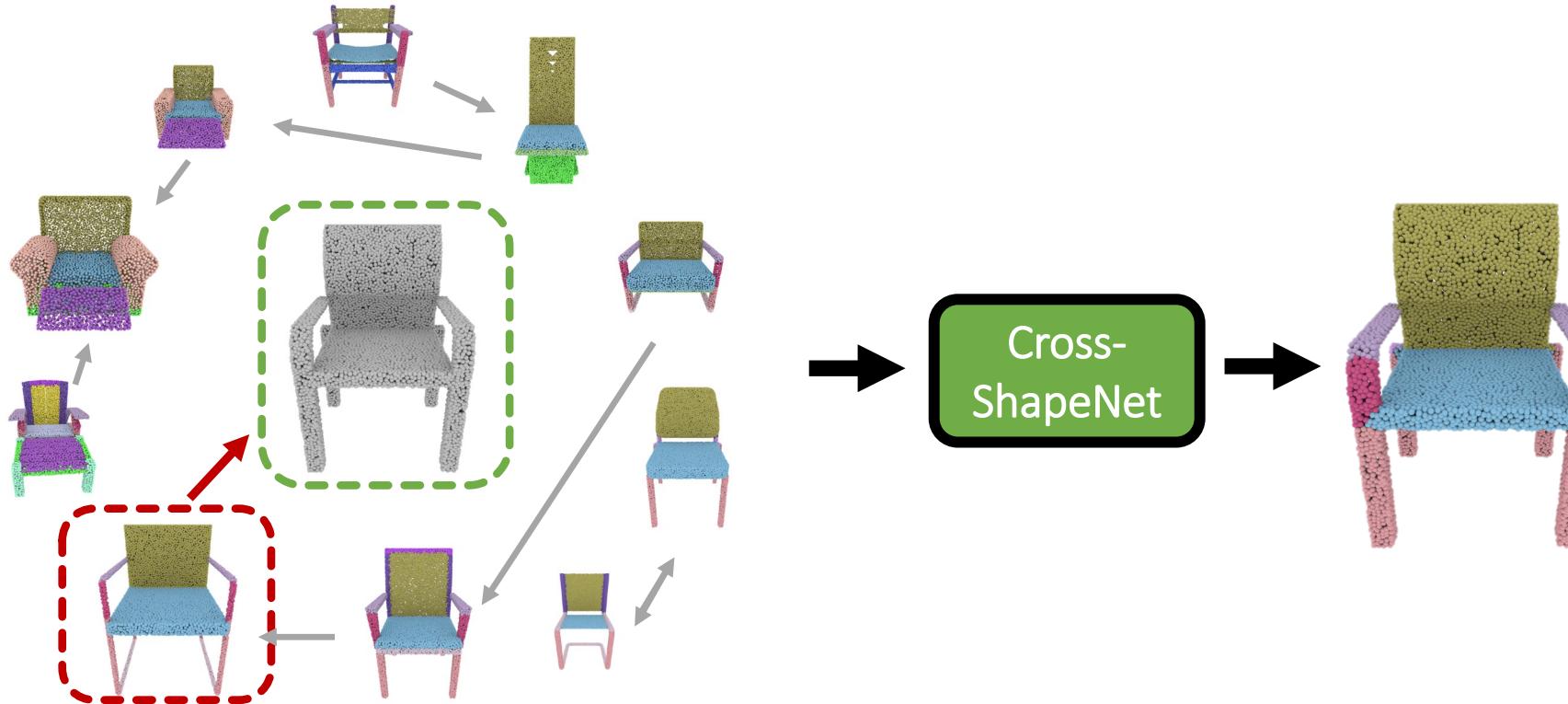


Cross-Shape Attention for Part Segmentation of 3D Point Clouds



Marios Loizou^{†1}

Melinos Averkiou¹

Siddhant Garg^{†2}

Evangelos Kalogerakis²

Dmitry Petrov^{†2}

¹University of Cyprus / CYENS CoE

²University of Massachusetts Amherst

Goal: learn more coordinated feature representations



test shape

Goal: learn more coordinated feature representations



test shape



training collection

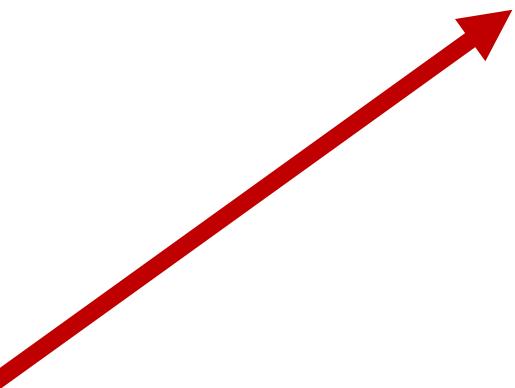
Goal: learn more coordinated feature representations



