This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Heat Diffusion based Multi-scale and Geometric Structure-aware Transformer for Mesh Segmentation

Chi-Chong Wong
University of Macau
amilton.wong@connect.um.edu.mo

Abstract

Triangle mesh segmentation is an important task in 3D shape analysis, especially in applications such as digital humans and AR/VR. Transformer model is inherently permutation-invariant to input, which makes it a suitable candidate model for 3D mesh processing. However, two main challenges involved in adapting Transformer from natural languages to 3D mesh are yet to be solved, such as i) extracting the multi-scale information of mesh data in an adaptive manner; ii) capturing geometric structures of mesh data as the discriminative characteristics of the shape. Current point based Transformer models fail to tackle such challenges and thus provide inferior performance for discretized surface segmentation. In this work, heat diffusion based method is exploited to tackle these problems. A novel Transformer model called MeshFormer is proposed, which i) integrates Heat Diffusion method into Multi-head Self-Attention operation (HDMSA) to adaptively capture the features from local neighborhood to global contexts; ii) applies a novel Heat Kernel Signature based Structure Encoding (HKSSE) to embed the intrinsic geometric structures of mesh instances into Transformer for structure-aware processing. Extensive experiments on triangle mesh segmentation validate the effectiveness of the proposed MeshFormer model and show significant improvements over current state-of-the-art methods.

1. Introduction

Discretized surface semantic segmentation is a task to semantically classify the labeling of each discrete element in 3D discretized surface. Such discrete element can be triangle face in mesh input [22, 24, 37] or 3D point in point cloud input [11, 20, 26, 28, 34, 48–50, 52]. It is an essential task in many applications for 3D vision and computer graphics, such as 3D human body analysis, digital humans and AR/VR, etc. In this work, we mainly focus on mesh representation as input, as point cloud representation can be regarded as the special case of mesh which discards the surface connectivity.

The challenges in learning mesh representation involves its inherent characteristics such as irregularity and unorderedness. Following the success in NLP [7, 15, 44] and 2D computer vision domain [16, 29, 43, 46], Transformer model such as [20, 32, 50, 52] has been adopted as an effective model for processing 3D point cloud input due to its inherent capability in processing unordered point sets. Along such direction, it is natural to adapt the Transformer model to the mesh input, which is also one of the most common representation for 3D input modality. However, adapting Transformer model from natural languages to mesh input, with respect to the specific characteristics of mesh structures, involves many challenges, such as i) extracting the multi-scale information of mesh data in an adaptive manner; ii) capturing geometric structures of mesh representation as the discriminative characteristics. Sufficiently capturing such two essential information is the prerequisite for accurate mesh based semantic segmentation task. In retrospect to the recent point cloud based Transformer models [20, 32, 52], it is found that all these methods did not provide effective approaches to tackle the challenges mentioned above, and thus provided a limited increment in segmentation accuracy for mesh input.

Recent work Swin Transformer [29] applies multiple fixed-size windows for limiting the attention computation along scale-varying regions, and hence provides a hierarchical Transformer model to extract multi-scale information for dense prediction task. However, such approach which uses fixed-size windows is only applicable to regular 2D images input, while it is infeasible for irregular input such as 3D meshes with diverse shapes. In order to adapt the Transformer model for adaptively extracting multi-scale information for irregular mesh input, it is essential to extend its core operation, self-attention, to have capability in progressively comparing the feature similarity from local neighborhood to global range of the mesh input, in the form of intrinsic geometry of surface [9]. In fact, several seminal works [12, 13] had studied on using heat diffusion method to

intrinsically communicate the interactions with the neighbouring vertices on discretized surface. Such heat diffusion method is able to compute the geodesic distance on mesh in a stable and accurate manner, which is essential to capture the intrinsic locality in discretized surface.

In this work, we opt for the heat diffusion method to propose a novel self-attention operation called Heat Diffusion based Multi-head Self-Attention (HDMSA), which adaptively limits the self-attention computation within multiple heat diffusion ranges to capture the multi-scale surface features from local neighborhood to global contexts. This extension facilitates the construction of multi-scale Transformer encoder in the proposed MeshFormer model.

The second challenging issue in adapting Transformer model for mesh input is to encode the geometric structural information of mesh as a supplement for shape-specific inductive bias of Transformer model. In Non-Euclidean domain, very recent works such as SAN [25], GNN-LSPE [17] methods have investigated on injecting the eigenfunctions, which are derived from graph Laplacian operator, into positional encoding as a kind of graph-specific inductive bias to traditional Transformer model. This approach exploits the spectral information of graph Laplacian to capture the structure of input modality, and thus is considered as a promising candidate for processing mesh input. However, this approach suffers from the issue of eigenfunction sign ambiguity, that means eigenfunction with either positive or negative sign still satisfies the original eigenproblem and is associated with the same eigenvalue, which lowers the discriminative power in the extracted structural information. In this work, instead of directly applying the spectral information, a novel Heat Kernel Signature based Structure Encoding (HKSSE) module is proposed, which effectively captures the intrinsic geometric structural information of the mesh while bypassing the issue of eigenfunction sign ambiguity. Moreover, it provides a more powerful way to capture the more advanced geometric information, i.e., the symmetry in geometry structure, which is very common in human body shapes. As a result, this capability brought to Transformer model reinforces a structure-aware segmentation prediction for mesh input.

