Cross-Shape Attention for Part Segmentation of 3D Point Clouds

Marios Loizou†1, Siddhant Garg†2, Dmitry Petrov†2, Melinos Averkiou1 and Evangelos Kalogerakis2

1 University of Cyprus/CYENS CoE  2 University of Massachusetts Amherst

Abstract
We present a deep learning method that propagates point-wise feature representations across shapes within a collection for the purpose of 3D shape segmentation. We propose a cross-shape attention mechanism to enable interactions between a shape’s point-wise features and those of other shapes. The mechanism assesses both the degree of interaction between points and also mediates feature propagation across shapes, improving the accuracy and consistency of the resulting point-wise feature representations for shape segmentation. Our method also proposes a shape retrieval measure to select suitable shapes for cross-shape attention operations for each test shape. Our experiments demonstrate that our approach yields state-of-the-art results in the popular PartNet dataset.

CCS Concepts
• Computing methodologies → Shape representations; • Computer systems organization → Neural networks;

1. Introduction
Learning effective point-based representations is fundamental to shape understanding and processing. In recent years, there has been significant research in developing deep neural architectures to learn point-wise representations of shapes through convolution and attention layers, useful for performing high-level tasks, such as shape segmentation. The common denominator of these networks is that they output a representation for each shape point by weighting and aggregating representations and relations with other points within the same shape.

† Equal contribution.

In this work, we propose a cross-shape attention mechanism that enables interaction and propagation of point-wise feature representations across shapes of an input collection. In our architecture, the representation of a point in a shape is learned by combining representations originating from points in the same shape as well as other shapes. The rationale for such an approach is that if a point on one shape is related to a point on another shape e.g., they lie on geometrically or semantically similar patches or parts, then cross-shape attention can promote consistency in their resulting representations and part label assignments. We leverage neural attention to determine and weigh pairs of points on different shapes. We integrate these weights in our cross-shape attention scheme to learn...
more consistent point representations for the purpose of semantic shape segmentation.

Developing such a cross-shape attention mechanism is challenging. Performing cross-attention across all pairs of shapes becomes prohibitively expensive for large input collections of shapes. Our method learns a measure that allows us to select a small set of other shapes useful for such cross-attention operations with a given input shape. For example, given an input office chair, it is more useful to allow interactions of its points with points of another structurally similar office chair rather than a stool. During training, we maintain a sparse graph (Figure 1), whose nodes represent training shapes and edges specify which pairs of shapes should interact for training our cross-shape attention mechanism. At test time, the shape collection graph is augmented with additional nodes representing test shapes. New edges are added connecting them to training shapes for propagating representations from relevant training shapes.

We tested our cross-shape attention mechanism on two different backbones to extract the initial point-wise features per shape for the task of part segmentation: a sparse tensor network based on MinkowskiNet [CGS19] and the octree-based network MID-FC [WYZ*21]. For both backbones, we observed that our mechanism significantly improves the point-wise features for segmentation. Compared to the MinkowskiNet baseline, we found an improvement of 3.1% in mean part IoU in the PartNet benchmark [MZC*19a]. Compared to MID-FC, we found an improvement of +1.3% in mean part IoU, achieving a new state-of-the-art result in PartNet (MID-FC: 60.8% → Ours: 62.1%).

In summary, our main technical contribution is an attention-based mechanism that enables point-wise feature interaction and propagation within and across shapes for more consistent segmentation. Our experiments show state-of-the-art performance on the recent PartNet dataset.

2. Related work

We briefly overview related work on 3D deep learning for point clouds. We also discuss cross-attention networks developed in other domains.