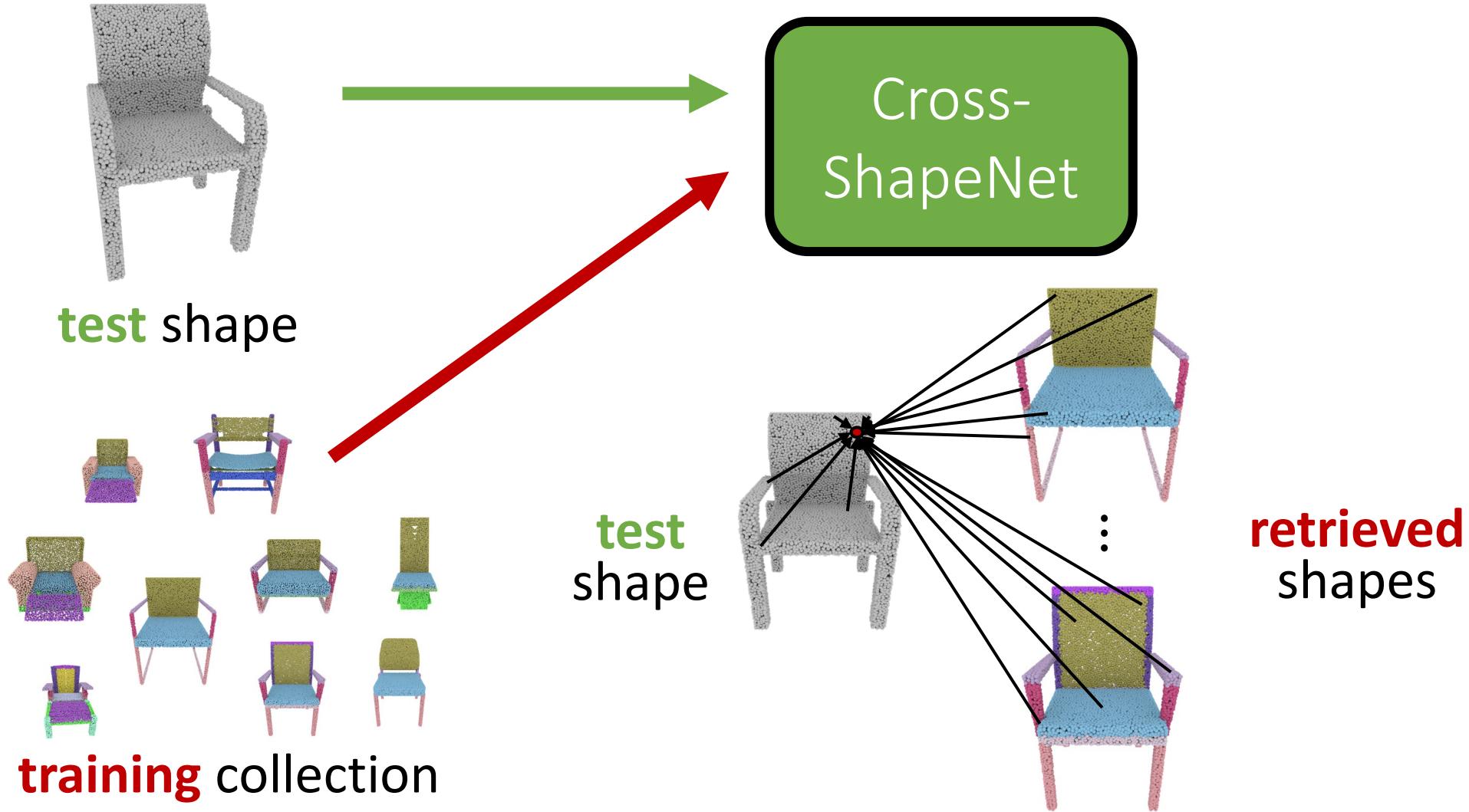
test shape



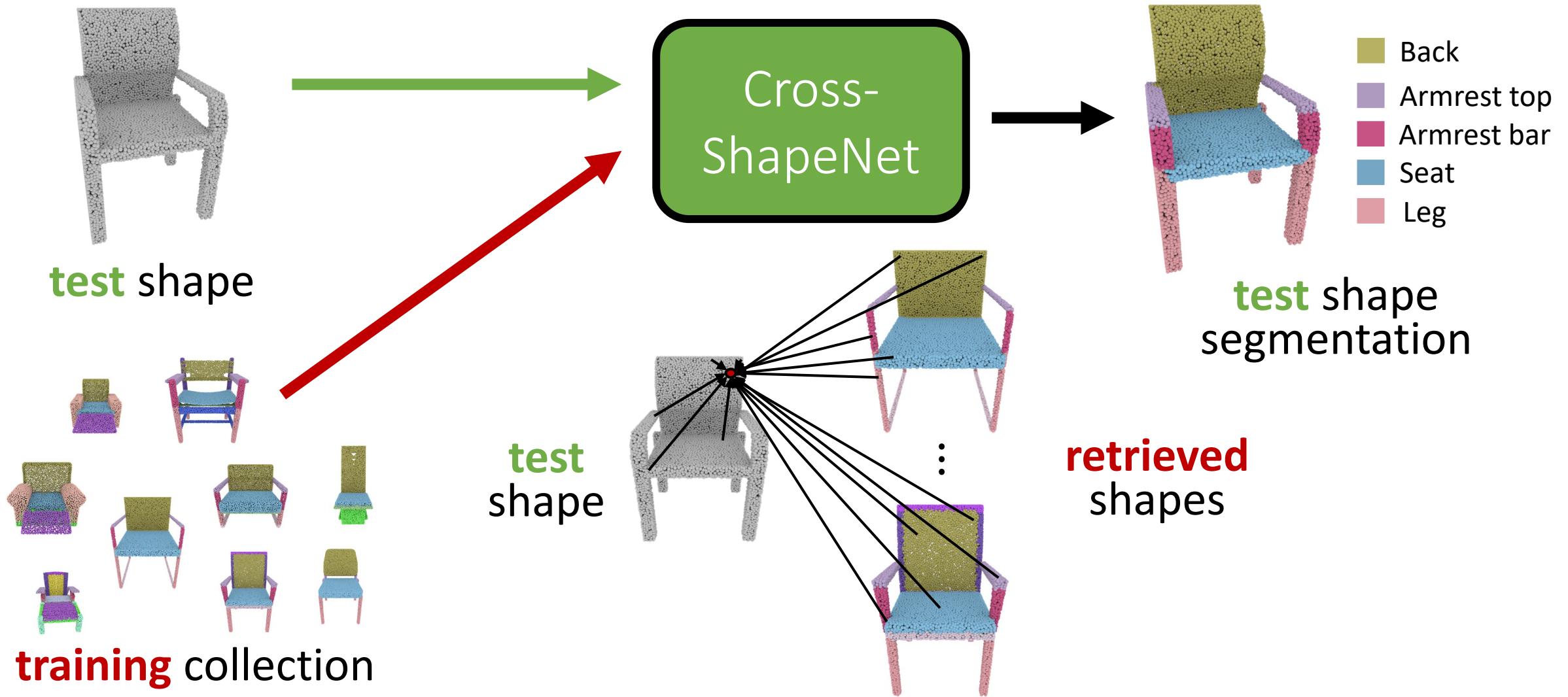
training collection



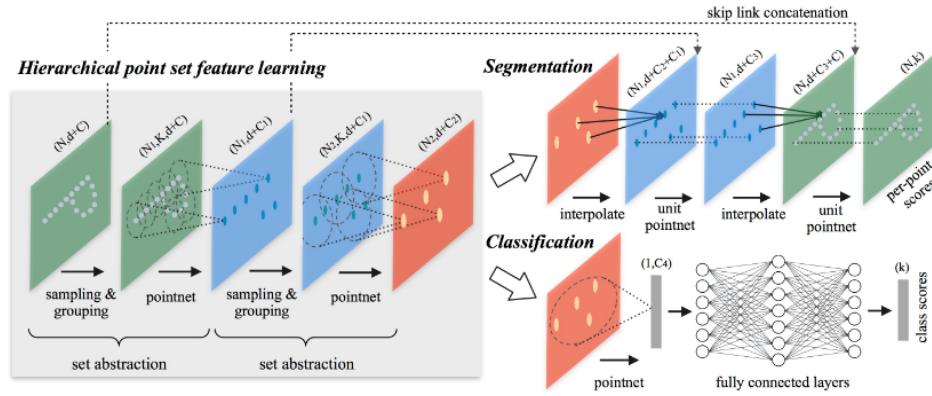
