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

Marios Loizou†1
Melinos Averkiou1
1University of Cyprus / CYENS CoE

Siddhant Garg†2
Evangelos Kalogerakis2
2University of Massachusetts Amherst

Dmitry Petrov†2

† Equal Contribution
Goal: learn more coordinated feature representations
Goal: learn more coordinated feature representations

test shape

training collection
Goal: learn more coordinated feature representations

*test shape*

*training collection*

Cross-ShapeNeXt
Goal: learn more coordinated feature representations
Goal: learn more coordinated feature representations

test shape

Cross-ShapeNet

retrieved shapes

test shape segmentation

training collection
**Prior work**: Point-based networks

PointNet++ [Qi et al. 2017]
**Prior work**: Point-based networks

- **PointNet**++ [Qi et al. 2017]
- **PointNeXt** [Qian et al. 2022]
- **CloserLook3D** [Liu et al. 2020]
Prior work: GCNs for non-Euclidean data

DGCNN [Wang et al. 2019]

DeepGCNs [Li et al. 2023]
Prior work: Volumetric networks

O-CNN [Wang et al. 2017]

MinkowskiNet [Choy et al. 2019]
Prior work: Attention is All You Need

Transformer [Vaswani et al. 2017]

PointTransformer v1/v2 [Zhao et al. 2021, 2022]
Why use **attention** for 3D representations?

Encode points such that their features capture **relations** wrt the rest of the shape.
Why use **attention** for 3D representations?

Encode points such that their features capture **relations** wrt the rest of the shape.
Why use **attention** for 3D representations?

Encode points such that their features capture **relations** wrt the rest of the shape.
Why use **attention** for 3D representations?

Encode points such that their features capture **relations** wrt the rest of the shape.
Why use **attention** for 3D representations?

Encode points such that their features capture **relations** wrt the rest of the shape.
Why use **attention** for 3D representations?
Why use **attention** for 3D representations?

- **low attention**
- **key points**
- **query point**
- **high attention**

Attention scores range from 0 to 1.
Why use **attention** for 3D representations?
Why use **attention** for 3D representations?

**new query representation** = **attention** representations

**key value representations**

**key points**

**query point**

**attention scores**
Why use **attention** for 3D representations?

- **query** representation from multiple attention heads
- **Add & Norm**
- **Feed-forward**
- **final query representation**

**query** point

**key points**

Attention scores

0

1
Motivation: Long-range interactions across shapes

No interactions across shapes
Motivation: Long-range interactions across shapes

test shape

Cross-ShapeNet

shapes from input collection

output part labels
**Key challenge**: Retrieve compatible shapes
Key challenge: Retrieve compatible shapes

Shape Collection

Cross-ShapeNet

test shape
Key challenge: Retrieve compatible shapes

Shape Collection

Cross-ShapeNet

test shape
Key challenge: Combine multiple shapes
Key challenge: Combine multiple shapes
**Key challenge**: Combine multiple shapes

[Diagram showing cross-shape attention with test shapes and retrieved shapes]
Pipeline

Shape Collection
Pipeline

Shape Collection
Pipeline

Shape Collection
Pipeline

Shape Collection

Cross-shape attention
Pipeline

Shape Collection

Cross-shape attention
Shape Collection

Cross-shape attention
Pipeline

Shape Collection

Cross-shape attention
Cross-Shape Attention

**query** shape $S_m = \{p_i\}_{i=1}^M$

**key** shape $S_n = \{p_j\}_{j=1}^N$
Cross-Shape Attention

**query** shape $S_m = \{p_i\}_{i=1}^M$

$X_m \in \mathbb{R}^{M \times D}$

**key** shape $S_n = \{p_j\}_{j=1}^N$

$X_n \in \mathbb{R}^{N \times D}$
Cross-Shape Attention

\[ X_m \in \mathbb{R}^{M \times D} \]

Backbone point representations

\[ X_n \in \mathbb{R}^{N \times D} \]

Transformer [Vaswani et al. 2017]
Cross-Shape Attention

\[ X_m \in \mathbb{R}^{M \times D} \]

\[ W_Q \in \mathbb{R}^{D \times D} \]

\[ Q_m = W_Q \cdot X_m \]

\[ X_n \in \mathbb{R}^{N \times D} \]

\[ W_K \in \mathbb{R}^{D \times D} \]

\[ K_n = W_K \cdot X_n \]