2.1. 3D deep learning for processing point clouds

Several different types of neural networks have been proposed for processing point sets over the recent years. After the pioneering work of PointNet [QSMG17; QYSG17], several works further investigated more sophisticated point aggregation mechanisms to better model the spatial distribution of points [LCL18; SGS19; LKM19; LHC*20; DLD*21; ZCW*20; LLG*22; ZYX*22; QLP*22]. Alternatively, point clouds can be projected onto local views [SMKL15; QSN*16; KAMC17; HKC*17] and processed as regular grids through image-based convolutional networks. Another line of work converts point representations into volumetric grids [WSK*15; MS15; DCS*17; RWS*18; SWL19; LTLH19] and processes them through 3D convolutions. Instead of uniform grids, hierarchical space partitioning structures (e.g., kd-trees, octrees, lattices) can be used to define regular convolutions [RUG17; KL17; WLG*17; WSLT18; SJS*18; WYZ*21]. Another type of networks incorporate point-wise convolution operators to directly process point clouds [LBS*18; HTY18; XLCT18; LFM*19; GWL18; AML18; HRV*18; WSM*18; XFX*18; WQF19; KZH19; TQD*19]. Alternatively, shapes can be treated as graphs by connecting each point to other points within neighborhoods in a feature space. Then graph convolution and pooling operations can be performed either in the spatial domain [WSL*19; SFYT18; LYYD19; WSS18; ZHW*19; LFXP19; LB19; LMQ*23; JZL*19; XSW*19; WHH*19; HWW*19; LKM19; LAK20], or spectral domain [YSGG17; BM*15; MBMR16; MBM*17]. Attention mechanisms have also been investigated to modulate the importance of graph edges and point-wise convolutions [XLCT18; XSW*19; WHH*19; YHYJ19]. Graph neural network approaches have been shown to model non-local interactions between points within the same shape [WSL*19; LMQ*23; XSW*19; HWW*19]. Finally, several recent works [ZZJ*21; GCL*21; EBD21; ML21; YTR*22; XWL*21; PWT*22; LLL*22; YGX*23; SEH*23] introduced a variety of transformer-inspired models for point cloud processing tasks. None of the above approaches have investigated the possibility of extending attention across shapes. A notable exception are the methods by Wang et al. [WS19] and Cao et al. [CPBM21] that propose cross-attention mechanisms across given pairs of point cloud instances representing different transformations of the same underlying shape for the specific task of rigid registration. Our method instead introduces cross-attention across shapes within a large collection without assuming any pre-specified shape pairs. Our method aims to discover useful pairs for cross-shape attention and learns representations by propagating them within the shape collection. Our method shows that the resulting features yield more consistent 3D shape segmentation than several other existing point-based networks.

2.2. Cross-attention in other domains

Our method is inspired by recent cross-attention models proposed for video classification, image classification, keypoint recognition, and image-text matching. Wang et al. [WGGH18] introduced non-local networks that allow any image query position to perceive features of all the other positions within the same image or across frames in a video. To avoid huge attention maps, Huang et al. proposes a “criss-cross” attention module [HHW*19] to maintain sparse connections for each position in image feature maps. Cao et al. [CXL*19] simplifies non-local blocks with query-independent attention maps [CXL*19]. Lee et al. [LCH*18] propose cross-attention between text and images to discover latent alignments between image regions and words in a sentence. Hou et al. [HCB*19] models the semantic relevance between class and query feature maps in images through cross-attention to localize more relevant image regions for classification and generate more discriminative features. Sarlin et al. [SDMR19] learns keypoint matching between two indoor images from different viewpoints by leveraging self-attention and cross-attention to boost the receptive field of local descriptors and allow cross-image communication. Chen et al. [CFP21] propose cross-attention between multiscape representations for image classification. Finally, Doersch et al. [DGZ20] introduced a CrossTransformer model for few-shot learning on images. Give an unlabeled query image, their model computes local cross-attention similarities with a number of labeled images and then infers class membership.

Our method instead introduces attention mechanisms across 3D...
shapes. In contrast to cross-attention approaches in the above domains, we do not assume any pre-existing paired data. The usefulness of shape pairs is determined based on a learned shape compatibility measure.

3. Method

Given an input collection of 3D shapes represented as point clouds, the goal of our method is to extract and propagate point-based feature representations from one shape to another, and use the resulting representations for 3D semantic segmentation. To perform the feature propagation, we propose a Cross-Shape Attention (CSA) mechanism. The mechanism first assesses the degree of interaction between pairs of points on different shapes. Then it allows point-wise features on one shape to influence the point-wise features of the other shape based on their assessed degree of interaction. In addition, we provide a mechanism that automatically selects shapes (“key shapes”) to pair with an input test shape (“query shape”) to execute these cross-shape attention operations. In the following sections, we first discuss the CSA layer at test time (Section 3.1). Then we discuss our retrieval mechanism to find key shapes given a test shape (Section 3.2), our training (Section 3.3), test stage (Section 3.4), and finally our network architecture details (Section 3.5).

3.1. Cross-shape attention for a shape pair

The input to our CSA layer is a pair of shapes represented as point clouds: \( S_m = \{ p_i \}_{i=1}^{P_m} \) and \( S_n = \{ p_j \}_{j=1}^{P_n} \), where \( p_i, p_j \in \mathbb{R}^3 \) represent 3D point positions and \( P_m, P_n \) are the number of points for each shape respectively. Our first step is to extract point-wise features for each shape.