Goal: learn more coordinated feature representations



Goal: learn more coordinated feature representations

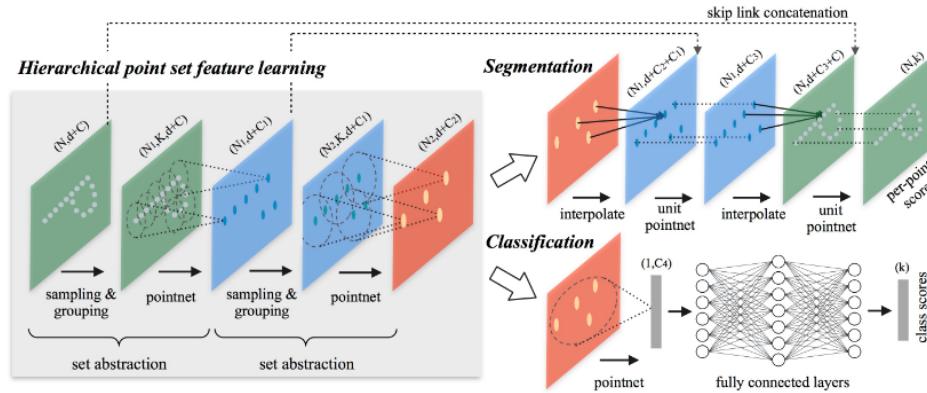


Prior work: Point-based networks

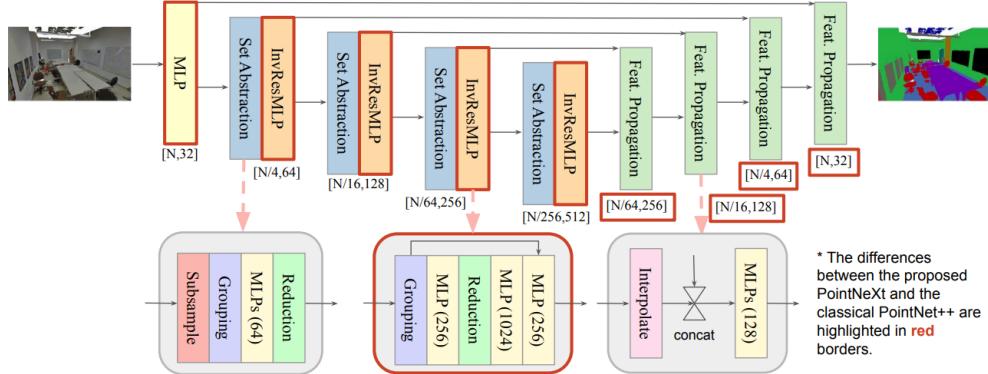


PointNet++ [Qi et al. 2017]