To the best of our knowledge, the proposed model is the first mesh based Transformer model which integrates heat diffusion methods to tackle the discretized surface semantic segmentation problem. The main contributions of this work are summarized as follows:

1. With the heat diffusion extension, the proposed multi-head self-attention operation allows intrinsic communication for vertices on mesh input from local neighborhood to global context and thus is able to capture multi-scale mesh features.

2. A novel heat kernel signature based structure encoding module is applied to embed the mesh intrinsic geometric structures into Transformer for providing structure-aware segmentation output.

3. The proposed work demonstrates the feasibility on the extension of generic Transformer model structure for 3D mesh input with heat diffusion methods.

2. Related Works

In this section, the three main related techniques: deep learning methods on triangular meshes, Transformer on point clouds and heat diffusion methods are discussed.

2.1. Deep Learning on 3D Meshes

With the significant success in image recognition tasks [19, 23, 38, 42], deep convolutional neural network (CNN) models have been extended to be applied on processing the 3D mesh input in computer graphics domain. The early pioneering works [5, 31] generalize the CNN operation on mesh structure. Such methods although provide promising results, require complex geometric tools and thus suffer from great difficulties in training. Recent works such as MeshCNN [22] and HodgeNet [39] exploit the connectivity in mesh. However, these methods are often not robust to variations in mesh structure and lack of capability in capturing multi-scale features of meshes. Also it is difficult to implement HodgeNet [39] for GPU acceleration and thus its deployment is limited. Spectral convolution approach [14, 51] which constructs the equivalent convolution operation on spectral domain is another promising direction. However, such approach is limited by its low generalization to other shapes and inefficiency. LapCluster [35] exploits spectral clustering method to capture the mesh features, but the advanced structure of mesh still cannot be well captured. Recent method SubdivNet [24] tries to bring traditional geometric processing tools such as surface subdivision and simplification into deep CNN models, but it still suffers from highly complex and irregular mesh input and is incapable to capture the globally long-range contexts.

2.2. Transformer on Point Clouds

Transformer [44] and its variants have become the most leading models for various NLP tasks [8, 15, 44] since its first launch. Recently, Transformer models are transferred to 2D image domain [10, 16, 29, 46] and 3D vision domain [20, 50, 52]. Its powerful representation learning via self-similarity comparisons and the extracted global contextual information provide competitive or better performance than the long-standing CNN based models. As 3D point cloud data is an orderless representation, Transformer model inherently becomes the ideal candidate for processing orderless point cloud representation due to its inherent order-agnostic property. Point cloud based Transformer

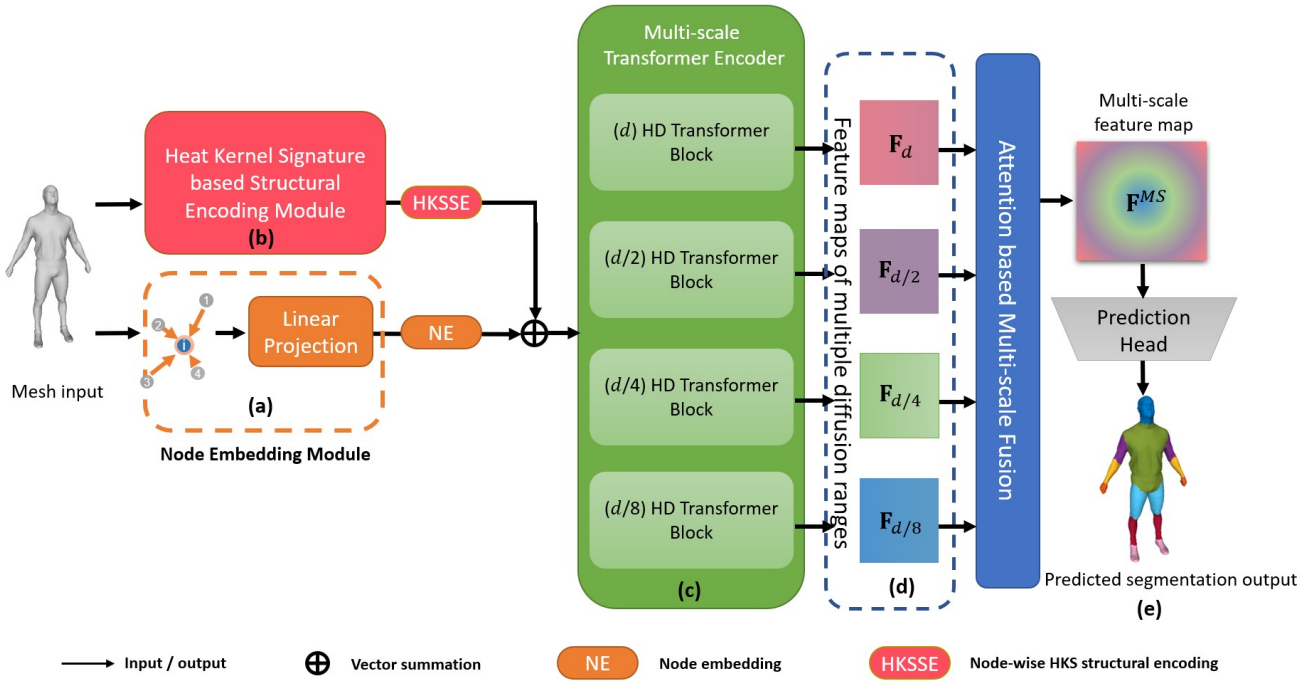

Figure 1. The entire architecture of MeshFormer.

model such as [20, 32, 50, 52] has been adopted as an effective model for processing 3D point cloud input due to its inherent capability in processing unordered point sets. Along such direction, it is natural to adapt the Transformer model to the mesh input, which is also one of the most common representations for 3D input modality. However, those methods [20, 32, 50, 52] do not provide effective ways to capture the multi-scale features of 3D input. Also, such methods only use a simple positional encoding, e.g. node embedding and thus lose the essential geometric structural information. The very recent work MeshMAE [27] applies the pre-training on a large-scale dataset, but it still suffers from limited performance on segmentation task since it only uses a patch embedding without the consideration on an effective geometric structural embedding.