Transformer [Vaswani et al. 2017]
Cross-Shape Attention

\[ X_m \in \mathbb{R}^{M \times D} \]

\[ W_Q \in \mathbb{R}^{D \times D} \]

\[ Q_m \in \mathbb{R}^{M \times D} \]

Backbone point representations

\[ X_n \in \mathbb{R}^{N \times D} \]

\[ W_K \in \mathbb{R}^{D \times D} \]

\[ K_n \in \mathbb{R}^{N \times D} \]

Intermediate representations

Transformer [Vaswani et al. 2017]
Cross-Shape Attention

Backbone point representations

\[ X_m \in \mathbb{R}^{M \times D} \]

Query Transformation

\[ W_Q \in \mathbb{R}^{D \times D} \]

Intermediate representations

\[ Q_m \in \mathbb{R}^{M \times D} \]

Key Transformation

\[ W_K \in \mathbb{R}^{D \times D} \]

\[ K_n \in \mathbb{R}^{N \times D} \]

Value Transformation

\[ W_V \in \mathbb{R}^{D \times D} \]

\[ V_n = W_V \cdot X_n \]

Transformer [Vaswani et al. 2017]
Cross-Shape Attention

Backbone point representations

Intermediate representations

Key-value representations

Transformer [Vaswani et al. 2017]

\[ \mathbf{X}_m \in R^{M \times D} \]

\[ \mathbf{W}_Q \in R^{D \times D} \]

\[ \mathbf{Q}_m \in R^{M \times D} \]

\[ \mathbf{X}_n \in R^{N \times D} \]

\[ \mathbf{W}_K \in R^{D \times D} \]

\[ \mathbf{K}_n \in R^{N \times D} \]

\[ \mathbf{W}_V \in R^{D \times D} \]

\[ \mathbf{V}_n \in R^{N \times D} \]
Cross-Shape Attention

Transformer [Vaswani et al. 2017]
Cross-Shape Attention

\[ Q_m \in \mathbb{R}^{M \times D} \]
\[ K_n \in \mathbb{R}^{N \times D} \]

Transformer [Vaswani et al. 2017]
Cross-Shape Attention

\[ Q_m \in \mathbb{R}^{M \times D} \]
\[ K_n \in \mathbb{R}^{N \times D} \]

\[ \forall i = 1, \ldots, M \]
\[ \forall j = 1, \ldots, N \]

Transformer [Vaswani et al. 2017]
Cross-Shape Attention

\[
\text{softmax} \left( \frac{Q_m^T K_n}{\sqrt{D}} \right) = A_{m,n} \in R^{M \times N}
\]

Transformer [Vaswani et al. 2017]
Cross-Shape Attention

\[ A_{m,n} \cdot V_n = X_m^{(CSA)} \in \mathbb{R}^{M \times D} \]

- **Cross-shape attention matrix**
- **Key shape value representations**
- **Cross-shape attention representations**
Cross-Shape Attention

\[ A_{m,n} \cdot V_n = X_{m}^{(CSA)} \in \mathbb{R}^{M \times D} \]

- **Cross-shape attention matrix**
- **Key shape value representations**
- **Cross-shape attention representations**

**query shape**

**key shape**
Cross-Shape Attention for multiple shapes

query shape

key shape
Cross-Shape Attention for multiple shapes
Cross-Shape Attention for multiple shapes

query shape...

key shapes $\mathcal{C}(m)$
Cross-Shape Attention for multiple shapes

- $C(m)$: set of compatible key shapes
- $c(m, n)$: compatibility function between query shape $S_m$ and key shape $S_n$

