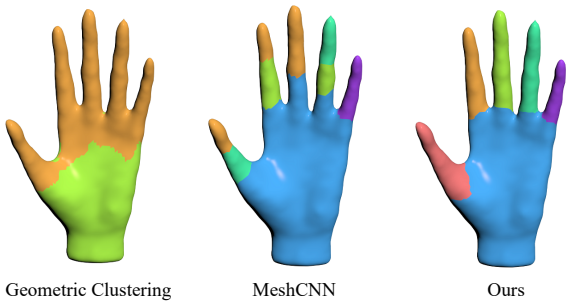
A Text-Driven Attentive Approach for 3D Shape Segmentation

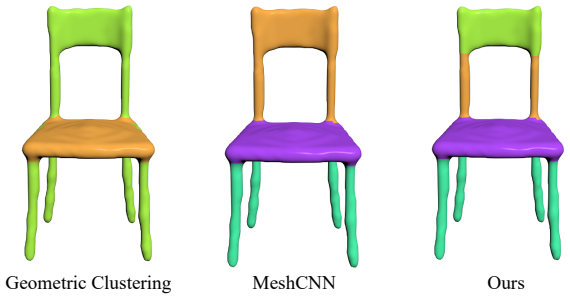
Zhenyu Shu, Chenyu Zhu*, and Shiqing Xin

***Abstract*—3D shape segmentation is an essential task in computer graphics and is widely used in many applications. It plays a critical role in understanding the structure and semantics of 3D models. Traditional approaches to 3D shape segmentation primarily rely on geometric features to partition models into meaningful components. However, these methods often struggle when the geometric characteristics of different parts are similar, resulting in ambiguous segmentation outcomes. To address this fundamental limitation, we introduce a novel text-driven multimodal framework that systematically integrates textual semantics with geometric analysis for enhanced 3D shape segmentation. Our approach leverages a pre-trained language model with prefix tuning to bridge the semantic granularity gap between part-level annotations and face-level segmentation, while a specialized mesh self-attention module captures contextual relationships among neighboring faces. We design an attention-based text-driven integration mechanism that dynamically weights multimodal features, complemented by a Laplace-Adaptive Attention Module (LAAM) that better handles the distributions of geometric features. Through contrastive learning, we align textual and geometric representations in a shared semantic space, enabling effective disambiguation of geometrically similar but semantically distinct parts. We also contribute the Fine-grained HumanBody benchmark for comprehensive evaluation. Extensive experiments on Princeton Segmentation Benchmark, COSEG, ShapeNetCore, and our proposed benchmark demonstrate that our method significantly outperforms existing approaches, achieving superior segmentation accuracy while effectively resolving geometric ambiguities through semantic understanding.**

***Index Terms*—3D Shape Segmentation, Attention Mechanism, Contrastive Learning**



(a) Results of different methods for a hand model



(b) Results of different methods for a chair model

Fig. 1. Comparative segmentation results on hand and chair models. From left to right: (1) geometric feature clustering, (2) traditional geometry-only method (MeshCNN [4]), and (3) our text-driven approach. Traditional methods struggle with geometrically similar parts, while our multimodal framework achieves precise segmentation by leveraging semantic textual information.

I. Introduction

3D shape segmentation represents a cornerstone technique in computer vision and computer graphics, with far-reaching applications spanning medical image analysis [1], [2], computer-aided design [3], robotics, and digital content creation. The fundamental goal of 3D segmentation is to decompose complex three-dimensional shapes into semantically meaningful sub-parts characterized by distinct geometric or functional properties, thereby enabling sophisticated downstream applications, including shape analysis, editing, synthesis, and understanding. This decomposition process significantly enhances the comprehension of inherent shape attributes and structural relationships, making it indispensable for numerous practical applications.

Despite decades of research progress, 3D shape segmentation remains fundamentally challenging due to several inherent complexities. First, the intrinsic diversity and semantic ambiguity of model components necessitate consistent and meaningful label assignment across varying geometric contexts. Second, accurate identification of part boundaries often depends on subtle geometric cues that require sophisticated integration of multi-scale local and global features. Third, and most critically, many real-world objects contain parts with highly similar geometric characteristics but distinct semantic meanings, a scenario where purely geometry-based approaches frequently fail to achieve satisfactory results.

Traditional 3D shape segmentation methodologies [5]–[10] typically follow a well-established three-stage pipeline: (1) feature extraction, where each mesh face is represented by concatenated hand-crafted geometric descriptors; (2) clustering or classification, where machine learning algorithms operate

Zhenyu Shu is with School of Computer and Data Engineering, NingboTech University, Ningbo, PR China.
E-mail: shuzhenyu@nit.zju.edu.cn (Zhenyu Shu)
Chenyu Zhu is with College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China. Corresponding author.
E-mail: chenyuzhu_paper@163.com (Chenyu Zhu)
Shiqing Xin is with School of Computer Science and Technology, ShanDong University, Jinan, PR China.
Manuscript received month day, year; revised month day, year.

on these feature vectors to assign semantic labels; and (3) label projection, where computed labels are mapped back to the original 3D mesh surfaces. While this paradigm has achieved reasonable success, it suffers from fundamental limitations that constrain its effectiveness in complex scenarios.

The primary limitation of existing approaches lies in their exclusive reliance on hand-crafted geometric descriptors, which predominantly capture local shape characteristics while struggling to encode global contextual information effectively. This constraint becomes particularly problematic when handling shapes with complex topologies, noise artifacts, or incomplete geometric data. More critically, purely geometric approaches fail catastrophically when confronted with parts exhibiting similar geometric properties but serving different functional or semantic roles, such as distinguishing between individual fingers on a hand or differentiating chair legs from chair backs in certain orientations.

To address these fundamental limitations, we introduce a novel text-driven multimodal framework for 3D shape segmentation that represents the systematic integration of textual semantics with geometric analysis in this paper. Our core insight is that textual descriptions provide complementary semantic information that can effectively disambiguate geometrically similar parts, thereby overcoming the primary weakness of existing purely geometric approaches. As illustrated in Figure 1, our method demonstrates superior performance compared to traditional geometry-only approaches, particularly in challenging scenarios involving geometrically similar but semantically distinct parts. By incorporating textual information that describes part semantics (e.g., "finger of a hand," "leg of a chair"), our framework can successfully distinguish between parts that confound traditional methods.

Our proposed framework introduces several novel technical contributions that enable effective multimodal learning for 3D segmentation. We employ a Prefix-Tuning mechanism [11] that addresses the fundamental mismatch between part-level textual annotations and face-level geometric segmentation requirements. By introducing trainable prefix vectors that enhance textual feature diversity while preserving pre-trained linguistic knowledge, we bridge the semantic granularity gap that has hindered previous multimodal approaches. We design a specialized Mesh Self-Attention module that captures spatial dependencies between neighboring mesh faces, overcoming the limitation of traditional geometric descriptors that operate in isolation. Additionally, we propose a sophisticated text-driven integration module that employs attention mechanisms to dynamically weight the contribution of textual features based on their relevance to local geometric contexts. Furthermore, we introduce a novel Laplace-Adaptive Attention Module (LAAM) based on Laplace distributions that better captures the heavy-tailed and sparse nature of geometric features compared to traditional Gaussian-based approaches, significantly improving the model's ability to focus on discriminative features while suppressing noise.

Our research makes several significant contributions to the field of 3D shape analysis:

- We propose a novel dataset, the Fine-grained HumanBody benchmark, for benchmarking text-driven 3D shape segmentation methods, which builds upon the original Human Body benchmark. Our benchmark consists of 100 3D models, where each 3D model was segmented into 12 distinct parts, incorporating more comprehensive structural details to facilitate detailed analysis.
- We present a novel text-driven framework for 3D model segmentation that systematically addresses the challenges of multimodal integration through four innovative components: Prefix Tuning mechanism for bridging semantic granularity gaps, Mesh Self-Attention Module for capturing geometric context, Text-Driven Module for adaptive multimodal fusion, and Laplace-Adaptive Attention Module for distribution-aware feature weighting. Through contrastive learning alignment, the proposed approach enables effective disambiguation of geometrically ambiguous parts while significantly improving segmentation accuracy.
- Comprehensive experimental results from public benchmark datasets show that our proposed method surpasses traditional 3D segmentation techniques in semantic segmentation, demonstrating our text-driven framework's effectiveness.

The remainder of this paper is structured as follows: Section II provides a comprehensive review of related work in 3D shape segmentation, attention mechanisms, and multimodal learning. Section III presents our proposed text-driven framework with detailed explanations of each innovative component. Section IV reports extensive experimental results, including comparative analyses, ablation studies, and performance evaluation across multiple benchmark datasets. Finally, Section VI concludes the paper with a discussion of implications and future research directions.

II. Related Work

This section reviews related work on our proposed method, including attention mechanisms commonly employed in computer vision, traditional 3D shape segmentation methods, and deep learning-based approaches.

A. Attention mechanism

The human visual system selectively focuses on critical elements for efficient analysis of complex scenes. Inspired by this capability, researchers have incorporated attention mechanisms into computer vision to enhance performance. The attention mechanism functions as a dynamic selection process, adaptively assigning different weights to features based on their relevance.

Channel Attention. Squeeze-and-Excitation Networks (SENet) [12] introduced an effective channel-wise attention mechanism utilizing global pooling and fully connected layers. This approach enables SENet to learn the significance of each feature channel, enhancing useful features while diminishing less relevant ones. Building on this, Efficient Channel Attention (ECA-Net) [13] refines the SENet module by proposing a local cross-channel interaction strategy without dimensionality reduction, adaptively determining the appropriate size for one-dimensional convolution kernels.

Spatial Attention. GE-Net [14] utilizes deep convolutions to encode spatial information, effectively capturing contextual relationships between features. Double Attention Networks [15] collect critical features from the entire spatial domain into a compact set and adaptively distribute them to each position. Global-Context Networks (GC-Net) [16] propose a query-independent simplified structure to reduce parameters and computations. CCNet [17] determines connections between target feature pixels and every other point in the feature map, adjusting the importance of target pixel features.

B. Traditional 3D Shape Segmentation

Early 3D shape segmentation research focused on extracting significant features from 3D shapes using advanced algorithms. Traditional methods relied on geometric features such as surface normals and curvature to capture local and global shape characteristics. Traditional machine learning techniques, such as Support Vector Machines and Random Forests, were utilized to classify different regions [18], [19].

Boundary-based methods. These methods identify boundaries between different parts by detecting boundary features such as curvature and normal vectors on the model's surface. While practical for models with distinct geometric features, they are noise-sensitive. Katz et al. [20] proposed a hierarchical segmentation approach based on fitting primitives to the surface. Hachani et al. [21] presented an implicit segmentation approach using Morse theory and curvature analysis with Shape Diameter Function (SDF). Additionally, Yamauchi et al. [22] developed mesh segmentation techniques based on pose-invariant boundary detection.