2.3. Heat Diffusion Methods

Heat diffusion method has been widely adopted in numerical computation on 3D meshes for its capability in capturing the intrinsic surface properties. Seminal works such as [12, 13] had applied the heat diffusion mechanism to compute the geodesic distance on mesh in stable manner. And thus it is invaluable to exploit such approach to extract intrinsic locality on discretized surface. Heat Kernel Signatures (HKS) [40] and its follow-up works [3, 6] are proved to be a suitable shape descriptor to capture the geometric structural information, while bypassing the issue of eigenfunction sign ambiguity. Inspired by these promising methods, a novel MeshFormer model is proposed to integrate heat diffusion mechanism into Transformer model for capturing the intrinsic multi-scale surface features and geometric structural information for mesh based representation learning.

3. Method

In this section, the details of the proposed Transformer model: MeshFormer for 3D mesh segmentation are presented. The proposed MeshFormer comprises two novel essential modules: i) Heat Diffusion based Multi-head Self-Attention (HDMSA) for adaptively capturing the mesh features from local neighborhood to global contexts, ii) Heat Kernel Signature based Structure Encoding (HKSSE) applied to embed the intrinsic geometric structures into Transformer framework for structure-aware mesh processing. The entire architecture of MeshFormer is illustrated in Fig.1. The details of each proposed module are described in the following sections.

3.1. Extracting Mesh Locality from Heat Diffusion Method

In order to allow the communication with features in local neighborhood situated in diverse 3D mesh structures, the heat diffusion method [12] is applied to effectively extract the communication range on the intrinsic geometry of discretized surface [9]. The heat diffusion equation describes how a quantity, e.g. feature value f defined on a manifold

is propagated across the time period t,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\mathbf{x}, t) = \Delta f(\mathbf{x}, t) \tag{1}$$

The mesh geodesic distance starting from each reference vertex is used to measure the communication range, that is the locality. To compute such mesh geodesic distance, we need to first solve the heat diffusion equation. The resulting solution $f(\mathbf{x}, t)$ is called the heat kernel, which represents the heat distribution over time. To relate the heat kernel distribution $f(\mathbf{x}, t)$ with the desired geodesic distance function $d(\mathbf{x}, \mathbf{x}')$, the normalized gradient operation is first applied to heat kernel distribution $f(\mathbf{x}, t)$, then the resulting vector field $X = -\nabla f / |\nabla f|$ (along gradient descent direction) is plugged into the Poisson equation $\Delta d = \nabla \cdot X$, where the notation Δ and $\nabla\cdot$ denotes the Laplace-Beltrami operator and divergence operator, respectively. After solving such equation, we could obtain the desired geodesic distance d.

To convert the Eq. (1) into discretized setting in time, a backward Euler step $(f_t - f_0)/t \approx \Delta f_t$ is applied to obtain the following linear equation:

$$(\mathrm{Id} - t\Delta) f_t = f_0 \tag{2}$$

where Id denotes the identity matrix. f_0 and f_t represent the initial value and the discrete value at a short time step t, respectively. Since smaller backward Euler time step may leads to large discretization error, a more suitable strategy is to set the time step as the square of the average edge length of mesh h, i.e. $t = h^2$.

To further discretize the Eq. (2) in spatial setting, piecewise linear elements on triangle mesh based on finite element method (FEM) are exploited, and the Laplace-Beltrami operator in smooth setting is approximated by cotangent Laplacian [36] in discrete setting. Fig. 3 illustrates the different geodesic distances on a mesh computed by heat diffusion method mentioned above, which is then used to extract the locality region for the self-attention operation. As the maximum geodesic distance d_i starting from each source vertex v_i can be represented as its global communication range, the portion of the maximum geodesic distance is used to describe the locality range from heat diffusion, e.g. half of global range is denoted as $d_i/2$.

3.2. Heat Diffusion based Multi-head Self-Attention (HDMSA)

An input triangle mesh is presented as a pair (V, F), where V and F denotes a list of vertices $V = \{v_1, ..., v_n\}$ and a list of triangular faces $F = \{\mathrm{fa}_1, ..., \mathrm{fa}_m\}$, respectively. Each triangular face fa_j comprises the indices of three vertices for storing the triangle connectivity. We use $p(v_i) = p_i \in \mathbb{R}^3$ to denote the 3D position of each vertex.

Node Embedding Each input mesh is first evenly-sampled into M vertices. For each sampled vertex, a K-nearest neighbor (e.g. K=16) grouping is applied for creating a group of neighboring vertices, as illustrated in Fig.1(a). Such group, now considered as node, works as the basic processing element, analogous to the word token in Transformer for NLP task. As the raw coordinate of each node is only 3-dim, a linear projection layer is then applied on such group to obtain M node embeddings (NE) with higher dimensions. Note that the number of sampled vertices M is regarded as the sequence length of input for Transformer. However, as the input sequence length M in our case is generally larger than the one in NLP, the globally pair-wise comparison computations in general self-attention operation are unaffordable.