Cross-shape attention output

$$X'_m = \sum_{n \in \{C(m), m\}} c(m, n)A_{m,n}V_n$$
Compatibility function

\[ X'_{m}^{(SSA)} \in \mathbb{R}^{M \times D} \]

\[ X'_{n}^{(SSA)} \in \mathbb{R}^{N \times D} \]
Compatibility function

\[ X_m'(SSA) \in \mathbb{R}^{M \times D} \]

\[ \text{avg}_{i} X_{m,i}'(SSA) \]

\[ y_m(SSA) \in \mathbb{R}^{D} \]

\[ X_n'(SSA) \in \mathbb{R}^{N \times D} \]

\[ \text{avg}_{i} X_{n,i}'(SSA) \]

\[ y_n(SSA) \in \mathbb{R}^{D} \]
Compatibility function

\[ X'_{m}^{(SSA)} \in \mathbb{R}^{M \times D} \]

\[ \text{avg}_{i} X'_{m,i}^{(SSA)} \rightarrow \]

\[ y_{m}^{(SSA)} \in \mathbb{R}^{D} \]

\[ U_{Q} \in \mathbb{R}^{D \times D} \]

\[ u_{m} = U_{Q} y_{m}^{(SSA)} \]

\[ X'_{n}^{(SSA)} \in \mathbb{R}^{N \times D} \]

\[ \text{avg}_{i} X'_{n,i}^{(SSA)} \rightarrow \]

\[ y_{n}^{(SSA)} \in \mathbb{R}^{D} \]

\[ U_{K} \in \mathbb{R}^{D \times D} \]

\[ u_{n} = U_{K} y_{n}^{(SSA)} \]
Compatibility function

\[ \mathbf{u}_m \in \mathbb{R}^D \]

\[ \hat{\mathbf{u}}_m = \mathbf{u}_m / \|\mathbf{u}_m\| \]

\[ \mathbf{u}_n \in \mathbb{R}^D \]

\[ \hat{\mathbf{u}}_n = \mathbf{u}_n / \|\mathbf{u}_n\| \]

Cosine similarity

\[ s(m, n) = \hat{\mathbf{u}}_m \cdot \hat{\mathbf{u}}_n \]
Compatibility function

\[ \hat{u}_m \rightarrow s(m, n_1) \rightarrow \hat{u}_{n_1} \]

query shape
Compatibility function

query shape

\[ \hat{u}_m \]

\[ \hat{u}_{n_1} \]

\[ \hat{u}_{n_k} \]

key shapes \( C(m) \)
Compatibility function

query shape

$\hat{u}_m$

$s(m, m)$

$s(m, n_1)$

$s(m, n_k)$

$\hat{u}_{n_1}$

$\hat{u}_{n_k}$

key shapes $C(m)$
Compatibility function

\[ c(m, n) = \frac{e^{s(m, n)}}{\sum_{n \in \mathcal{C}(m)} e^{s(m, n)}} \]
Cross-Shape Attention for multiple shapes

Cross-shape attention

key shapes
Retrieve compatible shapes

Shape Collection

Cross-ShapeNet

query shape
Retrieve compatible shapes

Shape Collection

Cross-ShapeNet

query shape
Key shape retrieval
Key shape retrieval

$X'_n^1$  

$X'_n^2$  

$X'_n^3$  

$X'_n^4$  

$X'_n^5$  

$X'_n^6$  

$X'_n^7$  

$X'_n^8$
Key shape retrieval
Key shape retrieval

Cosine similarity

\[ S_{m,n_k} = X'_{m} \cdot (X'_{n_k})^\top \]
Key shape retrieval

\[ r_i(m, n_k) = \max_j S_{m,n_k}[i,j] \]
Key shape retrieval

\[ X'_{n_0}(SSA) \]

\[ X'_{n_1}(SSA) \]

\[ X'_{n_2}(SSA) \]

\[ X'_{n_3}(SSA) \]

\[ X'_{n_4}(SSA) \]

\[ X'_{n_5}(SSA) \]

\[ X'_{n_6}(SSA) \]

\[ X'_{n_7}(SSA) \]

\[ X'_{n_8}(SSA) \]

\[ X'_{m}(SSA) \]

\[ r(m, n_k) = \text{avg} r_i(m, n_k) \]
Key shape retrieval

\[ r(m, n_k) = \text{avg} r_i(m, n_k) \]
Key shape retrieval

\[ r(m, n_k) = \text{avg} r_i(m, n_k) \]
Key shape retrieval: Examples

query shapes

key shapes
PartNet dataset

[Mo et al. 2019]
PartNet dataset

[Mo et al. 2019]
Examples of shape collections

Table category

Chair category

5,707 training shapes

4,489 training shapes
Training details: Loss

\[ L_{CE} = - \sum_{p_i \in S_k} \hat{q}_i \log q_i \]