Region-growing-based methods. These methods use clustering or region-growing algorithms to compute similarity measures between adjacent regions, considering features such as color, texture, or shape. Points or faces with similar features are grouped into the same category. Although capable of handling complex topological structures, they are often computationally intensive. Attene et al. [23] proposed a hierarchical mesh segmentation approach based on fitting primitives. Vieira and Shimada [24] presented a surface mesh segmentation method using region growing with adaptive thresholds.

Graph-based methods. These approaches represent 3D models as graph structures, where vertices, edges, or faces correspond to graph nodes and edges. Segmentation is achieved by minimizing energy functions using clustering or partitioning algorithms such as graph cuts or minimum spanning trees. Golovinskiy and Funkhouser [19] proposed a randomized cuts algorithm for mesh segmentation. Random walk algorithms model segmentation as a random walk process from seed points.

C. Deep Learning on 3D Shape Segmentation

With rapid deep learning advancement, there has been increasing emphasis on deep learning-based 3D shape segmentation methods [25], [26]. These methods leverage powerful architectures including CNNs [27], PointNet [28], and Transformer-based models [29] to automatically extract high-level geometric features from point clouds, meshes, or volumetric data, significantly improving segmentation accuracy and generalization.

Guo et al. [30] and Yu et al. [31] effectively utilized Transformers and Recurrent Neural Networks for point labeling on 3D surfaces. MeshCNN [4] introduced specialized neural network architecture for meshes, employing edge convolution and edge pooling operations to maintain high resolution while reducing computational complexity. Lahav et al. [32] applied random walk algorithms to analyze global structure and local characteristics. Milano et al. [33] combined convolution operations on primal and dual meshes with new distance metrics. HodgeNet [34] computes lower-order eigendecompositions using sparse differential operators, while Laplacian2Mesh [35] transforms meshes into the Laplacian-Beltrami spectral domain.

Recent advances in vision-language models have catalyzed progress in multimodal 3D shape understanding. PointCLIP [36] adapted CLIP for point cloud understanding through multi-view depth map projections, enabling zero-shot and few-shot classification. Point-BERT [37] introduced masked point modeling with transformers for self-supervised pre-training, while ULIP [38] established contrastive alignment across image, text, and point cloud modalities for unified representations. OpenShape [39] further advanced this paradigm by training on large-scale multi-modal datasets for generalizable shape-text-image representations. However, these methods primarily focus on object-level classification or part-level point cloud labeling rather than fine-grained face-level mesh segmentation, and rely on simple projection or alignment strategies that do not explicitly address the semantic granularity mismatch between part-level annotations and face-level features—a fundamental challenge our work systematically addresses.

Most geometry-based methods rely on features such as face descriptors and edge dihedral angles. However, multi-view projection methods address segmentation by establishing mappings between 3D shapes and projected images. Wang et al. [40] and Kalogerakis et al. [41] employed projection matrices to project 3D shapes onto 2D images for image-based segmentation. PartSLIP [42] organizes point clouds by clustering, creates bounding boxes from multiple perspectives using GLIP, and evaluates clusters by scoring visible points.

To overcome limitations of existing geometry-only methods, we propose a novel text-driven 3D shape segmentation framework. This framework uses 3D model data with part captions as input, effectively extracts geometric and text features, and aligns them in semantic space through contrastive learning. By dynamically considering feature distributions, our method achieves superior segmentation performance.

III. METHOD

A. Overview

Our method introduces a novel multimodal approach that addresses the fundamental limitation of traditional 3D shape segmentation methods: their exclusive reliance on geometric

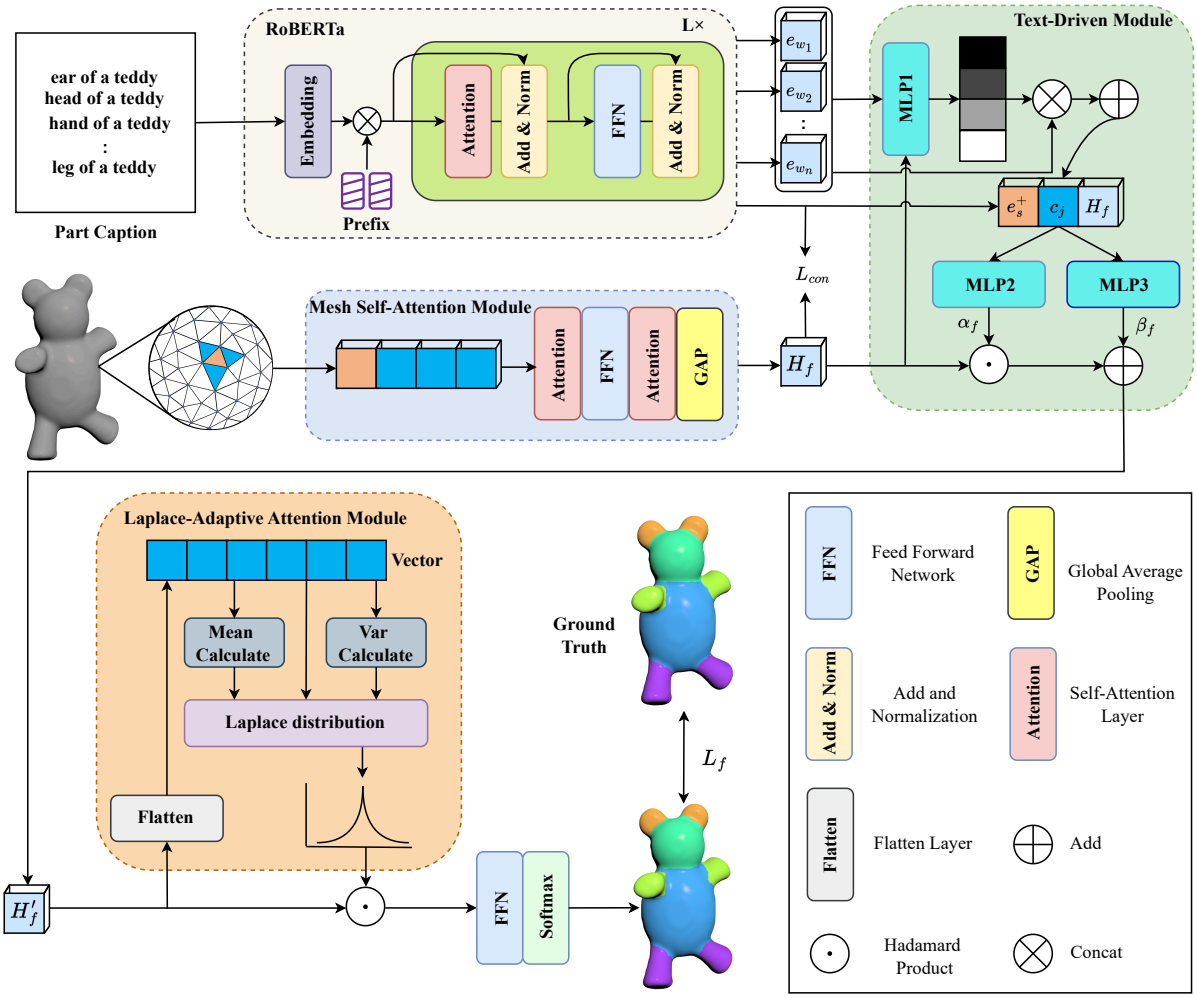


Fig. 2. Our segmentation framework consists of four main components: i) Prefix Tuning: The Prefix Tuning method extracts sentence and word-level features. ii) Mesh Self-Attention Module: This module incorporates additional contextual information by employing an attention mechanism that aggregates features from the central face and its adjacent faces. iii) Text-Driven Module: This module integrates sentence and word features with geometric features to facilitate 3D model segmentation. iv) Laplace-Adaptive Attention Module: This module dynamically adjusts the attention focus by considering the distribution of the input features. The primary network is trained using a combination of cross-entropy and contrastive loss, aligning its output with the ground-truth segmentation labels of fully annotated shapes. This training process ensures that the network learns to generate accurate segmentation predictions.

features. When geometric characteristics between different parts are similar, traditional methods often produce ambiguous segmentation results. To overcome this challenge, we propose a text-driven framework that leverages the complementary nature of textual descriptions to enhance geometric understanding.

The framework is trained on a 3D shape dataset augmented with part-level textual captions, following a carefully designed forward propagation sequence. As illustrated in Figure 2, our architecture comprises four innovative modules that work synergistically: (1) **Prefix Tuning** extracts enriched sentence and word features while bridging the semantic gap between part-level captions and face-level segmentation through learnable prefixes; (2) **Mesh Self-Attention Module** captures contextual relationships among neighboring faces using a spatial attention mechanism; (3) **Text-Driven Module** seamlessly integrates textual and geometric features through attention-based fusion; and (4) **Laplace-Adaptive Attention Module** dynamically adjusts feature importance based on the statistical distribution of input data.

B. Prefix Tuning: Bridging Semantic Granularity Gap

A critical challenge in text-driven 3D segmentation arises from the granularity mismatch: 3D models are annotated with part-level captions, while segmentation must be performed at the face level. This discrepancy creates alignment inconsistencies that can severely impact performance. Traditional approaches that directly map part captions to face features fail to capture the subtle variations within semantic parts.

To address this fundamental issue, we introduce an innovative Prefix Tuning mechanism, as illustrated in Figures 2 and 3. Our approach leverages a pre-trained RoBERTa model [43] as the backbone for textual feature extraction, with its parameters frozen to preserve pre-trained linguistic knowledge.

During forward propagation, the model processes both the part caption and the learned prefix simultaneously:

$$H^j = [E(x), Prefix^j], \tag{1}$$

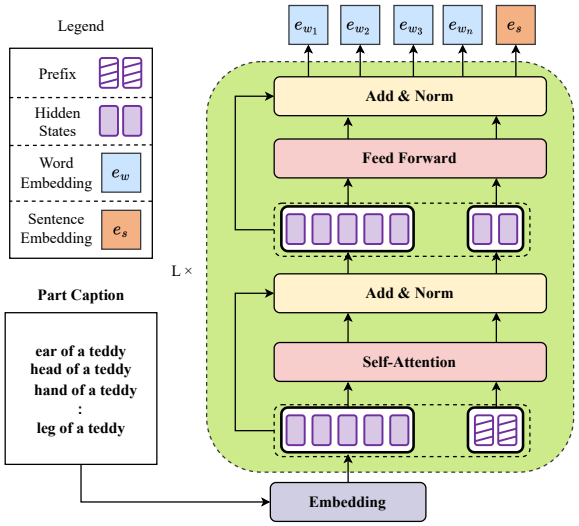


Fig. 3. Our innovative Prefix Tuning approach addresses the fundamental mismatch between part-level textual annotations and face-level geometric features. By incorporating trainable prefixes into the pre-trained RoBERTa model, we enable fine-grained distinction between faces within the same semantic part while maintaining semantic consistency.

where $E(\cdot)$ represents the embedding function, x denotes the input part caption, [,] represents the concatenation operation, and $Prefix \in \mathbb{R}^{p \times d}$ refers to the trainable prefix matrix. The superscript j indicates the j-th layer, p denotes the number of continuous prefix vectors, and d represents the hidden state dimensionality.