To tackle such issue, a novel Heat Diffusion based Multi-head Self-Attention (HDMSA) operation is proposed for limiting the comparison range while capturing the locality in 3D shape. The latter capability is also essential to learn the multi-scale features of mesh structure. The node embeddings are passed through the multi-scale Transformer encoder, which is composed of multiple heat diffusion (HD) Transformer blocks with diffusion regions from local neighborhood to global range, denoted as $d/2^k, (k = 0, 1, 2, 3)$, where d represents the maximum geodesic distance starting from each source vertex, as illustrated in Fig 1(c). An example of multiple diffusion ranges on a mesh example is illustrated in Fig.3.

HD Transformer block Each HD Transformer block comprises the core component, Heat Diffusion based Multi-head Self-Attention (HDMSA) module (will be described in the following section), and the general components in a standard Transformer block, such as a succeeding multi-layer perceptron (MLP) layer. The HDMSA module and MLP layer are all preceded by a LayerNorm (LN) layer for stabilizing the training process. And the residual connection is applied in the same style as a standard Transformer block. The whole HD Transformer block is illustrated in Fig. 2.

HD based Multi-head Self-Attention As the standard Transformer model [44] and the point cloud based variants [20,52] perform global self-attention by computing all pair-wise similarity comparisons among all the tokens, such intensive computation is unsuitable for dense prediction task. Instead, inspired by exploiting heat diffusion method to capture the intrinsic locality on the surface, a heat diffusion based multi-head self-attention is proposed to limit the similarity comparisons. The resulting feature maps in multiple diffusion ranges scales $(\mathrm{F}_{d/8}, \mathrm{F}_{d/4}, \mathrm{F}_{d/2}, \mathrm{F}_d)$ are further fused by an attention layer to obtain the multi-scale feature map F^{MS}, which is then attached with the prediction head

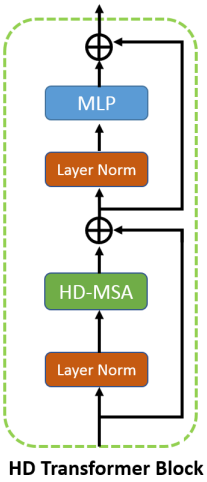

Figure 2. The construction of Heat Diffusion (HD) Transformer block.

to obtain the mesh semantic segmentation output, as illustrated in Fig. 1(c)-(e).

It is noted that no downsampling or pooling operation for mesh is required in our model, and consequently, the intrinsic geometric structure of mesh can be preserved to the highest degree.

3.3. Heat Kernel Signature based Structure Encoding (HKSSE)

To capture the structural information of a 3D mesh, it's natural to exploit the eigenvalues and the corresponding eigenfunctions derived from the Laplacian-Beltrami operator computed on the mesh data, like the recent methods such as SAN [25] and GNN-LSPE [17]. However, such approach cannot process the spectral information in a consistent manner and thus fails to provide sufficiently accurate structural information of input. In this work, instead of directly applying the spectral information like SAN [25] method which suffers from eigenfunction sign ambiguity issue, we further integrate a novel intrinsic geometric structural encoding module which exploits heat kernel signature [41] derived from the spectrum of 3D shape. This specific encoding module is called the heat kernel signature based structural encoding (HKSSE). The heat kernel signature [41] preserves all the intrinsic geometric information captured in heat kernels, which are the fundamental elements to describe the heat diffusion process on the manifold.

To describe the intrinsic geometric structure of each input mesh instance, the set of eigenvalues $\{\lambda_i\}$ and the corresponding eigenfunctions $\{\boldsymbol{\phi}_i\}$ of the cotangent Laplacian related to the mesh shape is first computed by eigendecomposition. It is noted that $\boldsymbol{\phi}_i = \{\phi_i(x)\}$ is a function defined on all the vertices ($x \in V$). Then the heat kernel signature $\mathrm{HKS}(x,t)$ for time scale t can be computed by the following equation:

$$\mathrm{HKS}(x,t) = \sum_{i=1}^{n} e^{-\lambda_i t}\phi_i^2(x) \tag{3}$$

where $t > 0$ denotes the heat diffusion temporal period and n is the number of vertices in a mesh. The HKS sums over all the spectrum, and applies a squared term on the eigenfunction. As a result, the spectral information of the shape are effectively captured without the eigenfunction sign ambiguity. And thus the set of all time-scale heat kernel signatures $\mathrm{HKS}(x) = \{\mathrm{HKS}(x,t), t \in \mathbb{R}^+\}$ fully characterizes the intrinsic structures of 3D shape in multi-scale manner. The Fig. 4 illustrates the comparison on a set of eigenfunctions and heat kernel signatures in multiple time scales. It is shown that heat kernel signatures are able to extract the more advanced geometric structures, such as the symmetric parts in human body, i.e. hands and legs, which is essential to be encoded into the Transformer model for geometric structure-aware processing for 3D mesh.

Then, the proposed encoding module HKSSE is attached to the Transformer model, together with node embedding, to build up the full MeshFormer model, as illustrated in Fig.1(b). As a result, the structure-specific inductive bias is captured for structure-aware mesh segmentation.