In our implementation, we experimented with two backbones for point-wise feature extraction: a sparse tensor network based on a modified version of MinkowskiNet [CGS19], and an octree-based network, called MID-FC [WYZ+21] (architecture details for the two backbones are provided in Section 3.5 and supplementary material). The output from the backbone is a per-point D-dimensional representation stacked into a matrix for each of the two shapes respectively: \( X_m \in \mathbb{R}^{P_m \times D} \) and \( X_n \in \mathbb{R}^{P_n \times D} \). The CSA layer produces new D-dimensional point-wise representations for both shapes:

\[
X_m' = f(X_m, X_n; \theta), \quad X_n' = f(X_n, X_m; \theta)
\]

where \( f \) is the cross-shape attention function with learned parameters \( \theta \) described in the next paragraphs.

Key and query intermediate representations. Inspired by Transformers [VSP+17], we first transform the input point representations of the first shape in the pair to intermediate representations, called “query” representations. The input point representations of the second shape are transformed to intermediate “key” representations. The keys will be compared to queries to determine the degree of influence of one point on another. Specifically, these transformations are expressed as follows:

\[
q_{m,j}^{(h)} = W_q^{(h)} x_{m,j}, \quad k_{n,j}^{(h)} = W_k^{(h)} x_{n,j}
\]

where \( x_{m,j} \) and \( x_{n,j} \) are point representations for the query shape \( S_m \) and key shape \( S_n \), \( W_q^{(h)} \) and \( W_k^{(h)} \) are \( D' \times D \) learned transformation matrices shared across all points of the query and key shape respectively, and \( h \) is an index denoting each different transformation (“head”). The dimensionality of the key and query representations \( D' \) is set to \([D/H] \) where \( H \) is the number of heads. These intermediate representations are stacked into the matrices \( Q_m^{(h)} \in \mathbb{R}^{P_m \times D'} \) and \( K_n^{(h)} \in \mathbb{R}^{P_n \times D'} \). Furthermore, the point representations of the key shape \( S_n \) are transformed to value representations as:

\[
v_{n,j}^{(h)} = W_{v}^{(h)} x_{n,j}
\]

where \( W_v^{(h)} \) is a learned \( D' \times D \) transformation shared across the points of the key shape. These are also stacked to a matrix \( V_n^{(h)} \in \mathbb{R}^{P_n \times D'} \).

Pairwise point attention. The similarity of key and query representations is determined through scaled dot product [VSP+17]. This provides a measure of how much one shape point influences the point on the other shape. The similarity of key and query representations is determined for each head as:

\[
A_{m,n}^{(h)} = \text{softmax} \left( \frac{Q_m^{(h)} \cdot (K_n^{(h)})^T}{\sqrt{D'}} \right)
\]

where \( A_{m,n}^{(h)} \in \mathbb{R}^{P_m \times P_n} \) is a cross-attention matrix between the two shapes for each head.

Feature representation updates. The cross-attention matrix is used to update the point representations for the query shape \( S_m \):

\[
\tilde{x}_{m,j}^{(h)} = \sum_{j=1}^{P_n} A_{m,n}^{(h)}(i,j) W_v^{(h)} x_{n,j}
\]

The point-wise features are concatenated across all heads, then a linear transformation layer projects them back to D-dimensional space and they are added back to the original point-wise features of the query shape:

\[
x_{m,j}' = x_{m,j} + W_d \left[ \tilde{x}_{m,j}^{(1)}, \tilde{x}_{m,j}^{(2)}, \ldots, \tilde{x}_{m,j}^{(H)} \right]
\]

where \( H \) is the number of heads and \( W_d \) is another linear transformation. The features are stacked into a matrix \( X_m' \in \mathbb{R}^{P_m \times D} \), followed by layer normalization [BKH16].

Self-shape attention. The pairwise attention of Equation 4 and update operation of Equation 5 can also be applied to a pair that consists of the shape and itself. In this case, the CSA layer is equivalent to Self-Shape Attention (SSA), enabling long-range interactions between shape points within the same shape.

Cross-shape attention for multiple shapes. We can further generalize the cross-shape operation in order to handle multiple shapes and also combine it with self-shape attention. Given a selected set of key shapes, our CSA layer outputs point representations for the query shape \( S_m \) as follows:

\[
x_{m,j}' = \sum_{n \in \{C(m),m\}} c(m,n) A_{m,n} V_n
\]

where \( C(m) \) is a set of key shapes deemed compatible for cross-shape attention with shape \( S_m \) and \( c(m,n) \) is a learned pairwise compatibility function between the query shape \( S_m \) and each key shape \( S_n \). The key idea of the above operation is to update point representations of the query shape \( S_m \) as a weighted average of
attention-modulated representations computed by using other key shapes as well as the shape itself. The compatibility function \( c(m, n) \) assesses these weights that different shapes should have for cross-shape attention. It also implicitly provides the weight of self-attention when \( \mathcal{S}_m = \mathcal{S}_n \).