The trainable prefixes serve as learnable context vectors that enable the model to differentiate between faces belonging to the same semantic part. Unlike traditional fine-tuning approaches that modify the entire pre-trained model, our prefix tuning preserves the linguistic representations while introducing task-specific adaptations. This design ensures consistency between training and inference phases by maintaining fixed positional encodings and segment embeddings.

Through the self-attention mechanism inherent in RoBERTa, the model captures complex relationships between words and learns to associate different prefix patterns with distinct geometric variations within the same semantic part. At the final layer, we extract the global sentence representation e_s from the [**CLS**] token and individual word vectors e_{w_i}, where $i \in \{1, n\}$ represents the word index and n denotes the total number of words in the textual description.

C. Mesh Self-Attention Module: Contextual Geometric Understanding

Traditional geometric feature descriptors operate independently on each face, failing to capture the contextual relationships that are crucial for accurate segmentation. To address this limitation, we design a Mesh Self-Attention Module that explicitly models the spatial dependencies between neighboring faces.

We employ five complementary geometric descriptors to characterize each face: Average Geodesic Distance (AGD) [44], Gaussian Curvature (GC) [45], Shape Diameter Function (SDF) [46], Scale-Invariant Heat Kernel Signature (SIHKS) [47], and Wave Kernel Signature (WKS) [48]. These descriptors capture different geometric properties: AGD measures geodesic distances, GC captures local curvature, SDF represents thickness, while SIHKS and WKS provide multi-scale geometric signatures. The resulting feature vectors have dimensions of 1, 1, 1, 19, and 100, respectively, which are concatenated to form a comprehensive 122-dimensional geometric descriptor v.

Our Mesh Self-Attention Module processes a central face along with its three neighboring faces, forming a local contextual neighborhood. The input consists of the central face feature v_0 and three neighboring face features v_1, v_2, and v_3, concatenated into a $122 \times 1 \times 4$ feature tensor. The self-attention mechanism computes contextual relationships through:

$$Q_i = v_i W^Q, \quad K_i = v_i W^K, \quad V_i = v_i W^V, \tag{2}$$

$$a_{ij} = \frac{\exp(Q_i K_j^T)}{\sum_{k=1}^{n} \exp(Q_i K_k^T)}, \tag{3}$$

$$v_i' = \sum_{j=1}^{n} a_{ij} V_j, \tag{4}$$

where $i \in \{0, 1, 2, 3\}$ indexes the faces in the local neighborhood. The attention weights a_{ij} capture the contextual importance of each neighboring face relative to the central face.

After obtaining attention-weighted features, we apply global average pooling to aggregate the local contextual information:

$$H_f = F_{pooling}(v_0', v_1', v_2', v_3'). \tag{5}$$

The resulting H_f represents a "neighborhood fused feature" that incorporates contextual information from the local neighborhood, enabling the model to understand each face within its spatial context rather than in isolation.

We employ contrastive learning to ensure semantic consistency between geometric and textual features. For each face, we encourage its global feature H_f to be similar to the corresponding part caption's sentence embedding e_s^+ while being dissimilar to non-corresponding captions:

$$L_{\text{con}}^i(H_{f_i}, e_s^+) = -\log\left(\frac{\exp(\cos(H_{f_i}, e_s^+)/\tau)}{\sum_{j=1}^{m} \exp(\cos(H_{f_i}, e_s^j)/\tau)}\right), \tag{6}$$

$$L_{\text{con}}^{total} = \frac{1}{|F|} \sum_{i=1}^{F} L_{\text{con}}^i(H_{fi}, e_s^+), \tag{7}$$

where F represents the total number of faces, τ is the temperature parameter, and m denotes the number of different part captions. This contrastive objective aligns geometric and textual representations in a shared semantic space. While our framework could theoretically incorporate larger neighborhoods (k-neighborhoods where $k > 1$), we empirically found that a 1-neighborhood configuration already captures sufficient contextual information for effective segmentation on the datasets we evaluated. Through preliminary experiments

on k-neighborhoods with $k \in \{1, 2, 3\}$ across our benchmark datasets, we observed that 1-neighborhood provides adequate geometric context for distinguishing between different parts, with the local spatial relationships being well-represented through the attention mechanism. Although extending to larger neighborhoods ($k > 1$) may potentially yield further performance improvements, the computational overhead increases significantly with larger k values due to the exponential growth in attention computations and memory requirements. Given that our 1-neighborhood approach already demonstrates superior performance compared to existing methods on these datasets (as shown in our experimental results), we adopt this configuration to maintain both effectiveness and computational efficiency.

D. Text-Driven Module: Multimodal Feature Integration

In our method, we innovatively integrate textual and geometric features through a Text-Driven Module. This module addresses the challenge of effectively combining heterogeneous feature modalities while preserving their complementary information.

The module takes three inputs: the neighborhood fused feature H_f, the corresponding sentence vector e_s^+, and the collection of word features e_w. Rather than simple concatenation, we employ an attention-based fusion mechanism that dynamically weights the contribution of each word based on its relevance to the current geometric context.

We first compute the semantic similarity between the neighborhood fused feature and each word embedding using cosine similarity, which is processed through a multi-layer perceptron (MLP) for non-linear transformation:

$$\alpha_i' = \frac{\exp(\gamma_0 \cos(H_f, e_{w_i}))}{\sum_{k=1}^{n} \exp(\gamma_0 \cos(H_f, e_{w_k}))}, \tag{8}$$

where γ_0 is a learnable temperature parameter that controls the sharpness of the attention distribution, and n represents the number of words in the part caption.

The attention weights α_i' form an attention map that indicates the relevance of each word to the current geometric context. This enables the model to focus on semantically relevant words while suppressing irrelevant information. The weighted word features are then aggregated to form a context-aware word representation:

$$c_j = \sum_{i=1}^{n} \alpha_i' \cdot e_{w_i}, \tag{9}$$

The final multimodal feature fusion employs an adaptive normalization mechanism inspired by Feature-wise Linear Modulation. This approach allows textual features to dynamically modulate geometric features:

$$\begin{aligned} H_f' = \alpha_f(\text{concat}(e_s^+, c_j, H_f)) \cdot \frac{H_f - \mu}{\sigma} \\ + \beta_f(\text{concat}(e_s^+, c_j, H_f)), \end{aligned} \tag{10}$$

where $\alpha_f(\cdot)$ and $\beta_f(\cdot)$ are learnable functions that generate scaling and shifting parameters based on the concatenated multimodal features. The neighborhood fused feature H_f is normalized using their batch statistics (μ and σ), and then modulated by textual information. In our implementation, $\alpha_f(\cdot)$ is realized as a normalization network, while $\beta_f(\cdot)$ is implemented as a linear projection layer.

This adaptive modulation mechanism enables textual features to directly influence the geometric representation, allowing the model to emphasize or suppress certain geometric characteristics based on semantic context.

E. Laplace-Adaptive Attention Module: Distribution-Aware Feature Weighting

We propose the Laplace-Adaptive Attention Module (LAAM) to account for the inherent variations in feature distributions across different 3D shapes and part types. Traditional attention mechanisms typically assume Gaussian distributions, which may not accurately model the heavy-tailed distributions commonly observed in geometric features. Our LAAM addresses this limitation by employing the Laplace distribution, which better captures the sparse and peaked nature of geometric feature distributions.

The LAAM operates by first computing the sample statistics of the input feature vector x:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2, \tag{11}$$

where N represents the feature dimension.

A key innovation of our LAAM is the introduction of a learnable offset δ that adjusts the sample mean to approximate better the population mean:

$$\phi = \mu + \delta. \tag{12}$$

This learned offset enables the model to dynamically adapt to different feature distributions, rather than relying solely on batch statistics. The input features are then normalized using the adjusted mean:

$$x_{\text{norm}} = \frac{x - \phi}{\sqrt{\sigma^2 + \epsilon}}, \tag{13}$$

where $\epsilon > 0$ ensures numerical stability.

The core of LAAM applies the Laplace probability density function to compute attention weights:

$$\text{LAAM}(x) = \frac{1}{2b} \exp\left(-\frac{|x_{\text{norm}}|}{b}\right), \tag{14}$$

where b is a learnable scale parameter. The Laplace distribution's exponential decay with absolute deviation makes it particularly suitable for geometric features, which often exhibit sparse activation patterns. The final output combines the original features with the computed attention weights through element-wise multiplication:

$$\text{Output} = x \odot \text{LAAM}(x), \tag{15}$$

where $\odot$ represents element-wise multiplication.

The LAAM's adaptive nature allows it to emphasize features that deviate significantly from the adjusted mean while suppressing noise and irrelevant variations. This distribution-aware attention mechanism proves particularly effective for 3D shape segmentation, where geometric features often exhibit non-Gaussian distributions.

Following the LAAM, the output is processed through a feed-forward network and a softmax layer for final label prediction. The cross-entropy loss quantifies the prediction accuracy:

$$L_f = -\frac{1}{|F|}\sum_{f\in F}\sum_{y} y \log p(y|h_f, x). \tag{16}$$

F. Training Algorithm and Optimization Strategy

Our training strategy combines two complementary objectives: semantic alignment through contrastive learning and accurate segmentation through cross-entropy loss. The final loss function balances these objectives:

$$L = \alpha L_f + (1-\alpha)L_{\text{con}}, \tag{17}$$

where $\alpha = 0.6$ empirically balances segmentation accuracy and semantic alignment. This weighting ensures that the model learns to produce accurate segmentations while maintaining semantic consistency between textual and geometric features.

Algorithm 1 outlines our complete training and inference procedures. During training, both geometric and textual features are processed to compute the contrastive loss, ensuring proper alignment in the shared embedding space. During inference, the model requires both geometric features from the 3D mesh and corresponding textual descriptions to perform segmentation, as the text-driven integration mechanism dynamically weights multimodal features to achieve accurate part-level predictions.

The training process ensures that geometric features are enriched with semantic understanding through contrastive learning alignment with textual representations. During inference, our method requires both geometric and textual inputs to achieve optimal segmentation performance, as the text-driven integration mechanism dynamically fuses multimodal features for accurate part-level predictions.