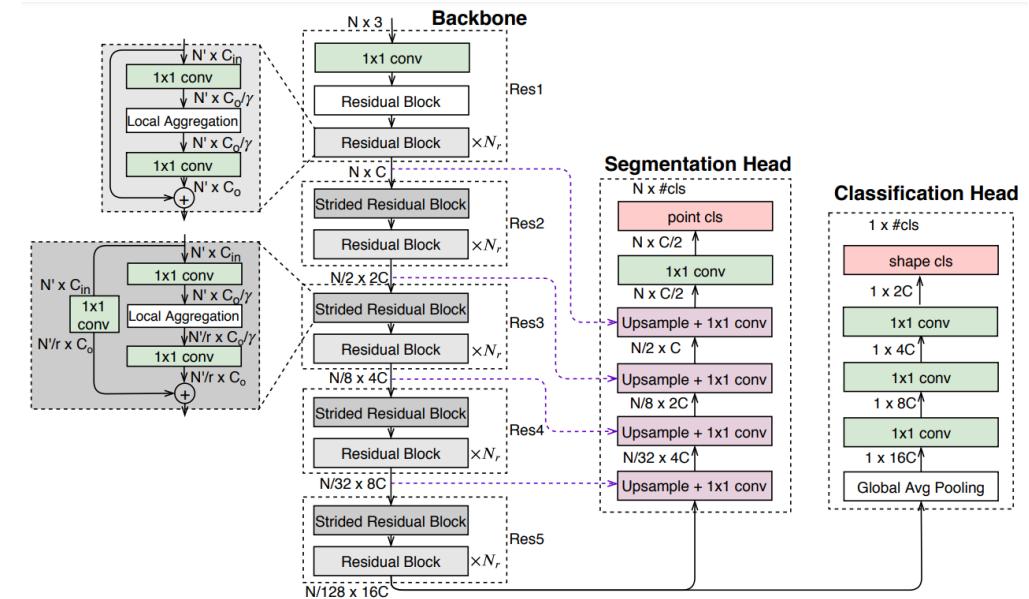
Prior work: Point-based networks



PointNet++ [Qi et al. 2017]

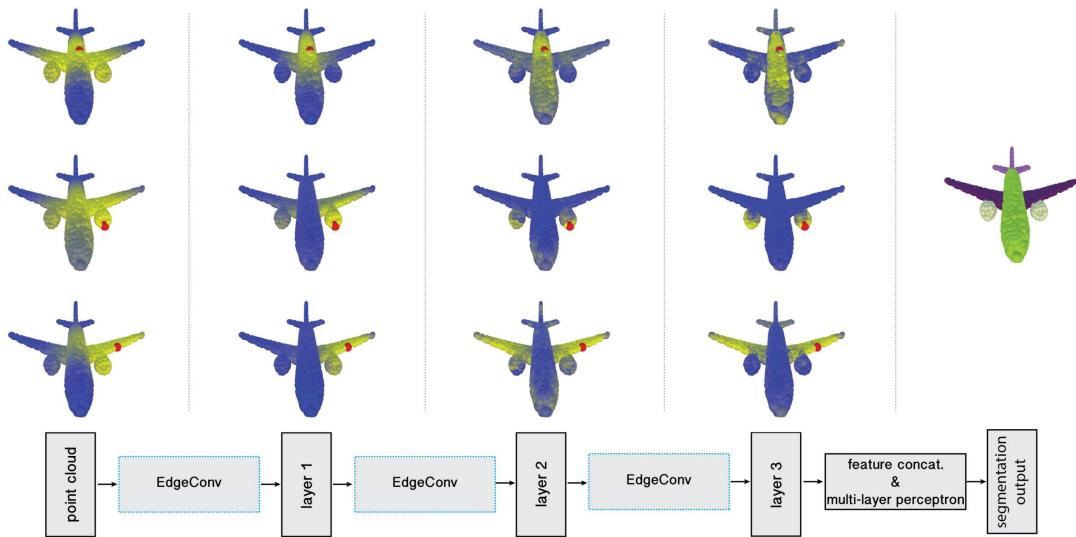


PointNeXt [Qian et al. 2022]