4. Experiments

Both the quantitative and qualitative experiments are conducted to verify the effectiveness of the proposed MeshFormer model in 3D mesh semantic segmentation task. The evaluations are performed on two diverse mesh based benchmark datasets, the HumanBody-Part [31] dataset and the COSEG [47] dataset. For measuring the model performance, we use the face based intersection over union (IoU) over parts, which is a widely-adopted evaluation metrics for mesh based segmentation task. Furthermore, the ablation studies on the core components of MeshFormer model is also provided.

4.1. Training Details

For training the MeshFormer model, AdamW [30] optimizer with default hyper-parameters is applied for optimizing the training loss. The initial learning rate is set to 0.001 with cosine annealing. All training settings are consistent for both HumanBody-Part [31] and COSEG [47] dataset. The experimental evaluations and ablation analysis were conducted using Nvidia 2080Ti GPU and the implementation of MeshFormer model is based on Pytorch library.

For comparison with point based methods, such as Point-Transfromer [52], and PCT [20] models, the mesh object is uniformly sampled into 4,096 points with the associated labels for training. For the accuracy evaluation, the nearest-face strategy is adopted to assign the label of each vertex

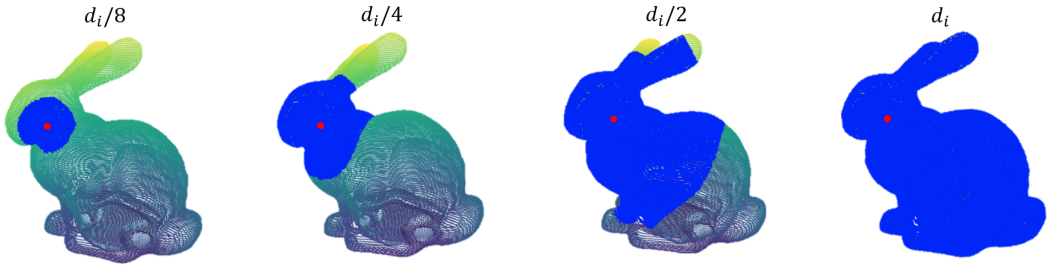

Figure 3. Multiple diffusion ranges (in blue color) on a mesh example (Stanford Bunny) from a source point i (in red color). The maximum geodesic distance d_i represents the global communication range for source point.

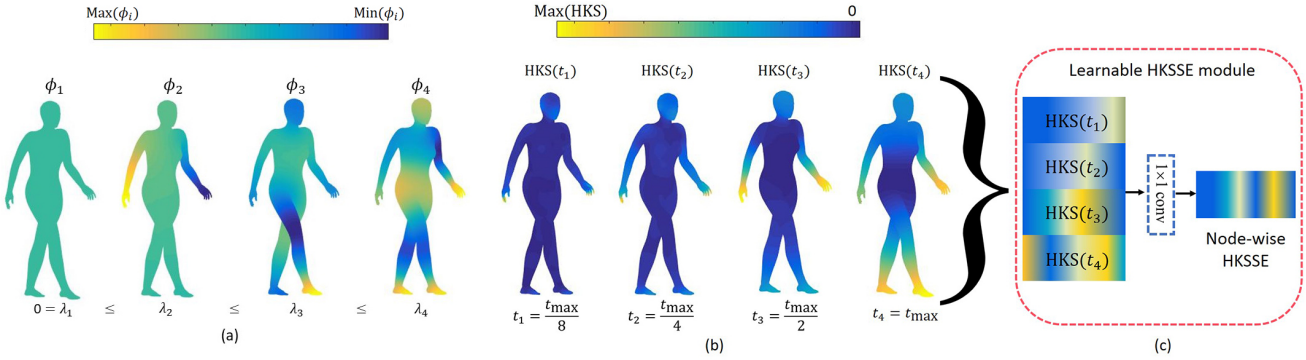

Figure 4. (a) Laplacian eigenfunctions of human body shape. (b) Heat kernel signatures in increasing time scales. (c) Learnable heat kernel signature based structural encoding (HKSSE) module.

to its nearest face. If the face is assigned with different labels among the three incident vertices, the majority voting is applied to choose a single label for the face.

4.2. Mesh Segmentation on HumanBody-Part Dataset

The HumanBody-Part dataset [31] for mesh segmentation evaluation is collected from SCAPE [2], FAUST [4], MIT [45], Adobe Fuse 2016 [1] and SHREC07-Human [18], and is annotated in 8 body parts. As this human body dataset contains meshes reconstructed from multiple scans on real human bodies in different persons and diverse poses, the resulting instances with a variety of mesh structures brings great challenges for accurate segmentation. For quantitative evaluation, the train/test split from MeshCNN [22] setting is used (381 mesh instances for train set and 18 mesh instances for test set).

Quantitative and Qualitative Results: In Table 1, the results of proposed MeshFormer model are displayed with comparisons with several state-of-the-art methods on Humanbody-Part dataset. The part-wise Intersection-over-Union (IoU) is used as the accuracy metric in quantitative evaluation. It is observed that the proposed MeshFormer provides the best accuracy with 94.2% score in part-wise IoU, outperforming both point-based and mesh-based models, especially the current Transformer models for 3D input (Point-Transformer [52], PCT [20]).

The qualitative segmentation results are illustrated in Fig.5(a). It is shown that the MeshFormer model provides accurate segmentation output with structure-aware property, and thus can handle symmetric parts well on the human bodies. Fig.5(b) provides the qualitative comparison to point based Transformer models(PCT, PCT+HKSSE). The predicted output of PCT suffers from ignoring the intrinsic geometric structural information and thus gives erroneous segmentation on symmetric body parts, e.g. hands, lower arms. After equipped with HKSSE to capture intrinsic geometric structures, the extended PCT gives accurate segmentation on symmetric parts. However, it still suffers from mis-segmentation around the joints, while MeshFormer provides coherent segmentation with the help of sufficient correlations captured by HDMSA.