Algorithm 1 Training and Inference Process of Our Method

Input: 3D model dataset F with part caption annotations, trainable prefix $Prefix$, frozen RoBERTa model, Adam optimizer hyperparameters β_1, β_2, learning rate l_r, batch size M.
Output: Predicted label for each face on test 3D shapes.
Training Process:
For number of training iterations do:
  For $f = 1, \ldots, |F|$ do:
    // Prefix Tuning: Extract textual features
    Concatenate part caption embedding with trainable prefix;
    Forward through frozen RoBERTa to obtain e_s, e_w;
    // Mesh Self-Attention: Extract geometric features
    Input face features to obtain neighborhood fused feature H_f;
    Calculate contrastive learning loss L_{con};
    // Text-Driven Module: Multimodal fusion
    Compute attention weights for word features based on H_f;
    Aggregate weighted word features: c_j;
    Fuse textual and geometric features: $H'_f = \text{Fusion}(e_s, c_j, H_f)$;
    // Laplace-Adaptive Attention
    Apply Laplace attention weighting to H'_f;
    Obtain predicted label $p(y|h_f, x)$;
    Calculate cross-entropy loss L_f;
    Calculate total loss: $L = \alpha L_f + (1-\alpha)L_{\text{con}}$;
    Update parameters (prefix, text-driven, LAAM);
  End for
End for
Inference Process:
For $f = 1, \ldots, |F|$ do:
  // Prefix Tuning
  Concatenate part caption with learned prefix, encode via RoBERTa: e_s, e_w;
  // Mesh Self-Attention
  Extract neighborhood fused feature H_f;
  // Text-Driven Module
  Compute attention-weighted word features and fuse with H_f: H'_f;
  // Laplace-Adaptive Attention
  Apply Laplace attention weighting;
  Predict segmentation label $p(y|h_f, x)$ via trained classifier.
End for

IV. Experiments

This section describes how we conduct our experiments and provides a comprehensive overview of the datasets utilized in our evaluation. We present both qualitative and quantitative results of our method on four benchmark datasets: Princeton Shape Benchmark (PSB), COSEG, ShapeNetCore, and our proposed Fine-grained HumanBody. We compare our method's performance against the latest state-of-the-art techniques in 3D shape segmentation and conduct extensive ablation studies to evaluate the impact of individual components within our framework.

Evaluation Metrics. In this study, we employ the area-weighted accuracy computation method as suggested in [49] and [50]:

$$\text{Accuracy} = \frac{\sum_{i\in T} t_i \mathbf{u}(l_i)}{\sum_{i\in T} t_i}, \tag{18}$$

where T represents the set of faces in the testing 3D shapes, t_i denotes the area of face i, and l_i is the predicted label of face i. The function $u(l_i)$ equals 1 if the prediction is correct and 0 otherwise. This area-weighted metric provides a more meaningful evaluation than simple face-counting accuracy, as it considers the contribution of each face proportional to its surface area.

A. Experimental Datasets

We assess the effectiveness of our proposed method through a comprehensive evaluation on four benchmark datasets: PSB, COSEG, ShapeNetCore, and Fine-grained HumanBody. These datasets provide diverse challenges for 3D shape segmentation,

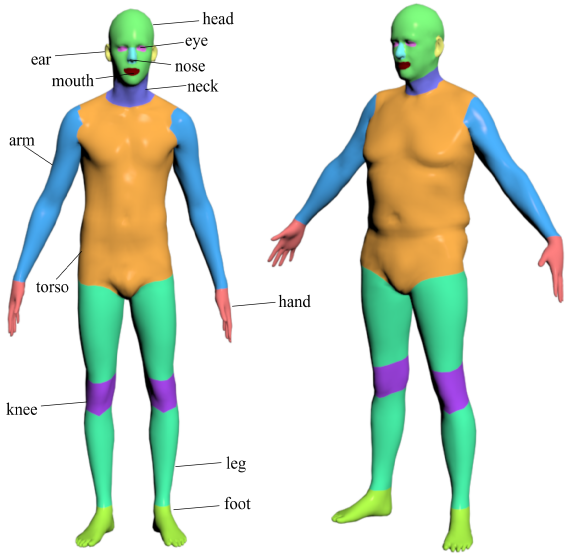


Fig. 4. Our newly proposed Fine-grained HumanBody Benchmark contains 100 meticulously annotated human shapes with fine-grained part segmentation, providing a challenging testbed for evaluating text-driven 3D shape segmentation methods.

covering various object categories, geometric complexities, and segmentation granularities.

Fine-grained HumanBody Dataset. We introduce a new benchmark called **Fine-grained HumanBody**, consisting of 100 human 3D models with detailed part annotations. Specifically, we randomly sampled 100 models from the original Human Body dataset [51] and performed fine-grained segmentation, dividing each model into 12 distinct anatomical parts, as illustrated in Figure 4. This dataset presents a particularly challenging scenario for text-driven segmentation, as human body parts often share similar geometric characteristics while having distinct semantic meanings. The Fine-grained HumanBody benchmark is publicly available at Google Drive[1] under an academic research license.

PSB Dataset. The PSB dataset was initially proposed by Chen et al. [52] and comprises 19 object categories, each containing 20 3D shapes, totaling 380 models. To ensure fair evaluation, we use the ground truth segmentation labels provided by Kalogerakis et al. [49] as the evaluation standard. The dataset covers diverse object types, including humans, animals, furniture, and tools, making it ideal for evaluating the generalization capability of segmentation algorithms.

COSEG Dataset. The COSEG dataset [53] is organized into two evaluation subsets: (1) a smaller subset consisting of 190 shapes across 8 categories (Candelabra, Chairs, Fourleg, Goblets, Guitars, Irons, Lamps, Vases), and (2) a larger subset containing 200 Tele-aliens, 400 chairs, and 300 vases. Like the PSB dataset, most COSEG shapes were preprocessed to establish an appropriate mesh topology suitable for geometric processing tasks. This dataset focuses on co-segmentation scenarios where consistent part labeling across similar shapes is crucial.

[1]https://drive.google.com/file/d/14s1st3K-BLb16znCupD0pOndjy6h9KME/view

ShapeNetCore Dataset. The ShapeNetCore dataset is a curated subset of the larger ShapeNet repository described in [54]. It encompasses 16 diverse object categories, including aeroplanes, bags, caps, cars, chairs, earphones, guitars, knives, lamps, laptops, motors, mugs, pistols, rockets, skateboards, and tables. This dataset provides comprehensive coverage of common household and industrial objects with varying geometric complexity. Because all 3D shapes in the original ShapeNetCore dataset are presented in point cloud, we therefore convert each shape in the dataset from point cloud to mesh first and conduct all experiments on the converted dataset to ensure consistency and fairness.

Textual Annotation Strategy. To adapt our text-driven approach across all four datasets, we provide part captions for every model. Since part labels and object categories are already available during dataset annotation, we automatically generate textual descriptions using predefined rules rather than requiring manual annotation. Each part of a 3D model is annotated with a simple textual description following the pattern "[part name] of a [object category]", such as "ear of a teddy", "handle of a cup", or "wing of an airplane". This rule-based text generation approach ensures consistent annotation across all datasets while significantly reducing annotation costs compared to manual text labeling. By providing these automatically generated textual annotations for 3D models, we effectively capture the relationships between textual and geometric information across multiple modalities, thereby overcoming the limitations of methods that rely solely on geometric features.

B. Main Results

We evaluate the effectiveness of our proposed text-driven framework through comprehensive experiments on four benchmark datasets. To ensure fair and consistent evaluation, we adhere to the performance metrics and experimental protocols established in prior literature, maintaining dataset-specific training-to-testing ratios as detailed below. It is important to note that the selection of comparative methods varies across datasets. This is primarily to ensure fair and reproducible comparisons. Instead of re-implementing competing approaches, we directly adopt officially reported results from the original publications. Since not all methods report results on every dataset, the available baselines naturally differ across tables.

On the PSB dataset, we compare our method against seven state-of-the-art fully supervised techniques: ShapeBoost [49], Hu et al. [56], PartSLIP [42], PointCLIP [36], ShapePFCN [41], MeshCNN [4], and MeshWalker [32]. Following established protocols, we maintain a 6:4 training-to-testing split across all experiments. As presented in Table I, our proposed method achieves a superior average accuracy of 96.27%, substantially outperforming all competing approaches. Notably, our method demonstrates the highest accuracy in 13 out of 16 object categories, with particularly significant improvements in challenging categories such as Human (94.38% vs. 93.80% by ShapePFCN), Hand (92.43% vs. 88.70% by ShapeBoost), and Bird (93.00% vs. 87.90% by ShapeBoost). The substantial performance gain in the Hand

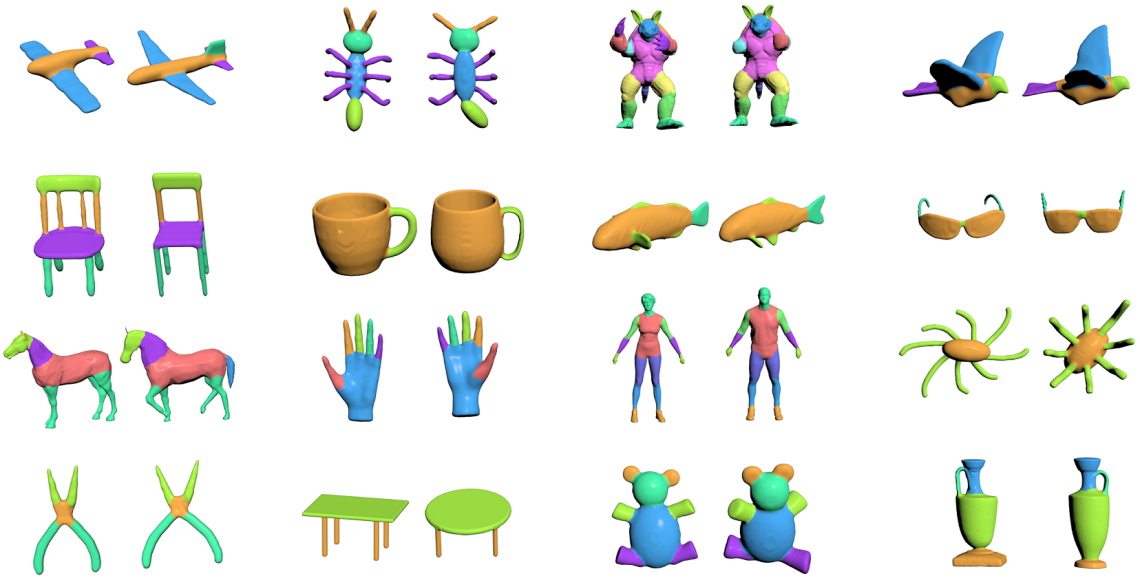


Fig. 5. Several experimental results showcasing the performance of our method on the PSB dataset.