**Compatibility function.** To compute the compatibility function, we first extract a global, \( D \)-dimensional vector representation \( y_m \) and \( y_n \) for the query shape \( \mathcal{S}_m \) and each key shape \( \mathcal{S}_n \) respectively through mean-pooling on their self-shape attention representations:

\[
y_{m}^{(\text{SSA})} = \text{avg}(X_{m, i}^{(\text{SSA})}) = \text{avg}_{i}(A_{m,n}V_m)
\]

(8)

\[
y_{n}^{(\text{SSA})} = \text{avg}(X_{n, i}^{(\text{SSA})}) = \text{avg}_{i}(A_{n,n}V_n)
\]

(9)

In this manner, the self-attention representations of both shapes provide cues for the compatibility between them expressed using their scaled dot product similarity \([\text{VSP}^*17]\):

\[
u_m = U_q y_{m}^{(\text{SSA})}
\]

\[
u_n = U_q y_{n}^{(\text{SSA})}
\]

\[
s(m, n) = \hat{u}_m \cdot \hat{u}_n^\top
\]

(10)

where \( U_q \) and \( U_k \) are learned \( D \times D \) transformations for the query and key shape respectively, and \( \hat{u}_m = u_m / ||u_m||, \hat{u}_n = u_n / ||u_n|| \).

The final compatibility function \( c(m, n) \) is computed as a normalized measure using a softmax transformation of compatibilities of the shape \( m \) with all other shapes in the set \( \mathcal{C}(m) \), including the self-compatibility:

\[
c(m, n) = \frac{\exp(s(m, n))}{\sum_{n \in \mathcal{C}(m)} \exp(s(m, n))}
\]

(11)

### 3.2. Key shape retrieval

To perform cross-shape attention, we need to retrieve one or more key shapes for each query shape. One possibility is to use the mea-
here we describe the two backbones (MinkNetHRNet, MID-FC).

### 3.4. Inference

During inference, we create a graph connecting each test shape with its K key shapes retrieved by the measure of Eq. 14. We then perform graph update and fine-tuning of our network is performed two times in our implementation.

### 3.5. Architecture

Here we describe the two backbones (MinkNetHRNet, MID-FC).

| Variant          | avg part IoU |
|------------------|--------------|
| MinkResUNet      | 46.8         |
| MinkHRNet        | 48.0         |
| MinkHRNetCSN-SSA | 48.7         |
| MinkHRNetCSN-K1  | 49.9         |
| MinkHRNetCSN-K2  | 49.7         |
| MinkHRNetCSN-K3  | 47.2         |
| MID-FC-SSA       | 61.8         |
| MID-FC-CSN-K1    | 61.9         |
| MID-FC-CSN-K2    | 61.9         |
| MID-FC-CSN-K3    | 62.0         |
| MID-FC-CSN-K4    | 62.1         |
| MID-FC-CSN-K5    | 62.0         |

**Table 1:** Ablation study for all our variants in PartNet.

**MinkNetHRNet.** The first backbone is a variant of the sparse tensor network based on MinkowskiNet [CGS19]. We note that our variant performed better than the original MinkowskiNet for 3D segmentation [CGS19], as discussed in our experiments. In a preprocessing step, we normalize the point clouds to a unit sphere and convert them to a sparse voxel grid (voxel size = 0.05). After two convolutional layers, the network branches into three stages inspired by the HRNet [WSC21], a network that processes 2D images in a multi-resolution manner. In our case, the first stage consists of three residual blocks processing the sparse voxel grid in its original resolution. The second stage downsamples the voxel grid by a factor of 2 and processes it through two other residual blocks. The third stage further downsamples the voxel grid by a factor of 2 and processes it through another residual block. The multi-resolution features from the three stages are combined into one feature map through upsampling following [WSC21]. The resulting feature map is further processed by a 1D convolutional block. The sparse voxel features are then mapped back to points as done in the original MinkowskiNet [CGS19]. Details about the architecture of this backbone are provided in the supplementary.

