segmentation

Our method demonstrates no significant domain gap in real-world noisy scenarios. Figure 10 shows a visual comparison with PartSLIP and PartSLIP++ on two point clouds captured by consumer-level RGB-D sensors, where the surface is at relatively low definition and the model is incomplete. In the first example, all methods are able to accurately segment the clothing, but our method shows the best segmentation ability of the book. For the second example, both PartSLIP and PartSLIP++ are unable to segment the 'skin' and 'glasses' categories. In contrast, our method can accurately segment these classes as well as the small area for the watch. More details are explained in the supplementary.

5. Ablation Study

HumanCLIP in Segmentation. Table 2 and Figure 6 present the advantage of HumanCLIP in extracting discriminative embeddings on human-related mask-caption data. We further ablate the module within our framework to assess its contribution to overall performance. Figure 11 illustrates the segmentation results when replacing the HumanCLIP with pre-trained AlphaCLIP model. In the example on the left, while AlphaCLIP accurately segments areas such as the jacket, pants, and hands, it struggles to distinguish the inner layer of clothing. In the example on the right, AlphaCLIP results in noisy segmentation across the legs and the right side of the body. Conversely, using HumanCLIP enables precise segmentation of body parts and clothing, as well as neighboring objects like a binder.

Number of views. To evaluate how the number of views affects the segmentation quality, we increase the number of views from 2 to 16 and compare the performance on the CTD dataset. The results are shown in Table 5. We observe that as we increase the number of views, the segmentation quality improves as it can better mitigate noisy mask proposals from affecting the final result. However, more views also require more time to preprocess, so we select 8 views to strike a good balance between quality and efficiency. It's important to note that all comparison methods utilize more than 8 views. The difference in the number of views does not confer an advantage to our method.

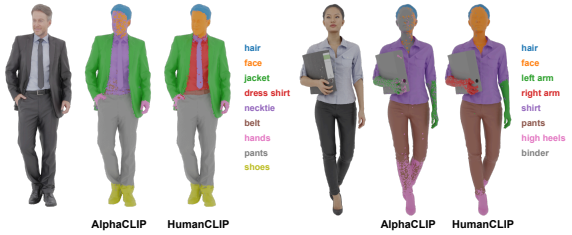


Figure 11. Visual comparison with AlphaCLIP in 3D human segmentation on the RenderPeople dataset.

Table 5. Effect of the number of views on the segmentation quality.

Metrics	Number of Views				
	2	4	8	10	16
Accuracy	91.48	92.31	93.29	93.44	93.71
mAcc.	91.43	92.09	93.21	93.49	93.61
mIoU.	80.49	81.43	83.36	83.87	84.24

Limitations. One limitation of this method is the slow runtime for a single inference caused by applying SAM to every view. This can make it difficult to apply our method to dynamic 3D humans. Another limitation is that we have to manually adjust the threshold to conduct promptable segmentation for different text inputs. In future works, we plan on applying our method to generate labeled data to train a model that can efficiently compute in 3D space.

6. Conclusion

In this paper, we present the first open-vocabulary method for 3D human segmentation. We introduce a novel HumanCLIP model and a MaskFusion module, which efficiently transfer the knowledge from 2D pre-trained vision-language models to the segmentation of 3D human data. Our method can seamlessly conduct semantic segmentation based on arbitrary user-defined text queries. The experimental results show that our method outperforms existing open-vocabulary 3D segmentation methods on five 3D human datasets. Additionally, we show that our method can be directly applied to various 3D representations including points clouds, meshes, and 3D Gaussian Splatting.

Acknowledgments. This research was supported by the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program in part (P0019797).

References

[1] Ahmed Abdelreheem, Ivan Skorokhodov, Maks Ovsjanikov, and Peter Wonka. Satr: Zero-shot semantic segmentation of 3d shapes. *arXiv preprint arXiv:2304.04909*, 2023. 2, 6

[2] Dimitrije Antić, Garvita Tiwari, Batuhan Ozcomlekci, Riccardo Marin, and Gerard Pons-Moll. Close: A 3d clothing segmentation dataset and model. In *International Conference on 3D Vision (3DV)*, 2024. 2

[3] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5420–5430, 2019. 2, 6

[4] Federica Bogo, Javier Romero, Matthew Loper, and Michael J Black. Faust: Dataset and evaluation for 3d mesh registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3794–3801, 2014. 2

[5] Giuseppe Cartella, Alberto Baldrati, Davide Morelli, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Openfashionclip: Vision-and-language contrastive learning with open-source fashion data. In *International Conference on Image Analysis and Processing*, pages 245–256. Springer, 2023. 2