- \( S_k \): shape \( k = \{ p_i \}_{i=1}^{P_k} \)
- \( \hat{q}_i \): ground-truth one-hot label vector for point \( p_i \)
- \( q_i \): predicted label probabilities for point \( p_i \)
Training details: Backbones

MinkowskiNet [Choy et al. 2019]

MID-FC [Wang et al. 2021]
Training details: Backbones

HRNet [Wang et al. 2021]
Training details: Collection graph

Shape Collection
Training details: Collection graph
Training details: Collection graph

Collection graph
Training details: Collection graph

Collection graph
Inference: Collection graph

Collection graph

test shape
Inference: Collection graph

Collection graph

test shape
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Part IoU</th>
</tr>
</thead>
</table>

Results: MinkowskiNet variants

<table>
<thead>
<tr>
<th>Method</th>
<th>Part IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinkHRNet</td>
<td>48.0</td>
</tr>
</tbody>
</table>
**Results**: MinkowskiNet variants

<table>
<thead>
<tr>
<th>Method</th>
<th>Part IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinkHRNet</td>
<td>48.0</td>
</tr>
<tr>
<td>MinkHRNetCSN-SSA</td>
<td>48.7</td>
</tr>
</tbody>
</table>

+0.7%
Results: MinkowskiNet variants

<table>
<thead>
<tr>
<th>Method</th>
<th>Part IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinkHRNet</td>
<td>48.0</td>
</tr>
<tr>
<td>MinkHRNetCSN-SSA</td>
<td>48.7</td>
</tr>
<tr>
<td>MinkHRNetCSN-K1</td>
<td>49.9</td>
</tr>
<tr>
<td>MinkHRNetCSN-K2</td>
<td>49.7</td>
</tr>
</tbody>
</table>
**Results:** MinkowskiNet variants

<table>
<thead>
<tr>
<th>Method</th>
<th>Part IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinkHRNet</td>
<td>48.0</td>
</tr>
<tr>
<td>MinkHRNetCSN-SSA</td>
<td>48.7</td>
</tr>
<tr>
<td>MinkHRNetCSN-K1</td>
<td>49.9</td>
</tr>
<tr>
<td>MinkHRNetCSN-K2</td>
<td>49.7</td>
</tr>
</tbody>
</table>

+ 1.2%
Results: MinkowskiNet variants

Ground truth

MinkHRNet

- Pillow
- Mattress
- Stretcher
- Leg
- Horizontal bar
- Vertical bar
- Bed post
- Ladder vertical bar
- Rung
Results: MinkowskiNet variants

Ground truth

MinkHRNet

MinkHRNetCSN-SSA

- Pillow
- Mattress
- Stretcher
- Leg
- Horizontal bar
- Vertical bar
- Bed post
- Ladder vertical bar
- Rung
Results: MinkowskiNet variants

Ground truth

MinkHRNetCSN-SSA

MinkHRNetCSN-K1

- Pillow
- Mattress
- Stretcher
- Leg
- Horizontal bar
- Vertical bar
- Bed post
- Ladder vertical bar
- Rung
### Results: MID-FC variants

<table>
<thead>
<tr>
<th>Method</th>
<th>Part IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MID-FC</td>
<td>60.8</td>
</tr>
</tbody>
</table>
## Results: MID-FC variants

<table>
<thead>
<tr>
<th>Method</th>
<th>Part IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MID-FC</td>
<td>60.8</td>
</tr>
<tr>
<td>MID-FC-CSN-SSA</td>
<td>61.8</td>
</tr>
</tbody>
</table>

+1.0%
## Results: MID-FC variants

<table>
<thead>
<tr>
<th>Method</th>
<th>Part IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MID-FC</td>
<td>60.8</td>
</tr>
<tr>
<td>MID-FC-CSN-SSA</td>
<td>61.8</td>
</tr>
<tr>
<td>MID-FC-CSN-K1</td>
<td>61.9</td>
</tr>
<tr>
<td>MID-FC-CSN-K2</td>
<td>61.9</td>
</tr>
<tr>
<td>MID-FC-CSN-K3</td>
<td>62.0</td>
</tr>
<tr>
<td><strong>MID-FC-CSN-K4</strong></td>
<td><strong>62.1</strong></td>
</tr>
<tr>
<td>MID-FC-CSN-K5</td>
<td>62.0</td>
</tr>
</tbody>
</table>