4.3. Mesh Segmentation on COSEG Dataset

The proposed MeshFormer is also evaluated on the shape COSEG dataset [47], which is a mesh dataset with a wide variety of diverse shapes. Three largest categories in COSEG, such as vases, chairs, and tele-aliens are selected for the quantitative evaluation. These three categories contains 200, 400 and 300 instances respectively, and are an-

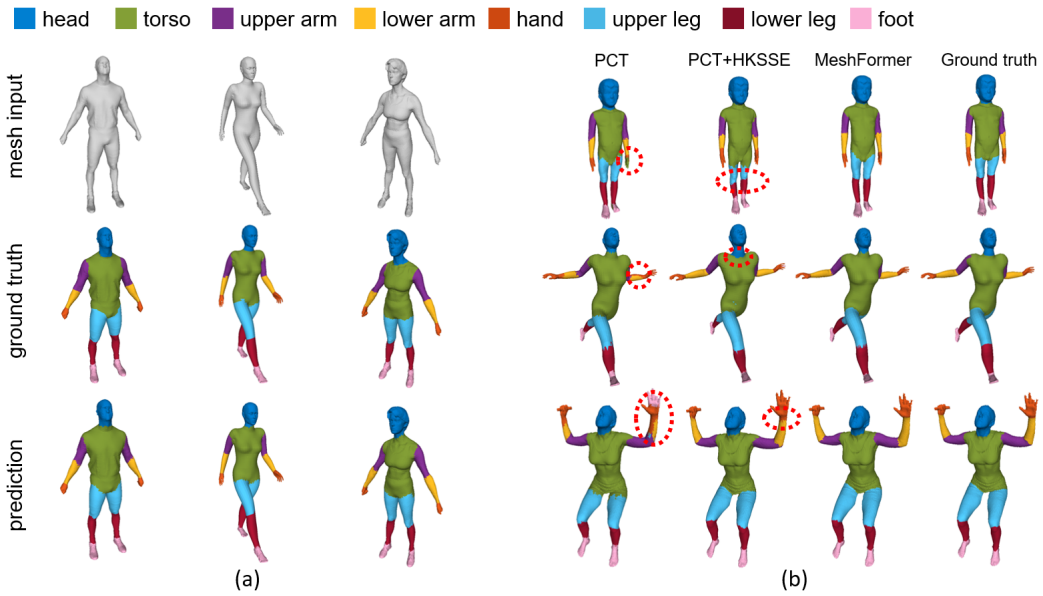

Figure 5. (a) Qualitative results of MeshFormer on the test set of HumanBody-Part [31] dataset. (b) Qualitative comparison to point based Transformer models(PCT, PCT+HKSSE). The dashed circles highlight the erroneous segmentations.

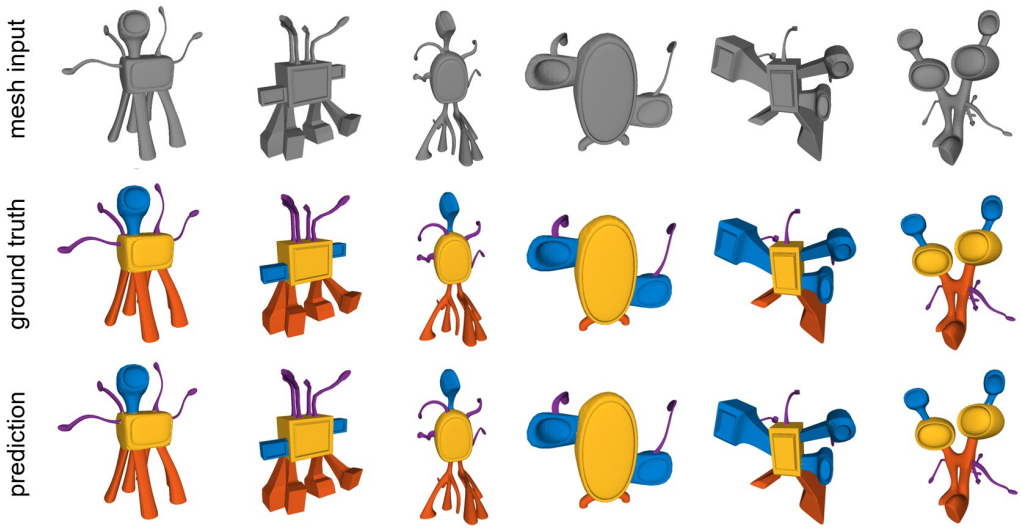

Figure 6. Qualitative results of MeshFormer on the test set of tele-aliens category in COSEG [47] dataset.

notated in 3 to 4 parts labels. The train/test split from SubdivNet [24] setting is used (randomly split the train/test set based on a ratio of $4:1$).

Quantitative and Qualitative Results: In Table 2, the quantitative evaluations of MeshFormer are given with comparisons to several state-of-the-art methods on COSEG dataset. The comparison results illustrated that the proposed MeshFormer model outperforms current state-of-the-art methods such as MeshCNN [22], PD-MeshNet [33], MeshMAE [27] and SubdivNet [24], as reported by the part-wise IoU for each category.