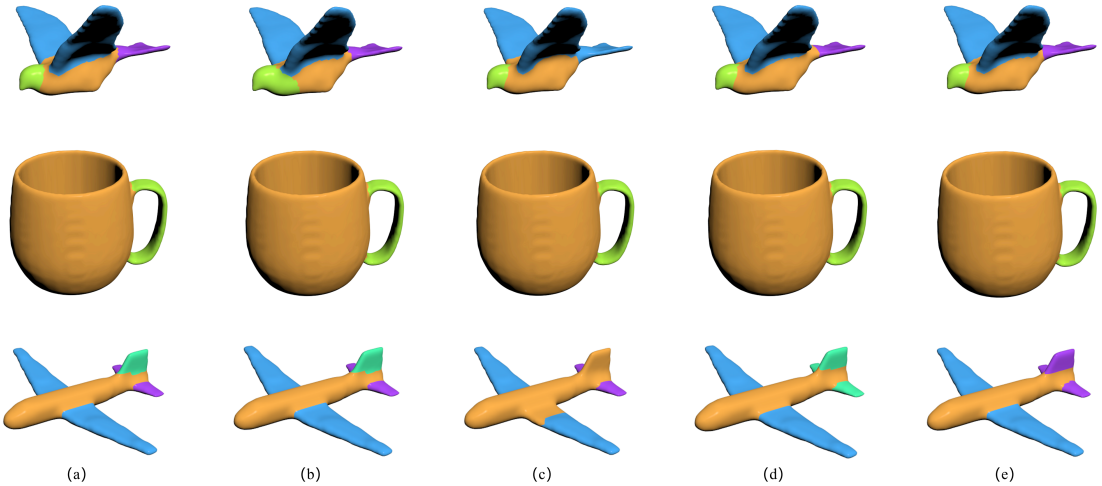


Fig. 6. Comparison with the other three segmentation algorithms. The methods displayed are as follows: (a) Ground truth, (b) Our method, (c) MeshCNN, (d) ShapePFCN, and (e) MeshWalker.

category (as illustrated in Figure 1) exemplifies our method's strength in handling geometrically similar parts through textual disambiguation. Figure 5 presents qualitative results demonstrating our algorithm's superior segmentation quality across diverse 3D shapes. Figure 6 provides a qualitative comparison of our method against MeshCNN, ShapePFCN, and MeshWalker on the PSB dataset.

On the small COSEG dataset comprising 190 shapes across 8 categories, we evaluate our method against three established approaches [55], [57], [58]. As detailed in Table II, our method achieves the highest average accuracy of 95.50%, outperforming all comparative methods. Specifically, our approach demonstrates superior performance in 5 out of 8 categories, with notable improvements in Lamps (94.37% vs. 92.41% by the second-best method) and Irons (93.58% vs. 91.22%). The underperformance in three categories (Goblets, Guitars, and Vases) can be attributed to the limited semantic diversity within these object classes, where textual features provide minimal additional discriminative power beyond geometric characteristics.

We conduct extensive evaluation on the large COSEG dataset, comparing our approach with seven recent methods: MeshCNN [4], MeshWalker [32], PDMeshNet [33], HodgeNet [34], SubdivNet [59], Laplacian2Mesh [35], and DGNet [60]. The experimental results in Table III demonstrate that our method achieves exceptional performance with an average accuracy of 98.90%, surpassing all competing approaches. Our method achieves the highest accuracy in Tele-aliens (99.02%) and Chairs (98.92%), while maintaining competitive performance in Vases (98.76%). Qualitative visualization results for the small COSEG and large COSEG datasets are provided in the supplementary materials.

On our newly proposed Fine-grained HumanBody benchmark, we compare our method with four established ap-

TABLE I
COMPARING OUR METHOD WITH OTHER SUPERVISED METHODS ON THE PSB DATASET. ALL EXPERIMENTS ARE CONDUCTED ON MESH REPRESENTATIONS. BASELINE RESULTS ARE CITED FROM [55].

Methods	ShapeBoost [49]	Hu [56]	PartSLIP [42]	PointCLIP [36]	ShapePFCN [41]	MeshCNN [4]	MeshWalker [32]	Ours (Acc.)	Ours (mIoU)
Human	93.20%	70.40%	92.12%	90.68%	93.80%	74.76%	87.02%	**94.38±0.82%**	89.24%
Cup	99.60%	97.40%	99.48%	**99.51%**	93.70%	95.86%	99.54%	99.43±0.31%	97.85%
Glasses	97.20%	**98.30%**	95.87%	95.89%	96.30%	93.94%	96.11%	98.12±0.45%	94.67%
Airplane	96.10%	83.30%	95.63%	95.74%	92.50%	84.36%	96.20%	**97.28±0.58%**	93.15%
Ant	98.80%	92.90%	98.45%	98.67%	**98.90%**	91.83%	97.36%	99.03±0.37%	96.78%
Chair	98.40%	89.60%	98.29%	97.85%	98.10%	84.75%	97.61%	**99.72±0.28%**	97.93%
Octopus	98.40%	97.50%	98.21%	98.14%	98.10%	98.21%	97.86%	**99.02±0.41%**	96.54%
Table	**99.30%**	99.00%	99.33%	99.27%	**99.30%**	96.78%	99.33%	99.32±0.33%	97.68%
Teddy	98.10%	97.10%	97.18%	97.61%	96.50%	84.29%	95.57%	**98.43±0.56%**	94.81%
Hand	88.70%	91.90%	89.47%	90.15%	88.70%	68.83%	83.31%	**92.43±1.15%**	85.37%
Plier	96.20%	86.00%	96.01%	95.78%	95.70%	83.69%	92.24%	**97.45±0.68%**	93.22%
Fish	95.60%	85.60%	96.54%	94.82%	95.90%	89.05%	94.58%	**97.12±0.72%**	92.48%
Bird	87.90%	71.50%	91.67%	90.34%	86.30%	68.09%	**92.76%**	93.00±0.94%	86.75%
Armadillo	90.10%	87.30%	90.28%	91.15%	**93.30%**	50.24%	89.12%	93.75±0.87%	87.93%
Vase	85.80%	80.20%	89.95%	89.67%	85.70%	68.94%	84.56%	**91.24±1.08%**	84.56%
FourLeg	86.20%	88.70%	84.87%	86.48%	89.50%	68.73%	80.93%	**90.52±1.21%**	83.18%
Average	94.35%	88.54%	94.52%	94.48%	93.89%	81.40%	92.76%	**96.27±0.67%**	91.82%

Fig. 7. Several experimental results evaluated on our proposed Fine-grained HumanBody dataset.

TABLE II
COMPARING OUR METHOD WITH OTHER SUPERVISED METHODS ON THE SMALL COSEG DATASET. ALL EXPERIMENTS ARE CONDUCTED ON MESH REPRESENTATIONS. BASELINE RESULTS ARE CITED FROM THEIR PAPER.

Methods	Ref [55]	Ref [57]	Ref [58]	Ours (Acc.)	Ours (mIoU)
Candelabra	93.28%	93.40%	94.93%	**95.72±0.58%**	91.34%
Chairs	97.05%	96.64%	96.88%	**97.62±0.42%**	94.87%
Fourleg	92.10%	93.37%	92.44%	**94.14±0.71%**	89.52%
Goblets	95.60%	97.92%	**97.99%**	96.57±0.48%	93.18%
Guitars	98.45%	97.85%	**98.73%**	98.61±0.35%	96.23%
Irons	90.39%	88.69%	91.22%	**93.58±0.83%**	88.76%
Lamps	90.15%	92.41%	87.18%	**94.37±0.76%**	89.94%
Vases	86.88%	88.13%	91.25%	**93.41±0.89%**	88.45%
Average	92.99%	93.55%	93.83%	**95.50±0.63%**	91.54%

proaches: MeshCNN [4], MeshWalker [32], HodgeNet [34], and SubdivNet [59]. To ensure experimental rigor, we reproduce these baseline methods under identical conditions, maintaining an 8:2 training-to-testing ratio. Table IV presents the comparative results, where our method achieves 94.00% accuracy, representing substantial improvements over SubdivNet (90.68%), MeshWalker (89.13%), HodgeNet (86.91%), and MeshCNN (86.21%). The significant performance gains on this fine-grained human body segmentation task validate the effectiveness of incorporating textual semantics for distinguishing anatomically similar but functionally distinct body parts. Figure 7 illustrates representative segmentation results on this challenging dataset.

We conduct a comprehensive evaluation on the ShapeNetCore dataset, encompassing 16 diverse object categories: aeroplanes, bags, caps, cars, chairs, earphones, guitars, knives, lamps, laptops, motors, mugs, pistols, rockets, skateboards, and tables. Our method is compared against seven state-of-the-art approaches: ShapeBoost [49], Guo et al. [50], ShapePFCN [41], SEG-MAT [61], PartNet [31], PCT [30], and Point-BERT [37]. As demonstrated in Table V, our text-driven framework achieves superior performance with a mean accuracy of 91.2%, substantially outperforming all competing methods. The performance improvement over the second-best method, ShapePFCN (88.4%), represents a significant advancement of 2.8 percentage points. Our method demonstrates particularly notable improvements in several challenging categories: aeroplanes (93.4% vs. 90.3%), bags (97.1% vs. 94.6%),

TABLE III
COMPARING OUR METHOD WITH OTHER SUPERVISED METHODS ON THE LARGE COSEG DATASET. ALL EXPERIMENTS ARE CONDUCTED ON MESH REPRESENTATIONS. BASELINE RESULTS ARE CITED FROM [55].

Methods	MeshCNN [4]	MeshWalker [32]	PDMeshNet [33]	HodgeNet [34]	SubdivNet [59]	Laplacian2Mesh [35]	DGNet [60]	Ours (Acc.)	Ours (mIoU)
Tele-aliens	95.76%	98.70%	98.18%	96.03%	97.30%	95.00%	97.40%	**99.02±0.38%**	97.15%
Chairs	94.54%	98.60%	97.23%	95.68%	96.70%	96.60%	96.70%	**98.92±0.45%**	96.78%
Vases	93.49%	**99.90%**	95.36%	90.30%	96.70%	94.60%	97.00%	98.76±0.52%	96.42%
Average	94.60%	98.77%	96.92%	94.00%	96.90%	95.40%	97.03%	**98.90±0.45%**	96.78%

TABLE IV
THE ACCURACY OF SEGMENTATION ON OUR PROPOSED FINE-GRAINED HUMANBODY DATASET COMPARED WITH OTHER SUPERVISED METHODS. ALL EXPERIMENTS ARE CONDUCTED ON MESH REPRESENTATIONS. BASELINE RESULTS ARE OUR REPRODUCTIONS.

Method	Accuracy	Method	Accuracy
MeshCNN [4]	86.21%	MeshWalker [32]	89.13%
SubdivNet [59]	90.68%	HodgeNet [34]	86.91%
Ours (Acc.)	**94.00±0.78%**	Ours (mIoU)	88.92%

and tables (90.7% vs. 84.8%). The method achieves the highest accuracy in 13 out of 16 categories, with especially strong performance on objects with clear semantic part distinctions, such as mugs (98.1%) and bags (97.1%).