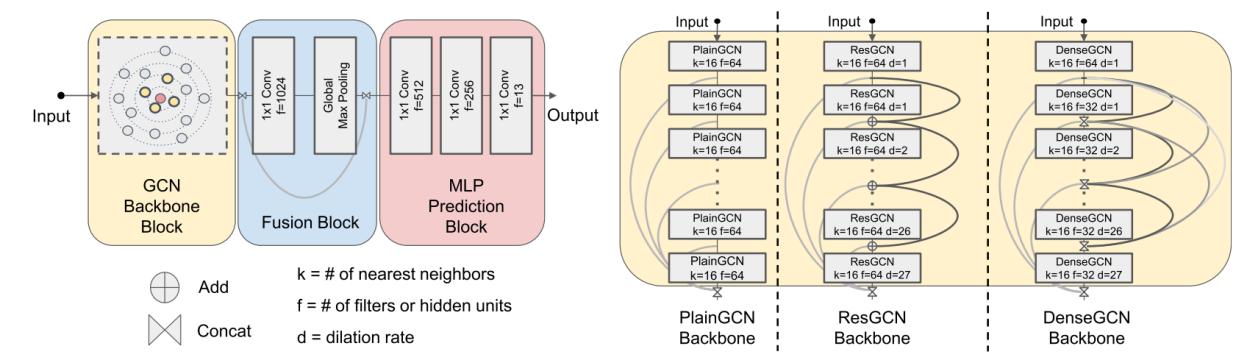


CloserLook3D [Liu et al. 2020]

Prior work: GCNs for non-Euclidean data

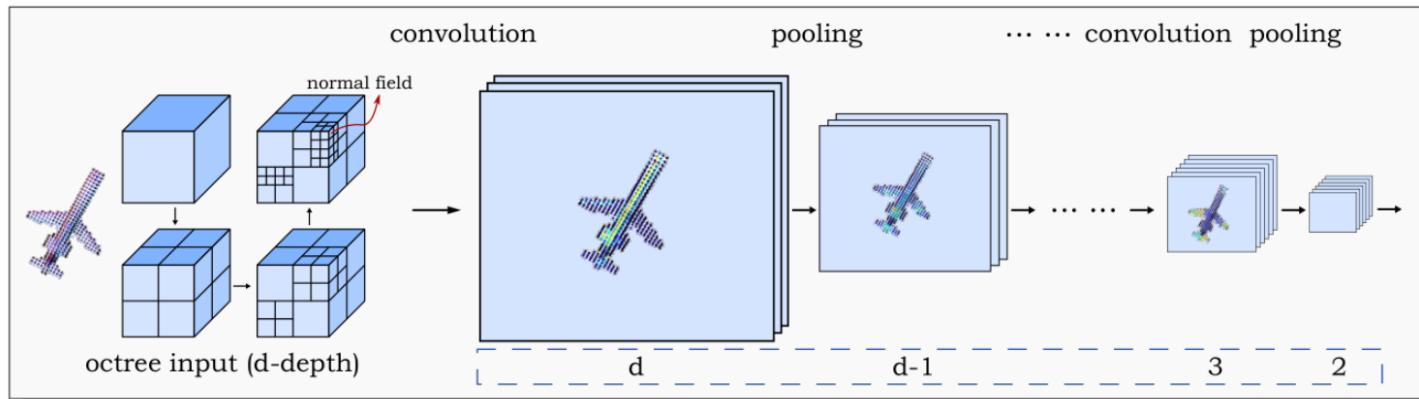


DGCNN [Wang et al. 2019]



DeepGCNs [Li et al. 2023]