**MID-FC.** The second variant utilizes an octree-based architecture based on the MID-FC network [WYZ21]. This network also incorporates a three-stage HRNet [WSC21] to effectively maintain and merge multi-scale resolution feature maps. To implement this architecture, each point cloud is first converted into an octree representation with a resolution of 644. To train this network, a self-supervised learning approach is employed using a multi-resolution instance discrimination pretext task with ShapeNetCore55 [WYZ21]. The training process involves two losses: a shape instance discrimination loss to classify augmented copies of each shape instance and a point instance discrimination loss to classify the same points on the augmented copies of a shape instance. This joint learning approach enables the network to acquire generic shape and point encodings that can be used for shape analysis tasks. Finally, the pre-trained network is combined with two fully-connected layers and our CSA layer. During training for our segmentation task, the HRNet is frozen, while we train only the two fully-connected layers and CSA layer for efficiency reasons. Details about the architecture of this backbone are provided in the supplementary.
3.6. Implementation details

We train our Cross Shape Network for each object category of PartNet [MZC*19b] separately, using the standard cross entropy loss, for 200 epochs. We set the batch size equal to 8 for all variants (SSA, K=1,2,3). For optimization we use the SGD optimizer [Rud16] with a learning rate of 0.5 and momentum = 0.9. We scale learning rate by a factor of 0.5, whenever the loss of the hold-out validation split saturates (patience = 10 epochs, cooldown = 10 epochs). For updating the shape graph for the training and validation split, we measure the performance of the validation split in terms of Part IoU. If it reaches a saturation point (patience = 10 epochs, cooldown = 5 epochs), we load the best model up to that moment, based on Part IoU performance, and update the graph for both splits. The graph is updated twice throughout our training procedure. For all layers we use batch normalization [IS15] with momentum = 0.02, except for the CSA module, where the layer normalization [BKH16] is adopted. We also refer readers to our project page marios2019.github.io/CSN includes our code and trained models.

4. Results

We evaluated our method for fine-grained shape segmentation qualitatively and quantitatively. In the next paragraphs, we discuss the used dataset, evaluation metrics, comparisons, and an analysis considering the computation time and size of our CSA layer.

Dataset. We use the PartNet dataset [MZC*19b] for training and evaluating our method according to its provided training, validation, and testing splits. Our evaluation focuses on the fine-grained level of semantic segmentation, which includes 17 out of the 24 object categories present in the PartNet dataset. We trained our network and competing variants separately for each object category.

Evaluation Metrics. For evaluating the performance of our method and variants, we used the standard Part IoU metric, as also
Figure 4: Qualitative comparisons for a few characteristic test shapes of PartNet between the original MID-FC network for 3D shape segmentation ("MID-FC") [WYZ*21], and CrossShapeNet (CSN) in case of using self-shape attention alone ("MID-FC-CSN-SSA") and using cross-shape attention with $K = 4$ key shape per query shape ("MID-FC-CSN-K4"). The last column shows the key shapes and their ordering, retrieved for each test shape.

Table 2: Comparisons with other methods reporting performance in PartNet. The column “avg.” reports the mean Part IoU (averaged over all 17 categories). The last column “#cat” counts the number of categories that a method wins over others.

Ablation. Table 1 reports the mean part IoU performance averaged the PartNet’s part categories for the original backbone ("MinkHRNet") and ("MID-FC"). We first observe that our backbone variant “MinkHRNet” improves over the original “MinkResUNet” proposed in [CGS19], yielding an improvement of 1.2% in mean Part IoU. Our variant based on self-shape attention alone (“MinkHRNetCSN-SSA”) further improves our backbone by 0.7% in Part IoU. We further examined the performance of our cross-shape attention (CSA layer) tested in the variants “MinkHRNetCSN-K1”, “MinkHRNetCSN-K2”, and

Ground Truth | MID-FC | MID-FC-CSN-SSA | MID-FC-CSN-K4 | Key shapes
--- | --- | --- | --- | ---

M. Loizou et al. / Cross-Shape Attention for Part Segmentation of 3D Point Clouds

© 2023 The Authors. Computer Graphics Forum published by Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.
“MinkHRNet-CSN-K3”, where we use $K = 1, 2, 3$ key shapes per query shape. Our CrossShapeNet with $K = 1$ (“MinkHRNet-CSN-K1”) offers the best performance on average by improving Part IoU by another 1.2% with respect to using self-shape attention alone. When using $K = 2$ key shapes in cross-shape attention, the performance drops slightly (−0.2% in Part IoU on average) compared to using $K = 1$, and drops even more when using $K = 3$. Thus, for the MinkowskiNet variants, it appears that the optimal number of key shapes is $K = 1$; we suspect that the performance drop for higher $K$ is due to the issue that the chance of retrieving shapes that are incompatible to the query shape is increased with larger numbers of retrieved key shapes.

We also observe improvements using the MID-FC backbone. Note that this backbone has higher performance than the MinkowskiNet variants due to its pretrained and fine-tuning strategies [WYZ*21]. Our variant based on self-shape attention alone (“MID-FC-SSA”) further improves the original MID-FC backbone by 1.0% in mean Part IoU. When using cross-shape attention, the optimal performance is achieved when using $K = 4$ key shapes (“MID-FC-CSN-K4”), which improves Part IoU by another 0.3% with respect to using self-shape attention alone. We note that the above improvements are quite stable – by repeating all experiments 15 times, the standard deviation of mean Part IoU is $\sigma = 0.03\%$. This means that the above differences are significant – even the improvement of 0.3% of “MID-FC-CSN-K4” over “MID-FC-SSA” is of scale 106.