The performance improvements across all datasets validate several key aspects of our approach: (1) The integration of textual semantics effectively resolves ambiguities in geometrically similar parts, as evidenced by substantial improvements in challenging categories like hands and human bodies; (2) Our attention-based multimodal fusion mechanism successfully leverages complementary information from both geometric and textual modalities; (3) The proposed LAAM effectively handles the non-Gaussian distributions commonly observed in geometric features; (4) The method demonstrates robust generalization across diverse object categories and geometric complexities, from simple shapes like mugs to complex articulated objects like chairs and human bodies. The primary distinction between our method and existing fully supervised approaches lies in our incorporation of textual semantics to complement geometric analysis. This multimodal approach proves particularly effective for segmenting 3D models where parts exhibit similar geometric characteristics but possess distinct semantic meanings, thereby addressing a fundamental limitation of purely geometry-based segmentation methods.

C. More Results

Robustness to Natural Language Descriptions. To evaluate the robustness of our method to different textual input modalities, we conduct additional experiments using natural language descriptions generated by GPT. Instead of rule-based patterns like "[part name] of a [object category]", we generate diverse natural language descriptions such as "the supporting leg that provides stability to the furniture" or "the curved armrest that provides comfort when seated", with varying lengths (8-20 words) and semantic richness. For each category in the PSB dataset, we generate 3-5 different natural language descriptions per part type to ensure diversity. As shown in Table VI, our method maintains robust performance with natural language descriptions, achieving 95.83% average accuracy compared to 96.27% with rule-based text. The marginal difference of only 0.44% demonstrates that our Prefix Tuning mechanism and attention-based text-driven integration effectively generalize to diverse linguistic expressions. Notably, challenging categories such as Hand (92.15% vs. 92.43%) and Human (94.02% vs. 94.38%) maintain highly competitive performance, confirming our method's robustness to different textual description styles and its suitability for real-world scenarios with varied natural language inputs.

Robustness to Text Variations. To evaluate our method's robustness to noisy textual inputs, we conduct experiments where 1-2 words are randomly deleted from rule-based text descriptions on the PSB dataset. As shown in Table VII, with 1-word deletion (e.g., "leg of chair" instead of "leg of a chair"), the average accuracy drops minimally from 95.95% to 94.75% (-1.20%). Even with 2-word deletion, where only partial text remains (e.g., "leg chair" or "a chair"), the average accuracy maintains 92.97%, representing only a 2.98% degradation. This robustness stems from two factors: (1) the frozen RoBERTa encoder captures semantic relationships that extract meaningful representations even from incomplete text, and (2) our Mesh Self-Attention Module provides geometric context that compensates when textual information is degraded. The graceful performance degradation confirms that our method effectively integrates multimodal information, with geometric features providing resilience against textual noise while textual semantics still contribute genuine discriminative value.

Failure Cases. Figure 8 illustrates representative failure cases. We observe two primary limitations: (1) Boundary fragmentation on objects with smooth geometric transitions (e.g., the cylinder), where the lack of distinct geometric edges leads to jagged segmentation boundaries; and (2) Semantic noise in anatomical transition regions (e.g., bird wings and human shoulders), where high geometric similarity between connected parts causes patchy over-segmentation. Future work aims to address these ambiguities by integrating Large Language Models (LLMs) to generate finer-grained semantic guidance.

Additional results, including computational efficiency analysis and visualization of contrastive learning alignment, are provided in the supplementary material.

TABLE V
COMPARING OUR METHOD WITH OTHER SUPERVISED METHODS ON THE SHAPENETCORE DATASET. ALL BASELINE EXPERIMENTS ARE CONDUCTED ON POINT CLOUD REPRESENTATIONS. BASELINE RESULTS ARE CITED FROM [55].

Category	ShapeBoost [49]	Guo et al. [50]	ShapePFCN [41]	SEG-MAT [61]	PartNet [31]	PCT [30]	Point-BERT [37]	Ours (Acc.)	Ours (mIoU)
aero	85.8	87.4	90.3	83.7	87.8	85.0	84.3	**93.4±0.72%**	88.6%
bag	93.1	91.0	94.6	-	86.7	82.4	84.8	**97.1±0.48%**	94.2%
cap	85.9	85.7	94.5	-	89.7	89.0	88.0	**95.2±0.61%**	91.4%
car	79.5	80.1	86.7	-	80.5	81.2	79.8	**88.2±0.89%**	82.7%
chair	70.1	66.8	82.9	80.3	**91.9**	**91.9**	91.0	90.4±0.78%	84.9%
eph.	81.4	79.8	84.9	82.1	75.7	71.5	81.7	**86.7±0.95%**	80.3%
guitar	89.0	89.9	91.8	90.9	91.8	91.3	91.6	**93.5±0.56%**	88.7%
knife	81.2	77.1	82.8	83.2	85.9	88.1	87.9	**90.1±0.84%**	84.5%
lamp	71.7	71.6	78.0	-	83.6	86.3	85.2	**87.9±0.91%**	82.1%
laptop	86.1	82.7	95.3	-	**97.0**	95.8	95.6	96.4±0.43%	93.5%
motor	77.2	80.1	87.0	-	74.6	64.6	75.6	**88.0±1.02%**	81.8%
mug	94.9	95.1	96.0	-	97.3	95.8	94.7	**98.1±0.38%**	95.7%
pistol	88.2	84.1	**91.5**	-	83.6	83.6	84.3	88.7±0.87%	82.4%
rocket	79.2	76.9	81.6	74.3	64.6	62.2	63.4	**82.1±1.18%**	75.3%
skate	91.0	89.6	91.9	79.6	78.4	77.6	76.3	**92.6±0.69%**	87.8%
table	74.5	77.8	84.8	80.4	85.8	83.7	81.5	**90.7±0.76%**	85.4%
mean	83.0	82.9	88.4	81.8	87.4	83.1	84.1	**91.2±0.75%**	86.2%

TABLE VI
ROBUSTNESS EVALUATION: SEGMENTATION ACCURACY COMPARISON OF RULE-BASED TEXT DESCRIPTION AND NATURAL LANGUAGE DESCRIPTION ON THE PSB DATASET.

Category	Rule-based Text	Natural Language	Difference
Human	**94.38%**	94.02%	-0.36%
Cup	**99.43%**	99.28%	-0.15%
Glasses	**98.12%**	97.95%	-0.17%
Airplane	**97.28%**	97.05%	-0.23%
Ant	**99.03%**	98.91%	-0.12%
Chair	**99.72%**	99.38%	-0.34%
Octopus	**99.02%**	98.87%	-0.15%
Table	**99.32%**	99.18%	-0.14%
Teddy	**98.43%**	98.21%	-0.22%
Hand	**92.43%**	92.15%	-0.28%
Plier	**97.45%**	97.23%	-0.22%
Fish	**97.12%**	96.88%	-0.24%
Bird	**93.00%**	92.71%	-0.29%
Armadillo	**93.75%**	93.52%	-0.23%
Vase	**91.24%**	90.97%	-0.27%
FourLeg	**90.52%**	90.19%	-0.33%
Average	**96.27%**	95.83%	-0.44%

TABLE VII
COMPARISON OF SEGMENTATION ACCURACY USING RULE-BASED VERSUS NATURAL LANGUAGE TEXT DESCRIPTIONS ON THE PSB DATASET.

Category	Clean Text	1-Word Deletion	2-Word Deletion	Accuracy Drop
Human	**94.38%**	93.12%	91.45%	-2.93%
Chair	**99.72%**	98.86%	97.53%	-2.19%
Hand	**92.43%**	90.87%	88.21%	-4.22%
Airplane	**97.28%**	96.15%	94.68%	-2.60%
Average	**95.95%**	94.75%	92.97%	-2.98%

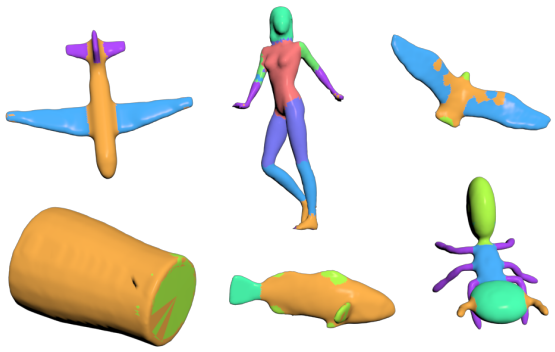


Fig. 8. Representative failure cases. The cylinder exhibits boundary fragmentation due to smooth geometry, while the bird and human models show semantic noise in anatomical transition regions.

TABLE VIII
ABLATION STUDY OF DIFFERENT INPUT FEATURE DESCRIPTORS.

Descriptors	SDF	SDF +AGD	SDF +AGD +GC	SDF +AGD +GC +WKS	SDF +AGD +GC +WKS +SIHKS
Human	72.47%	83.26%	88.63%	91.28%	**94.38%**
Ant	92.58%	94.15%	96.88%	97.79%	**99.03%**
Chair	88.35%	92.17%	95.43%	97.36%	**99.72%**
Hand	65.80%	68.47%	77.65%	86.44%	**92.43%**

D. Ablation study

In this section, we perform an ablation study to examine the critical elements of our approach and assess their influence on the overall effectiveness of the proposed algorithm. We provide an ablation study about the design choice of the Laplace-Adaptive attention module in the supplementary material.

Input Feature Descriptor. Our ablation experiments on fea-

TABLE IX
ABLATION STUDY EXPERIMENTS ON THE NETWORK FRAMEWORK.

Module	No Prefix	No Mesh Self-Attention	No Text-Driven	No Laplace-Attention	All Modules
PSB	92.43%	90.12%	87.66%	86.79%	**96.27%**
small COSEG	88.93%	85.10%	82.39%	82.10%	**95.50%**
Fine-grained HumanBody	90.51%	91.68%	87.46%	86.78%	**94.00%**

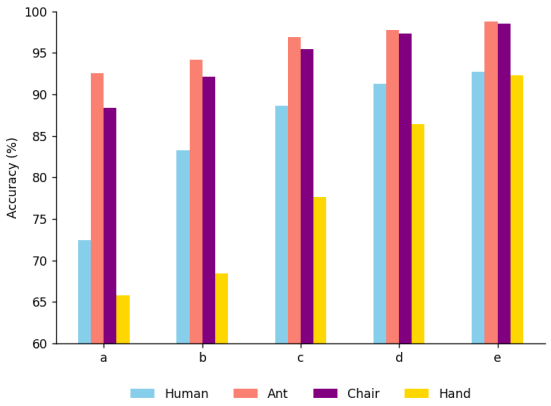


Fig. 9. Ablation Experiments on Feature Descriptors. The figure illustrates the influence of different feature descriptor selections on the accuracy of segmentation label predictions. The horizontal axis represents different combinations of feature descriptors: (a) SDF, (b) SDF + AGD, (c) SDF + AGD + GC, (d) SDF + AGD + GC + WKS, (e) SDF + AGD + GC + WKS + SIHKS.