The qualitative results of mesh segmentation on COSEG dataset are illustrated in Fig. 6. The tele-aliens set is selected for qualitative evaluation since the shapes of this category are full of complex and diverse geometric structures. It is shown that the MeshFormer model can predict mesh-based segmentation outputs which are very close to the ground-truth annotations. Noted that the proposed model is able to capture the structural and multi-scale information on 3D mesh and thus provides accurate semantics predictions on mesh objects with diverse geometric structures.

Method	Point or Mesh	Accuracy
Pointnet [11]	point	74.7
Pointnet++ [34]	point	82.3
DGCNN [48]	point	89.7
Point Transformer [52]	point	91.4
PCT [20]	point	91.7
Toric Cover [31]	mesh	88.0
PD-MeshNet [33]	mesh	86.9
SNGC [21]	mesh	91.0
MeshCNN [22]	mesh	92.3
SubdivNet [24]	mesh	93.0
MeshFormer	mesh	**94.2**

Table 1. The accuracy of mesh segmentation evaluated on HumanBody-Part dataset [31]. Metric is part-wise IoU (%).

Method	Vases	Chairs	Tele-aliens
MeshCNN [22]	85.2	92.8	94.4
PD-MeshNet [33]	81.6	90.0	89.0
MeshMAE [27]	97.0	97.2	97.9
SubdivNet [24]	96.7	96.7	97.3
MeshFormer	**97.5**	**97.8**	**98.1**

Table 2. The accuracy of mesh segmentation evaluated on COSEG dataset [47]. Metric is part-wise IoU (%).

4.4. Ablation Analysis

Here, the ablation analyses are presented to analyze further the proposed MeshFormer model. To validate the efficacy of the key components, the analyses are conducted on both Humanbody-Part [31] and Tele-aliens set of COSEG [47] dataset. The ablation result is listed in Table 3, starting from (a) Baseline to (f) full MeshFormer model.

(1). The effectiveness of HDMSA module. The proposed self-attention based on heat diffusion mechanism enables the model to directly extract the multi-scale intrinsic surface features. For the comparison, general multi-head self-attention ((a), (c) and (d)) only captures the similarity information through the globally pairwise comparisons, and thus lacks of the capability to capture the multi-scale features in surface. As a consequence, it provides inferior accuracy in mesh based segmentation.

(2). The effectiveness of HKSSE module. The heat kernel signature based structural encoding (HKSSE) module integrates the intrinsic geometric structural information into Transformer model as an effective supplement of 3D shape inductive bias. Therefore, the resulting MeshFormer is able to provide structure-aware segmentation for mesh input. By removing HKSSE module ((a) and (b)), the performance is significantly decreased due to the loss of the geometric structural information. LSE ((c) and (e)) refers to using the raw Laplacian eigenvector as structure encoding.

	IoU %	
	(Humanbody-Part)	(COSEG-TA)
(a). Baseline	87.8	92.8
(b). Baseline + HDMSA	89.2	94.1
(c). Baseline + LSE	88.9	93.9
(d). Baseline + HKSSE	90.3	95.3
(e). Baseline + HDMSA + LSE	92.6	96.3
(f). Baseline + HDMSA + HKSSE	**94.2**	**98.1**

Table 3. The part-wise mIoU scores of all ablated variants, starting from (a) Baseline to (f) full MeshFormer model.

These two options show positive improvement for 3D shape segmentation, but are still lower than HKSSE ((d) and (f)), since HKS can extract the more advanced geometric structures and sidestep the eigenfunction sign ambiguity issue.

The mesh based part-wise IoU scores of all ablated variants are compared in Table 3. We can conclude that: i) The most important impact comes from the HKSSE module, since the geometric structural information is essential in mesh based segmentation, especially for objects with advanced symmetric structures. ii) The role of HDMSA shows the next important impact in performance, especially for objects with diverse shapes.

5. Conclusion

In this work, a novel mesh based Transformer model called MeshFormer is proposed which exploits the heat diffusion mechanism to tackle several challenges in semantic segmentation for mesh input with diverse shapes and complex geometric structures. The proposed MeshFormer integrates heat diffusion into multi-head self-attention operation to extract the multi-scale intrinsic surface features. It also applies a learnable heat kernel based structure encoding to facilitate the mesh based Transformer model to reinforce the geometric structural correctness in prediction to provide structure-aware segmentation output.

Through these two improvements, the segmentation results for 3D mesh based objects (especially the objects with diverse shapes and symmetric structures such as human bodies) have significant gains in accuracy. The performance of MeshFormer is validated in terms of part-wise IoU scores over two challenging benchmarks. From the experiments, MeshFormer outperforms both point based and mesh based segmentation methods. The experimental evaluations also validate the contributions of MeshFormer: i) heat diffusion integrated into multi-head self-attention is an effective method to capture the intrinsic surface property from 3D mesh; ii) a more accurate and structure-aware semantic segmentation with sufficient geometric correctness for mesh objects with diverse shapes and advanced symmetric structures; iii) better performance than current state-of-the-art Transformer models for 3D input (e.g., Point-Transformer, PCT and MeshMAE models).