Comparisons with other methods. Table 2 includes comparisons with other methods reporting their performance on PartNet per category [XFX*18; QYSGi17; LMQ*23; LBS*18; LHC*20; WYZ*21] along with our best performing variants (“MinkHRNet-CSN-K1” and “MID-FC-CSN-K4”). Compared to the original MinkowskiNet (“MinkResUnet”), our “MinkHRNet-CSN-K1” variant achieves an improvement of 3.1% in terms of mean Part IoU in PartNet. Compared to “MID-FC”, our best variant (“MID-FC-CSN-K4”) also offers a noticeable improvement of 1.3% in mean Part IoU. To the best of our knowledge, the result of our best variant represents the highest mean Part IoU performance achieved in the PartNet benchmark so far. As it can been in the last column of Table 2, our method improves performance for 16 out of 17 categories.

Qualitative Results. Figures 3 shows qualitative comparisons for MinkowskiNet-based variants – specifically our best variant in this case using cross-shape attention with $K = 1$, self-shape attention, our MinkNetHR backbone, and the original MinkowskiNet. Our backbone often improves the labeling relative to the original MinkowskiNet (e.g., see bed mattress, or monitor base). Our cross-shape attention tends to further improve upon fine-grained details in the segmentation e.g., see the top of the bottle, the armrests in the chair, and the bottom of the blade, pushing the segmentation to be more consistent with the retrieved key shape shown in the inlet images. Figure 4 shows comparisons for the MID-FC-based variants, including using cross-shape attention with $K = 4$, self-shape attention, and the original MID-FC. We can drive similar conclusions – our method improves the consistency of segmentation especially for fine-grained details e.g., the lower bars of the table, the bottom

Number of parameters. Our CSA layer adds a relatively small overhead in terms of number of parameters. The MID-FC backbone has 1.8M parameters, the MinkHRNet has 24.8M parameters, while the CSA layer adds 0.4M parameters.

Computation. The cross-shape operation (Eq. 7) has linear time complexity wrt the number of used key shapes ($K$) and quadratic time complexity wrt the number of points per shape during both training and testing. To accelerate computation, subsets of points could be used as key points, as done in sparsified formulations of attention [ZGD*20; LSH*22]. The construction of the shape graph during training has quadratic time complexity wrt the number of the training shapes in the input collection. This is because the retrieval measure (Eq. 14) must be evaluated for all pairs of training shapes in our current implementation. For example, our CSA layer for $K = 1$ requires approximately 105 hours on a NVidia V100 to train for the largest PartNet category (3.5× more compared to training either backbone alone) due to the iterative graph construction and fine-tuning discussed in Section 3.3. A more efficient implementation could involve a more hierarchical approach e.g., perform clustering to select only a subset of training shapes as candidate key shapes for the retrieval measure. During inference, the time required per test shape exhibits linear complexity wrt the number of training shapes used for retrieval. In our experiments, testing time ranges from 0.7 sec (“Dish” class with the smallest number of shapes) to 7.8 sec (“Table” class with the largest number of shapes).

5. Conclusion

We presented a method that enables interaction of point-wise features across different shapes in a collection. The interaction is mediated through a new cross-shape attention mechanism. Our experiments show improvements of this interaction in the case of fine-grained shape segmentation.

Limitations. The performance increase comes with a higher computational cost at training and test time. It would be interesting to explore if further performance gains can be achieved through self-supervised pre-training [XGG*20; WYZ*21; SDR*22] that could involve a more hierarchical approach e.g., perform clustering to select only a subset of training shapes as candidate key shapes for the retrieval measure. During inference, the time required per test shape exhibits linear complexity wrt the number of training shapes used for retrieval. In our experiments, testing time ranges from 0.7 sec (“Dish” class with the smallest number of shapes) to 7.8 sec (“Table” class with the largest number of shapes).

Acknowledgments. This project has received funding from Adobe Research and the EU H2020 Research and Innovation Programme and the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy (GA 739578).

References

[AML18] ATZMON, MATAN, MARON, HAGGAI, and LIPMAN, YARON, “Point Convolutional Neural Networks by Extension Operators”. TOG 37.4 (2018) 2.

[BKH16] BA, JIMMY, KIROS, JAMIE RYAN, and HINTON, GEOFFREY E. “Layer Normalization”. arXiv:1607.06450 (2016) 3, 6.