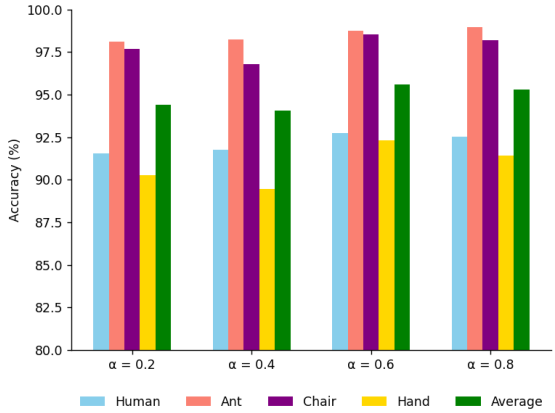


Fig. 10. Ablation Experiments on the ratio between cross-entropy loss and contrastive loss. The figure illustrates the influence of ratio settings on the accuracy of segmentation label predictions. When α is 0.6, the average segmentation accuracy is highest on the four classes we evaluated.

ture descriptors reveal a clear progressive improvement pattern as geometric feature richness increases. As demonstrated in Table VIII and Figure 9, the segmentation accuracy exhibits substantial gains when transitioning from a single descriptor to our complete 122-dimensional feature vector. Notably, the most dramatic improvements occur in challenging categories: the Hand category demonstrates a remarkable 26.52% accuracy increase (from 65.80% to 92.32%) when progressing from SDF alone to the full descriptor combination, while the Human category shows a 20.25% improvement (from 72.47% to 92.72%). This pronounced enhancement in anatomically complex objects validates our hypothesis that multi-scale geometric descriptors are particularly crucial for disambiguating parts with subtle geometric variations. The consistent upward trend across all categories, with each additional descriptor contributing meaningful performance gains, confirms that our comprehensive geometric representation effectively captures complementary shape characteristics essential for accurate segmentation.

Network Structure. To verify the contribution of each module in the proposed framework to the overall segmentation performance, we performed an ablation study on each module and assessed its influence on the segmentation results. Experiments were conducted on three datasets: PSB, small COSEG, and Fine-grained HumanBody. The quantitative results are summarized in Table IX. The findings from the ablation of different modules are as follows: When the Prefix Tuning component is excluded, segmentation results on all three datasets exhibit a slight decrease in accuracy, suggesting that introducing prefix tuning effectively reduces

TABLE X
ABLATION STUDY EXPERIMENTS ON THE WEIGHT BETWEEN CROSS-ENTROPY LOSS AND CONTRASTIVE LOSS.

α	0.2	0.4	0.6	0.8
Human	91.53%	91.78%	**94.38%**	92.54%
Ant	98.12%	98.25%	98.74%	**99.03%**
Chair	97.69%	96.78%	**99.72%**	98.19%
Hand	90.25%	89.45%	**92.43%**	91.43%
Average	94.40%	94.07%	**96.32%**	95.30%

the disparity between input data and enhances segmentation accuracy. When the Mesh Self-Attention module is omitted, a significant decline in segmentation accuracy is observed across all three datasets. Furthermore, when the Text-Driven components are removed, the segmentation results become less consistent, leading to confusion in identifying certain regions. Finally, when the Laplace-Attention module is not used, the performance of our network is the worst.

Parameter Settings. To evaluate the impact of parameter settings on the experimental outcomes, we conducted an ablation study on the parameters used in the experiments. First, we investigated the effect of the ratio between cross-entropy loss and contrastive loss on segmentation accuracy, experimenting with values of α ranging from 0.2 to 0.8. The segmentation accuracy of four models on the PSB dataset is presented in Table X. The results indicate that the average accuracy is optimum when α is set to 0.6. The corresponding qualitative results are shown in Figure 10.

Subsequently, we ablated the temperature coefficient used in contrastive learning to determine the optimal balance be-

TABLE XI
ABLATION STUDY EXPERIMENTS ON THE TEMPERATURE.

τ	0.05	0.1	0.2	0.3
Human	90.43%	**94.38%**	91.88%	91.56%
Ant	97.96%	**99.03%**	96.97%	97.12%
Chair	98.12%	**99.72%**	97.26%	97.38%
Hand	90.47%	**92.43%**	90.78%	92.03%
Average	94.25%	**96.39%**	94.22%	94.52%

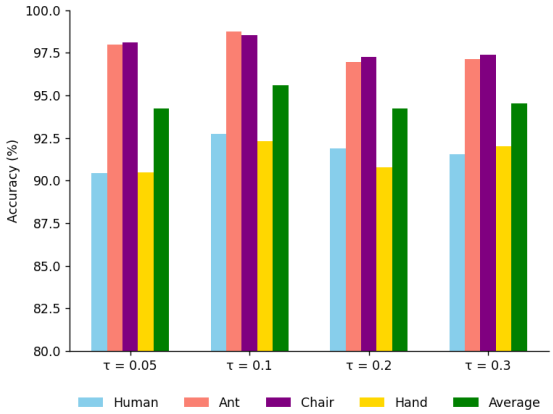


Fig. 11. Ablation Experiments on Temperature. The figure illustrates the influence of different contrastive loss temperature selections on the accuracy of segmentation label predictions. When τ is 0.1, segmentation accuracy reaches the highest in all four categories we evaluated.

tween geometric and textual features in the vector space. The experimental results are provided in Table XI, showing that the highest average segmentation accuracy was achieved when the temperature coefficient was set to 0.1. The qualitative results are illustrated in Figure 11.

Module Combination. To investigate the synergistic effects and interactions between different modules in our framework, we conduct comprehensive module combination experiments on the PSB dataset. As shown in Table XII, we evaluate individual modules, pairwise combinations, and the full model. Individual modules provide modest improvements: Mesh Self-Attention achieves the highest individual gain (+3.55%), while Prefix Tuning (+2.13%), Text-Driven (+1.88%), and LAAM (+1.36%) contribute smaller individual improvements. Notably, pairwise combinations reveal significant synergistic effects: Prefix + Text-Driven achieves +4.73% (exceeding the sum of individual contributions), demonstrating that semantic alignment and dynamic textual fusion work synergistically; Mesh Self-Attention + Text-Driven achieves +6.66%, showing strong complementarity between geometric context and textual semantics; and Mesh Self-Attention + LAAM achieves +6.08%, indicating effective interaction between geometric modeling and distribution-aware attention. The full model with all four modules achieves 96.27% (+9.48%), significantly outperforming all partial combinations. This comprehensive ablation reveals that: (1) textual components (Prefix + Text-Driven) provide substantial semantic disambiguation beyond individual contributions; (2) combining geometric context with textual features yields stronger improvements than either modality alone; (3) LAAM provides consistent improvements across all combinations through adaptive feature weighting; and (4) the full integration of all modules produces synergistic effects that substantially exceed the sum of individual contributions, demonstrating the effectiveness of our multimodal fusion design.

TABLE XII
MODULE COMBINATION ABLATION STUDY ON PSB DATASET TO REVEAL SYNERGISTIC EFFECTS BETWEEN COMPONENTS.

Module Combination	**PSB Accuracy**	**Improvement**
Baseline (no modules)	86.79%	-
Individual Modules		
+ Prefix Tuning only	88.92%	+2.13%
+ Text-Driven only	88.67%	+1.88%
+ Mesh Self-Attention only	90.34%	+3.55%
+ LAAM only	88.15%	+1.36%
Pairwise Combinations		
+ Prefix + Text-Driven	91.52%	+4.73%
+ Prefix + Mesh Self-Attention	92.43%	+5.64%
+ Mesh Self-Att. + Text-Driven	93.45%	+6.66%
+ Text-Driven + LAAM	91.78%	+4.99%
+ Prefix + LAAM	91.23%	+4.44%
+ Mesh Self-Att. + LAAM	92.87%	+6.08%
Full Model (All Four Modules)	96.27%	+9.48%

E. Performance

We implemented our algorithm using Matlab and Python. The experiments were conducted on a PC with an Intel Core i7 CPU, 32GB RAM, and an NVIDIA GeForce RTX 4090 GPU. The proposed method comprises two main phases: the training phase and the testing phase. The algorithm requires approximately 15 minutes to train on shapes from a single category during the training phase. In the testing phase, segmenting an unlabeled shape takes approximately 10 seconds. The process for handling shapes from one category of the PSB dataset takes around 30 minutes.

V. LIMITATIONS AND FUTURE WORK

Although the text-driven framework proposed in this paper is effective in various 3D shape segmentation tasks, it has some limitations. Firstly, the method heavily relies on the semantic richness of the textual descriptions. In categories with low semantic diversity (e.g., Goblets or Vases), the performance gain is limited because simple text offers minimal discriminative power beyond geometry. In the future, we aim to integrate Large Language Models (LLMs) to automatically generate fine-grained and context-aware captions to enhance performance in semantically sparse categories. Secondly, the Mesh Self-Attention module is currently limited to a 1-neighborhood configuration to maintain computational efficiency. Extending to larger receptive fields ($k > 1$) incurs significant computational overhead due to the exponential growth in attention calculations. We plan to work on optimizing the attention mechanism using sparse or linear-complexity strategies to capture broader geometric contexts efficiently in the future.

VI. Conclusion

In this paper, we proposed an innovative text-driven framework for 3D shape segmentation that fundamentally transforms traditional geometry-only approaches by systematically integrating textual semantics with geometric analysis. The core innovation of our work lies in addressing the intrinsic modality misalignment and feature distribution challenges in multimodal 3D understanding. Specifically, we introduced a Prefix Tuning mechanism that effectively bridges the semantic granularity gap between coarse part-level textual captions and fine-grained face-level geometric features. Furthermore, to tackle the non-Gaussian, heavy-tailed nature of geometric feature distributions—a critical aspect often overlooked by standard attention mechanisms—we developed the Laplace-Adaptive Attention Module (LAAM) to perform distribution-aware feature weighting. Through contrastive learning alignment, we established a unified semantic space that enables effective disambiguation of geometrically ambiguous parts. Additionally, we contributed the Fine-grained HumanBody benchmark to facilitate comprehensive evaluation of text-driven segmentation methods. Extensive experiments on multiple benchmarks demonstrate that our approach significantly outperforms existing state-of-the-art methods, validating the effectiveness of our text-driven innovations.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62172356, 61872321, 62272277), the Zhejiang Provincial Natural Science Foundation of China (LZ25F020012).

References

[1] A. Mademlis, P. Daras, A. Axenopoulos, D. Tzovaras, and M. G. Strintzis, "Combining topological and geometrical features for global and partial 3-D shape retrieval," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 819–831, 2008.