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[7] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023. 2

[8] Xin Chen, Anqi Pang, Wei Yang, Peihao Wang, Lan Xu, and Jingyi Yu. Tightcap: 3d human shape capture with clothing tightness field. *ACM Transactions on Graphics (TOG)*, 41 (1):1–17, 2021. 2, 6

[9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2, 3

[10] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 3

[11] Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. Contrastive language and vision learning of general fashion concepts. *Scientific Reports*, 12(1):18958, 2022. 2

[12] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 2, 3

[13] Qiujie Dong, Zixiong Wang, Manyi Li, Junjie Gao, Shuangmin Chen, Zhenyu Shu, Shiqing Xin, Changhe Tu, and Wenping Wang. Laplacian2mesh: Laplacian-based mesh understanding. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 2

[14] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 2

[15] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 932–940, 2017. 4, 5

[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)*, pages 770–785, 2018. 4, 5

[17] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: a network with an edge. *ACM Transactions on Graphics (ToG)*, 38(4):1–12, 2019. 2

[18] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. *arXiv preprint arXiv:2210.01055*, 2022. 2

[19] Andrej Jertec, David Bojanić, Kristijan Bartol, Tomislav Pribanić, Tomislav Petković, and Slavenka Petrak. On using pointnet architecture for human body segmentation. In *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 253–257. IEEE, 2019. 2

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 7

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 3, 4

[23] Kun Lan, Haoran Li, Haolin Shi, Wenjun Wu, Yong Liao, Lin Wang, and Pengyuan Zhou. 2d-guided 3d gaussian segmentation. *arXiv preprint arXiv:2312.16047*, 2023. 7

[24] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 2

[25] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu

Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2, 6

[26] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 2, 4

[27] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep human parsing with active template regression. *IEEE transactions on pattern analysis and machine intelligence*, 37(12):2402–2414, 2015. 4, 5

[28] Minghua Liu, Yinhao Zhu, Hong Cai, Shizhong Han, Zhan Ling, Fatih Porikli, and Hao Su. Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21736–21746, 2023. 2, 4, 6

[29] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 4, 5

[30] 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. 2

[31] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 4, 5

[32] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2

[33] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[35] Renderpeople, 2023. https://renderpeople.com/3d-people. 6

[36] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 3

[37] Nicholas Sharp, Souhaib Attaiki, Keenan Crane, and Maks Ovsjanikov. Diffusionnet: Discretization agnostic learning on surfaces. *ACM Transactions on Graphics (TOG)*, 41(3): 1–16, 2022. 2

[38] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13019–13029, 2024. 2, 4

[39] Ayça Takmaz, Jonas Schult, Irem Kaftan, Mertcan Akçay, Bastian Leibe, Robert Sumner, Francis Engelmann, and Siyu Tang. 3d segmentation of humans in point clouds with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1292–1304, 2023. 2, 3

[40] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 2

[41] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 1–18. Springer, 2020. 2, 6

[42] Takuma Ueshima, Katsuya Hotta, Shogo Tokai, and Chao Zhang. Training pointnet for human point cloud segmentation with 3d meshes. In *Fifteenth International Conference on Quality Control by Artificial Vision*, pages 72–77. SPIE, 2021. 2

[43] Ardian Umam, Cheng-Kun Yang, Min-Hung Chen, Jen-Hui Chuang, and Yen-Yu Lin. Partdistill: 3d shape part segmentation by vision-language model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2024. 2

[44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 5

[45] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 2

[46] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3173–3182, 2021. 2

[47] Wei Yang, Ping Luo, and Liang Lin. Clothing co-parsing by joint image segmentation and labeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3182–3189, 2014. 5

[48] 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. 6

[49] Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang,

Xiaodan Liang, and Hang Xu. Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15244–15253, 2023. 2

[50] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022. 2

[51] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 2

[52] Qiang Zhou, Yuang Liu, Chaohui Yu, Jingliang Li, Zhibin Wang, and Fan Wang. Lmseg: Language-guided multi-dataset segmentation. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[53] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024. 7

[54] Yuchen Zhou, Jiayuan Gu, Xuanlin Li, Minghua Liu, Yunhao Fang, and Hao Su. Partslip++: Enhancing low-shot 3d part segmentation via multi-view instance segmentation and maximum likelihood estimation. *arXiv preprint arXiv:2312.03015*, 2023. 2, 6

[55] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2639–2650, 2023. 2, 6