[BMMS*15] ROSCAINI, DAVIDE, MASCIT JONATHAN, MELZL SIMONE, et al. “Learning class-specific descriptors for deformable shapes using localized spectral convolutional networks”. CGF 34.5 (2015) 2.
[MRB16] Boscazi, Davide, Maschi, Jonathan, Rodolà, Emanuele, and Bronstein, Michaël. "Learning shape correspondence with anisotropic convolutional neural networks". *Proc. NIPS*. 2016.

[CFP21] Chen, Chun-Fu, Richard Fan, Quanfu, and Panda, Rameswar. "CrossVT: Cross-Attention Multi-Scale Vision Transformer for Image Classification". *Proc. ICCV*. 2021 2.

[CGS19] Choy, Christopher, Gwak, JunYoung, and Savarese, Silvio. "3D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks". *Proc. CVPR*. 2019 2.

[CPBM21] Cao, Anh-Quan, Puy, Gilles, Boulch, Alexandre, and Marlet, Renaud. "PCAM: Product of Cross-Attention Matrices for Rigid Registration of Point Clouds". *Proc. ICCV*. 2021 2.

[CXL*19] Cao, Yue, Xu, Jiariu, Lin, Stephen, et al. "GCNet: Non-local networks meet squeeze-excitation networks and beyond". *Proc. CVPR Workshops*. 2019 2.

[DCS*17] Dal, Angela, Chang, Angel X, Savva, Manolis, et al. "ScanNet: Richly-annotated 3D reconstructions of indoor scenes". *Proc. CVPR*. 2017 2.

[DGZ20] Doersch, Carl, Gupta, Ankush, and Zisserman, Andrew. "Cross-transformers: spatially-aware few-shot transfer". *Proc. NIPS*. 2020 2.

[DLD*21] Deng, Congyou, Litany, Or, Duan, Yueqi, et al. "Vector Neurons: a general framework for SO(3)-equivariant networks". *Proc. ICCV*. 2021 2.

[EBD21] Engel, Nico, Belagianis, Vasileios, and Dietmayer, Klaus. "Point transformer". *IEEE Access*. 9 (2021) 2.

[GCL*21] Guo, Meng-Hao, Cai, Jun-Xiong, Liu, Zheng-Ning, et al. "Pct: Point cloud transformer". *CVM*. 7 2 (2021) 2.

[GWL18] Groh, Farijan, Wieschollek, Patrick, and Lensch, Hendrik P. A. "Flex-Convolution (Million-Scale Point-Cloud Learning Beyond Grid-Worlds)". *Proc. ACCV*. 2018 2.

[HCB*19] Hou, Rubing, Chang, Hong, Bingpeng, MA, et al. "Cross Attention Network for Few-shot Classification". *Proc. NIPS*. 2019 2.

[HKC*17] Huang, Haibin, Kalogerakis, Evangelos, Chaudhuri, Siddharttha, et al. "Learning Local Shape Descriptors from Part Correspondences with Multiview Convolutional Networks". *TOG*. 37 1 (2017) 2.

[HR*18] Hermosilla, Pedro, Ritschel, Tobias, Vázquez, Pere-Pau, et al. "Monte Carlo Convolution for Learning on Non-Uniformly Sampled Point Clouds". *TOG*. 37 6 (2018) 2.

[HTY18] Hua, Bin-Shon, Tran, Minh-Khoi, and Yeung, Sai-Kit. "Pointwise convolutional neural networks". *Proc. CVPR*. 2018 2.

[HWH*19] Huang, Zilong, Wang, Xinggang, Huang, Lichao, et al. "CCNet: Cross-attention for semantic segmentation". *Proc. CVPR*. 2019 2.

[HWW*19] Han, Wenkai, Wen, Chenlu, Wang, Cheng, et al. "Point2Node: Correlation Learning of Dynamic-Node for Point Cloud Feature Modeling". *arXiv:1912.10775* (2019) 2.

[IJS15] Ioffe, Sergey and Szegedy, Christian. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". *Proc. ICML*. 2015 6.

[JZL*19] Jiang, Li, Zhiao, Hengshuang, Liu, Shu, et al. "Hierarchical point-edge interaction network for point cloud semantic segmentation". *Proc. CVPR*. 2019 2.

[KAMC17] Kalogerakis, Evangelos, Avetikian, Melinos, Maji, Subhransu, and Chaudhuri, Siddharttha. "3D shape segmentation with projective convolutional networks". *Proc. CVPR*. 2017 2.

[KL17] Klokov, Roman and Lemptisky, Victor. "Escape from cells: Deep KD-networks for the recognition of 3d point cloud models". *Proc. CVPR*. 2017 2.