[2] Z. Kuang, J. Yu, S. Zhu, Z. Li, and J. Fan, "Effective 3-D shape retrieval by integrating traditional descriptors and pointwise convolution," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3164–3177, 2019.

[3] Y. Yang, W. Xu, X. Guo, K. Zhou, and B. Guo, "Boundary-aware multidomain subspace deformation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 10, pp. 1633–1645, 2013.

[4] R. Hanocka, A. Hertz, N. Fish, R. Giryes, S. Fleishman, and D. Cohen-Or, "MeshCNN: A network with an edge," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1–12, 2019.

[5] J. Liu, Y. Chen, B. Ni, and Z. Yu, "Joint global and dynamic pseudo labeling for semi-supervised point cloud sequence segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5679–5691, 2023.

[6] Y. Su, X. Xu, and K. Jia, "Weakly supervised 3D point cloud segmentation via multi-prototype learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7723–7736, 2023.

[7] Z. Song, L. Zhao, and J. Zhou, "Learning hybrid semantic affinity for point cloud segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4599–4612, 2022.

[8] H. Shi, R. Li, F. Liu, and G. Lin, "Temporal feature matching and propagation for semantic segmentation on 3D point cloud sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7491–7502, 2023.

[9] X. Chang, H. Pan, W. Sun, and H. Gao, "A multi-phase camera-lidar fusion network for 3D semantic segmentation with weak supervision," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 3737–3746, 2023.

[10] M. Rong and S. Shen, "3D semantic segmentation of aerial photogrammetry models based on orthographic projection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7425–7437, 2023.

[11] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[13] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 534–11 542.

[14] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[15] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "Aˆ 2-nets: Double attention networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[16] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[17] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 603–612.

[18] L. Shapira, A. Shamir, and D. Cohen-Or, "Consistent mesh partitioning and skeletonisation using the shape diameter function," *The Visual Computer*, vol. 24, no. 4, pp. 249–259, 2008.

[19] A. Golovinskiy and T. Funkhouser, "Randomized cuts for 3D mesh analysis," in *ACM SIGGRAPH Asia 2008 papers*, 2008, pp. 1–12.

[20] S. Katz and A. Tal, "Hierarchical mesh decomposition using fuzzy clustering and cuts," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 954–961, 2003.

[21] M. Hachani, A. O. Zaid, and R. Khliwi, "Segmentation of 3D articulated meshes using shape diameter function and curvature information," in *2016 IEEE International Conference on Multimedia and Expo*, 2016, pp. 1–5.

[22] H. Yamauchi, S. Gumhold, R. Zayer, and H.-P. Seidel, "Mesh segmentation driven by gaussian curvature," *The Visual Computer*, vol. 21, no. 8, pp. 659–668, 2005.

[23] M. Attene, B. Falcidieno, and M. Spagnuolo, "Hierarchical mesh segmentation based on fitting primitives," *The Visual Computer*, vol. 22, no. 3, pp. 181–193, 2006.

[24] M. Vieira and K. Shimada, "Surface mesh segmentation and smooth surface extraction through region growing," *Computer Aided Geometric Design*, vol. 22, no. 8, pp. 771–792, 2005.

[25] Z. Shu, X. Shen, S. Xin, Q. Chang, J. Feng, L. Kavan, and L. Liu, "Scribble-based 3D shape segmentation via weakly-supervised learning," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 8, pp. 2671–2682, 2019.

[26] Z. Shu, C. Qi, S. Xin, C. Hu, L. Wang, Y. Zhang, and L. Liu, "Unsupervised 3D shape segmentation and co-segmentation via deep learning," *Computer Aided Geometric Design*, vol. 43, pp. 39–52, 2016.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[28] C. R. Qi, H. Su, K. Mo, and L. 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*, 2017, pp. 652–660.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[30] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "PCT: Point cloud transformer," *Computational Visual Media*, vol. 7, pp. 187–199, 2021.

[31] F. Yu, K. Liu, Y. Zhang, C. Zhu, and K. Xu, "PartNet: A recursive part decomposition network for fine-grained and hierarchical shape segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9491–9500.

[32] A. Lahav and A. Tal, "MeshWalker: Deep mesh understanding by random walks," *ACM Transactions on Graphics*, vol. 39, no. 6, pp. 1–13, 2020.

[33] F. Milano, A. Loquercio, A. Rosinol, D. Scaramuzza, and L. Carlone, "Primal-dual mesh convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 952–963, 2020.
[34] D. Smirnov and J. Solomon, "HodgeNet: Learning spectral geometry on triangle meshes," *ACM Transactions on Graphics*, vol. 40, no. 4, pp. 1–11, 2021.
[35] Q. Dong, Z. Wang, M. Li, J. Gao, S. Chen, Z. Shu, S. Xin, C. Tu, and W. Wang, "Laplacian2mesh: Laplacian-based mesh understanding," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 7, pp. 4349–4361, 2024.
[36] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li, "PointCLIP: Point cloud understanding by CLIP," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8552–8562.
[37] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-BERT: Pre-training 3D point cloud transformers with masked point modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 313–19 322.
[38] L. Xue, M. Gao, C. Xing, R. Martín-Martín, J. Wu, C. Xiong, R. Xu, J. C. Niebles, and S. Savarese, "ULIP: Learning a unified representation of language, images, and point clouds for 3D understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1179–1189.
[39] M. Liu, R. Shi, K. Kuang, Y. Zhu, X. Li, S. Han, H. Cai, F. Porikli, and H. Su, "OpenShape: Scaling up 3D shape representation towards open-world understanding," *Advances in Neural Information Processing Systems*, vol. 36, pp. 44 860–44 879, 2023.
[40] Y. Wang, M. Gong, T. Wang, D. Cohen-Or, H. Zhang, and B. Chen, "Projective analysis for 3D shape segmentation," *ACM Transactions on Graphics*, vol. 32, no. 6, pp. 1–12, 2013.
[41] E. Kalogerakis, M. Averkiou, S. Maji, and S. Chaudhuri, "3D shape segmentation with projective convolutional networks," in *Proceedings of IEEE Computer Vision and Pattern Recognition*, 2017, pp. 3779–3788.
[42] M. Liu, Y. Zhu, H. Cai, S. Han, Z. Ling, F. Porikli, and H. 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*, 2023, pp. 21 736–21 746.
[43] Y. Liu, "RoBERTa: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
[44] L. Shapira, S. Shalom, A. Shamir, D. Cohen-Or, and H. Zhang, "Contextual part analogies in 3D objects," *International Journal of Computer Vision*, vol. 89, no. 2, pp. 309–326, 2010.
[45] R. Gal and D. Cohen-Or, "Salient geometric features for partial shape matching and similarity," *ACM Transactions on Graphics*, vol. 25, no. 1, pp. 130–150, 2006.
[46] L. Shapira, A. Shamir, and D. Cohen-Or, "Consistent mesh partitioning and skeletonisation using the shape diameter function," *The Visual Computer*, vol. 24, no. 4, pp. 249–259, 2008.
[47] M. M. Bronstein and I. Kokkinos, "Scale-invariant heat kernel signatures for non-rigid shape recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1704–1711.
[48] D. Raviv, M. M. Bronstein, A. M. Bronstein, and R. Kimmel, "Volumetric heat kernel signatures," in *Proceedings of the ACM Workshop on 3D Object Retrieval*. ACM, 2010, pp. 39–44.
[49] E. Kalogerakis, A. Hertzmann, and K. Singh, "Learning 3D mesh segmentation and labeling," *ACM Transactions on Graphics*, vol. 29, no. 4, pp. 1–12, 2010.
[50] K. Guo, D. Zou, and X. Chen, "3D mesh labeling via deep convolutional neural networks," *ACM Transactions on Graphics*, vol. 35, no. 1, pp. 1–12, 2015.
[51] H. Maron, M. Galun, N. Aigerman, M. Trope, N. Dym, E. Yumer, V. G. Kim, and Y. Lipman, "Convolutional neural networks on surfaces via seamless toric covers," *ACM Transactions on Graphics*, vol. 36, pp. 1–10, 2017.
[52] X. Chen, A. Golovinskiy, and T. Funkhouser, "A benchmark for 3D mesh segmentation," *ACM Transactions on Graphics*, vol. 28, no. 3, pp. 1–12, 2009.
[53] Y. Wang, S. Asafi, O. van Kaick, H. Zhang, D. Cohen-Or, and B. Chen, "Active co-analysis of a set of shapes," *ACM Transactions on Graphics*, vol. 31, no. 6, pp. 1–10, 2012.
[54] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "ShapeNet: An information-rich 3D model repository," *arXiv preprint arXiv:1512.03012*, 2015.
[55] Z. Shu, T. Wu, J. Shen, S. Xin, and L. Liu, "Semi-supervised 3D shape segmentation via self refining," *IEEE Transactions on Image Processing*, vol. 33, pp. 2044–2057, 2024.
[56] R. Hu, L. Fan, and L. Liu, "Co-segmentation of 3D shapes via subspace clustering," *Computer Graphics Forum*, vol. 31, no. 5, pp. 1703–1713, 2012.
[57] Z. Shu, X. Sun, S. Xin, and L. Liu, "3D shape segmentation via attentive nonuniform downsampling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 12, pp. 12 184–12 196, 2024.
[58] Z. Shu, S. Li, S. Xin, and L. Liu, "3D shape segmentation with potential consistency mining and enhancement," *IEEE Transactions on Multimedia*, vol. 27, pp. 133–144, 2025.
[59] S.-M. Hu, Z.-N. Liu, M.-H. Guo, J.-X. Cai, J. Huang, T.-J. Mu, and R. R. Martin, "Subdivision-based mesh convolution networks," *ACM Transactions on Graphics*, vol. 41, no. 3, pp. 1–16, 2022.
[60] X.-L. Li, Z.-N. Liu, T. Chen, T.-J. Mu, R. R. Martin, and S.-M. Hu, "Mesh neural networks based on dual graph pyramids," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 7, pp. 4211–4224, 2024.
[61] C. Lin, L. Liu, C. Li, L. Kobbelt, B. Wang, S. Xin, and W. Wang, "SEG-MAT: 3D shape segmentation using medial axis transform," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 6, pp. 2430–2444, 2020.

Zhenyu Shu got his Ph.D. degree in 2010 at Zhejiang University, China. He is now working as a full professor at NingboTech University. His research interests include computer graphics, digital geometry processing, and machine learning. He has published over 40 papers in international conferences or journals.

Chenyu Zhu is a graduate student of the College of Computer Science and Technology at Zhejiang University. His research interests include image processing, computer graphics, and machine learning.

Shiqing Xin is a full professor at the Faculty of School of Computer Science and Technology in Shandong University. He received his Ph.D. degree in applied mathematics at Zhejiang University in 2009. His research interests include computer graphics, computational geometry and 3D printing.