Prior work: Volumetric networks

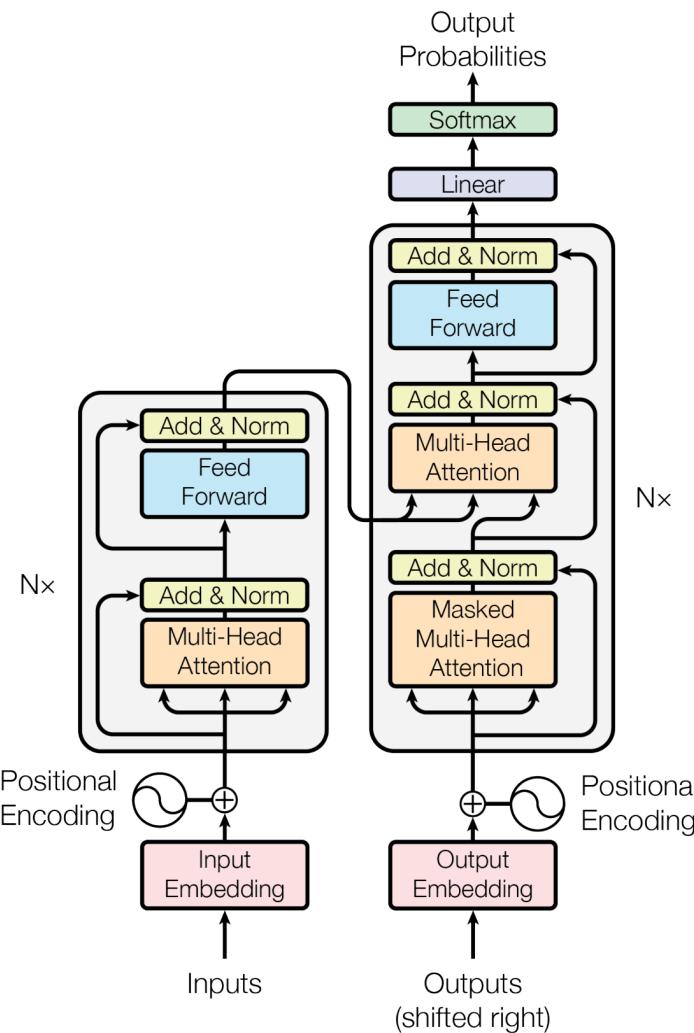


O-CNN [Wang et al. 2017]

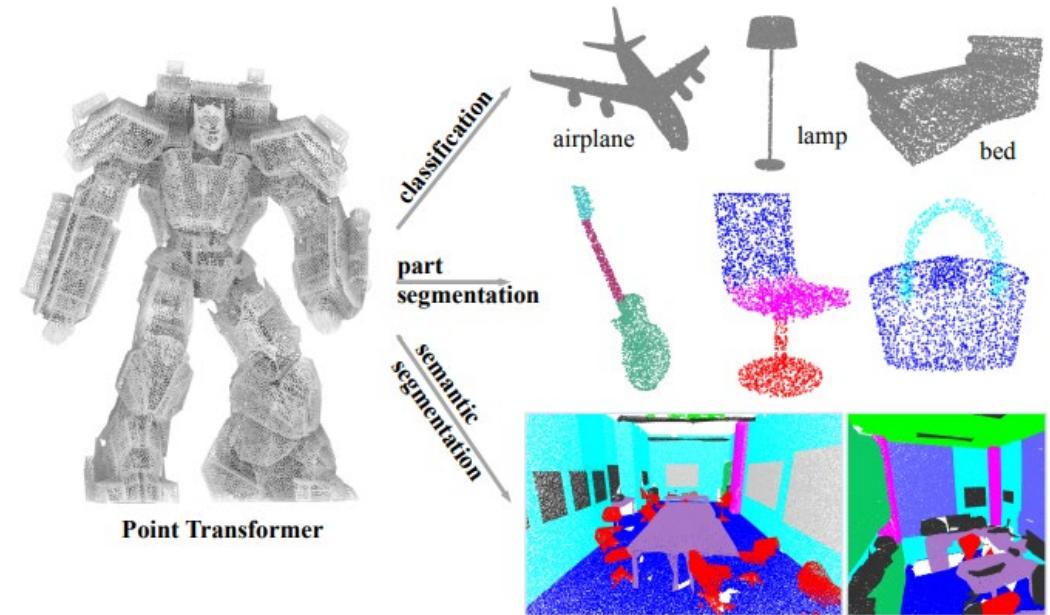


MinkowskiNet [Choy et al. 2019]

Prior work: Attention is All You Need



Transformer [Vaswani et al. 2017]



PointTransformer v1/v2 [Zhao et al. 2021, 2022]