[KZH19] Komarichev, Artem, Zhong, Zichun, and Hua, Jing. "A-CNN: Annularly convolutional neural networks on point clouds". *Proc. CVPR*. 2019 2.

[LAK20] Loizou, Marios, Averkiou, Melinos, and Kalogerakis, Evangelos. "Learning Part Boundaries from 3D Point Clouds". *CFG*. 50 5 (2020) 2.

[LB19] Landrieu, Loïc and Boussaha, Mohamed. "Point cloud oversegmentation with graph-structured deep metric learning". *Proc. CVPR*. 2019 2.

[LBS*18] Li, Yangyang, Bu, Rui Sun, Mengchoa, et al. "PointCNN: Convolution On X-Transformed Points". *Proc. NIPS*. 2018 2.

[LCH*18] Lee, Kuang-Huei, Chen, Xi, Hua, Gang, et al. "Stacked cross attention for image-text matching". *Proc. ECCV*. 2018 2.

[LCL18] Li, Jiashin, Chen, Ben, and Lee, Gim. "SO-Net: Self-Organizing Network for Point Cloud Analysis". *Proc. CVPR*. 2018 2.

[LFM*19] Liu, Yongcheng, Fan, Bin, Meng, Gaideng, et al. "DensePoint: Learning Densely Contextual Representation for Efficient Point Cloud Processing". *Proc. ICCV*. 2019 2.

[LFXP19] Liu, Yongcheng, Fang, Bin, Xian, Shiming, and Pan, Chunhong. "Relation-shape convolutional neural network for point cloud analysis". *Proc. CVPR*. 2019 2.

[LHC*20] Liu, Ze, Hu, Han, Cao, Yue, et al. "A Closer Look at Local Aggregation Operators in Point Cloud Analysis". *Proc. CVCC*. 2020 2.

[LKM19] LE, Eric-Tuan, Kokkinos, Iasonas, and Mitra, Niloy J. "Going Deeper with Point Networks". *arXiv:1907.00960* (2019) 2.

[LGO*22] Luo, Shitong, Li, Jiahan, Guan, Jiagui, et al. "Equivariant Point Cloud Analysis via Learning Orientations for Message Passing". *Proc. CVPR*. 2022 2.

[LLJ*22] Lai, Xin, Liu, Jianhui, Jiang, Li, et al. "Stratified Transformer for 3D Point Cloud Segmentation". *Proc. CVPR*. 2022 2.

[LMQ*23] Li, Guo Hao, Muller, Matthias, Qian, Guocheng, et al. "DeepGCNs: Making GCNs Go as Deep as CNNs". *TPAMI*. 45 6 (2023) 2.

[LST*22] Lui, Difen, Shetty, Sandesh, Hinz, Tobias, et al. "AS-SET: Autoregressive Semantic Scene Editing with Transformers at High Resolutions". *TOG*. 41 4 (2022) 8.

[LTHL19] Lui, Zhihan, Tang, Haotian, Lin, Yujun, and Han, Song. "Point-Voxel CNN for efficient 3D deep learning". *Proc. NIPS*. 2019 2.

[LYYD19] Lan, Shiyl, Yu, Ruchi, Yu, Gang, and Davis, Larry S. "Modeling local geometric structure of 3D point clouds using GeoCNN". *Proc. CVPR*. 2019 2.

[MBM*17] Monti, Federico, Boscaini, Davide, Masci, Jonathan, et al. "Geometric deep learning on graphs and manifolds using mixture model cnns". *Proc. CVPR*. 2017 2.

[ML21] Mazur, Kirill and Lemptisky, Victor. "Cloud Transformers: A Universal Approach to Point Cloud Processing Tasks". *Proc. ICCV*. 2021 2.

[MS15] Maturana, Daniel and Scherer, Sebastian. "VoxNet: A 3D convolutional neural network for real-time object recognition". *Proc. IHOS*. 2015 2.

[MZC*19a] Mo, Kaichun, Zhu, Shilin, Chang, Angel X, et al. "PartNet: A Large-scale Benchmark for Fine-grained and Hierarchical Part-level 3D Object Understanding". *Proc. CVPR*. 2019 2.

[MZC*19b] Mo, Kaichun, Zhu, Shilin, Chang, Angel X, et al. "PartNet: A Large-Scale Benchmark for Fine-Grained and Hierarchical Part-Level 3D Object Understanding". *Proc. CVPR*. 2019 5.

[PWT*22] Pang, Yatian, Wang, Wenxiaoy, Yao, Francis E. H., et al. "Masked Autoencoders for Point Cloud Self-supervised Learning". *arXiv*:2203.06604 (2022) 2.