+0.3%
Results: MID-FC variants
Results: MID-FC variants

Ground truth

MID-FC

MID-FC-CSN-SSA

- Bar
- Leg
- Board
- Shelf
Results: MID-FC variants

MID-FC-CSN-SSA

MID-FC-CSN-K4

Ground truth

Bar
Leg
Board
Shelf
## Results: Comparison with other methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Part IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResGCN-28 (Li et al. 2023)</td>
<td>45.1</td>
</tr>
<tr>
<td>CloserLook3D (Liu et al. 2020)</td>
<td>53.8</td>
</tr>
<tr>
<td>MinkResUNet (Choy et al. 2019)</td>
<td>46.8</td>
</tr>
<tr>
<td>MinkHRNetCSN-K1 (ours)</td>
<td>49.9</td>
</tr>
<tr>
<td>MID-FC (Wang et al. 2021)</td>
<td>60.8</td>
</tr>
<tr>
<td>MID-FC-CSN-K4 (ours)</td>
<td>62.1</td>
</tr>
</tbody>
</table>
**Results**: Comparison with other methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Part IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResGCN-28 (Li et al. 2023)</td>
<td>45.1</td>
</tr>
<tr>
<td>CloserLook3D (Liu et al. 2020)</td>
<td>53.8</td>
</tr>
<tr>
<td>MinkResUNet (Choy et al. 2019)</td>
<td>46.8</td>
</tr>
<tr>
<td>MinkHRNetCSN-K1 (ours)</td>
<td>49.9</td>
</tr>
<tr>
<td>MID-FC (Wang et al. 2021)</td>
<td>60.8</td>
</tr>
<tr>
<td>MID-FC-CSN-K4 (ours)</td>
<td><strong>62.1</strong></td>
</tr>
</tbody>
</table>
### Results: Comparison with other methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Part IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResGCN-28 (Li et al. 2023)</td>
<td>45.1</td>
</tr>
<tr>
<td>CloserLook3D (Liu et al. 2020)</td>
<td>53.8</td>
</tr>
<tr>
<td>MinkResUNet (Choy et al. 2019)</td>
<td>46.8</td>
</tr>
<tr>
<td>MinkHRNetCSN-K1 (ours)</td>
<td>49.9</td>
</tr>
<tr>
<td>MID-FC (Wang et al. 2021)</td>
<td>60.8</td>
</tr>
<tr>
<td>MID-FC-CSN-K4 (ours)</td>
<td><strong>62.1</strong></td>
</tr>
</tbody>
</table>

+1.3%
### Results: Comparison with other methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Part IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResGCN-28 (Li et al. 2023)</td>
<td>45.1</td>
</tr>
<tr>
<td>CloserLook3D (Liu et al. 2020)</td>
<td>53.8</td>
</tr>
<tr>
<td>MinkResUNet (Choy et al. 2019)</td>
<td>46.8</td>
</tr>
<tr>
<td>MinkHRNetCSN-K1 (ours)</td>
<td>49.9</td>
</tr>
<tr>
<td>MID-FC (Wang et al. 2021)</td>
<td>60.8</td>
</tr>
<tr>
<td>MID-FC-CSN-K4 (ours)</td>
<td>62.1</td>
</tr>
</tbody>
</table>

SOTA performance on the PartNet dataset

+1.3%
Summary

- Enable long range point feature interactions across shapes
Summary

- Enable long range point feature interactions across shapes
- Introduce a novel cross-shape attention mechanism
Summary

- Enable long range point feature interactions **across shapes**
- Introduce a **novel cross-shape attention** mechanism
- Retrieve **compatible shapes** for cross-shape attention
Summary

• Enable long range point feature interactions across shapes
• Introduce a novel cross-shape attention mechanism
• Retrieve compatible shapes for cross-shape attention
• SOTA performance on PartNet
Summary

Limitations:

- **Increased computational cost** due to shape retrieval
Limitations:

- **Increased computational cost** due to shape retrieval
- Currently no support for **multi-object scenes**
Thank you!

Acknowledgements:

Our project web page:
https://marios2019.github.io/CSN/