References

[1] Adobe. 2016. Adobe fuse 3d characters. https://www.mixamo.com. 6

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005. 6

[3] William Benjamin, Andrew Wood Polk, SVN Vishwanathan, and Karthik Ramani. Heat walk: Robust salient segmentation of non-rigid shapes. In *Computer Graphics Forum*, volume 30, pages 2097–2106. Wiley Online Library, 2011. 3

[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. 6

[5] Davide Boscaini, Jonathan Masci, Emanuele Rodolà, and Michael Bronstein. Learning shape correspondence with anisotropic convolutional neural networks. *Advances in neural information processing systems*, 29, 2016. 2

[6] Michael M Bronstein and Iasonas Kokkinos. Scale-invariant heat kernel signatures for non-rigid shape recognition. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1704–1711. IEEE, 2010. 3

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 1

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[9] James J. Callahan. *Intrinsic Geometry*, pages 257–328. Springer New York, New York, NY, 2000. 1, 3

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[11] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, July 2017. 1, 8

[12] Keenan Crane, Weischedel, Clarisse, and Max Wardetzky. The heat method for distance computation. *Communications of the ACM*, 60(11):90–99, 2017. 1, 3

[13] Keenan Crane, Clarisse Weischedel, and Max Wardetzky. Geodesics in heat: A new approach to computing distance based on heat flow. *ACM Transactions on Graphics (TOG)*, 32(5):1–11, 2013. 1, 3

[14] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016. 2

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 1, 2

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021. 1, 2

[17] Vijay Prakash Dwivedi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. Graph neural networks with learnable structural and positional representations. In *International Conference on Learning Representations*, 2022. 2, 5

[18] Daniela Giorgi, Silvia Biasotti, and Laura Paraboschi. Shape retrieval contest 2007: Watertight models track. *SHREC competition*, 8(7), 2007. 6

[19] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):142–158, 2016. 2

[20] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R. Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, Apr 2021. 1, 2, 3, 4, 5, 6, 8

[21] Niv Haim, Nimrod Segol, Heli Ben-Hamu, Haggai Maron, and Yaron Lipman. Surface networks via general covers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 632–641, 2019. 8

[22] 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. 1, 2, 6, 7, 8

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. Conf. Comput. Vis. Pattern Recognit.*, pages 770–778, 2016. 2

[24] Shi-Min Hu, Zheng-Ning Liu, Meng-Hao Guo, Jun-Xiong Cai, Jiahui Huang, Tai-Jiang Mu, and Ralph R Martin. Subdivision-based mesh convolution networks. *ACM Transactions on Graphics (TOG)*, 41(3):1–16, 2022. 1, 2, 7, 8

[25] Devin Kreuzer, Dominique Beaini, Will Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 5

[26] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed

points. In *Advances in Neural Information Processing Systems*, pages 820–830, 2018. 1

[27] Yaqian Liang, Shanshan Zhao, Baosheng Yu, Jing Zhang, and Fazhi He. Meshmae: Masked autoencoders for 3d mesh data analysis. In *European Conference on Computer Vision*, 2022. 3, 7, 8

[28] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8895–8904, 2019. 1

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. 1, 2

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[31] Haggai Maron, Meirav Galun, Noam Aigerman, Miri Trope, Nadav Dym, Ersin Yumer, Vladimir G Kim, and Yaron Lipman. Convolutional neural networks on surfaces via seamless toric covers. *ACM Trans. Graph.*, 36(4):71–1, 2017. 2, 5, 6, 7, 8

[32] Kirill Mazur and Victor Lempitsky. Cloud transformers: A universal approach to point cloud processing tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10715–10724, 2021. 1, 2, 3

[33] Francesco Milano, Antonio Loquercio, Antoni Rosinol, Davide Scaramuzza, and Luca Carlone. Primal-dual mesh convolutional neural networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 7, 8

[34] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, pages 5099–5108, 2017. 1, 8

[35] Yi-Ling Qiao, Lin Gao, Paul Rosin, Yu-Kun Lai, Xilin Chen, et al. Learning on 3d meshes with laplacian encoding and pooling. *IEEE Transactions on Visualization and Computer Graphics*, 2020. 2

[36] Martin Reuter, Silvia Biasotti, Daniela Giorgi, Giuseppe Patanè, and Michela Spagnuolo. Discrete laplace–beltrami operators for shape analysis and segmentation. *Computers & Graphics*, 33(3):381–390, 2009. 4

[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. 1

[38] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, 2017. 2

[39] Dmitriy Smirnov and Justin Solomon. Hodgenet: learning spectral geometry on triangle meshes. *ACM Transactions on Graphics (TOG)*, 40(4):1–11, 2021. 2

[40] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Proceedings of the Symposium on Geometry Processing*, SGP '09, page 1383–1392, Goslar, DEU, 2009. Eurographics Association. 3

[41] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009. 5

[42] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 2

[43] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021. 1

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1, 2, 4

[45] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *ACM SIGGRAPH 2008 papers*, pages 1–9. 2008. 6

[46] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 568–578, October 2021. 1, 2

[47] Yunhai Wang, Shmulik Asafi, Oliver Van Kaick, Hao Zhang, Daniel Cohen-Or, and Baoquan Chen. Active co-analysis of a set of shapes. *ACM Transactions on Graphics (TOG)*, 31(6):1–10, 2012. 5, 6, 7, 8

[48] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 2019. 1, 8

[49] Chi-Chong Wong and Chi-Man Vong. Persistent homology based graph convolution network for fine-grained 3d shape segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7098–7107, 2021. 1

[50] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3

[51] Li Yi, Hao Su, Xingwen Guo, and Leonidas J Guibas. Syncspeccnn: Synchronized spectral cnn for 3d shape segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2282–2290, 2017. 2

[52] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H.S. Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16259–16268, October 2021. 1, 2, 3, 4, 5, 6, 8