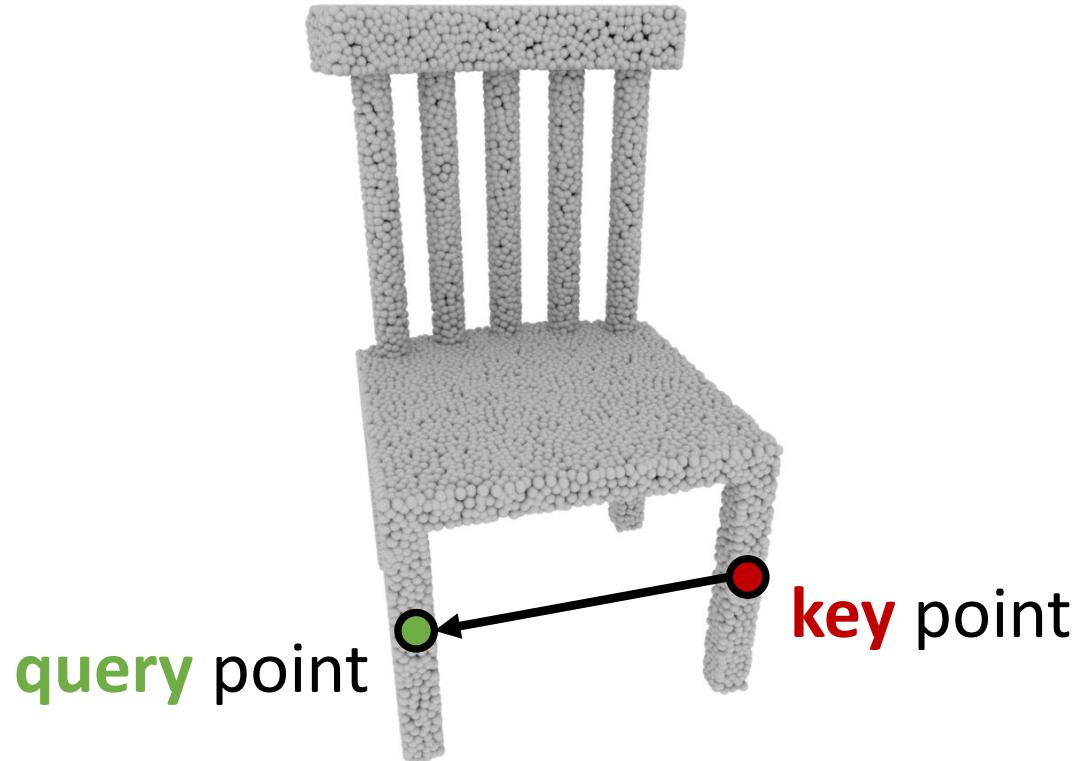
Why use attention for 3D representations?

Encode points such that their features capture **relations** wrt the rest of the shape



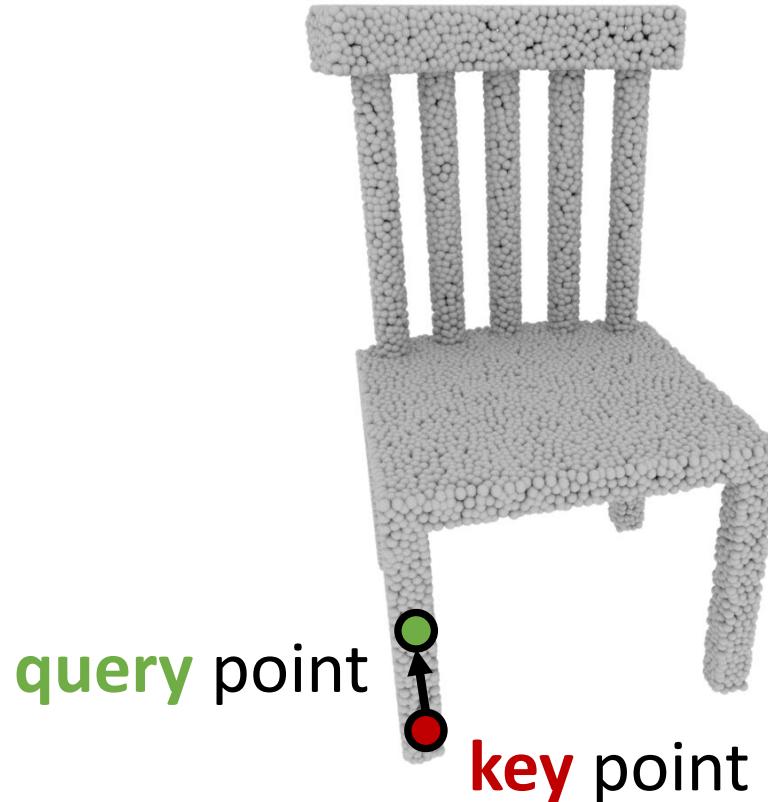
Why use attention for 3D representations?

Encode points such that their features capture **relations** wrt the rest of the shape



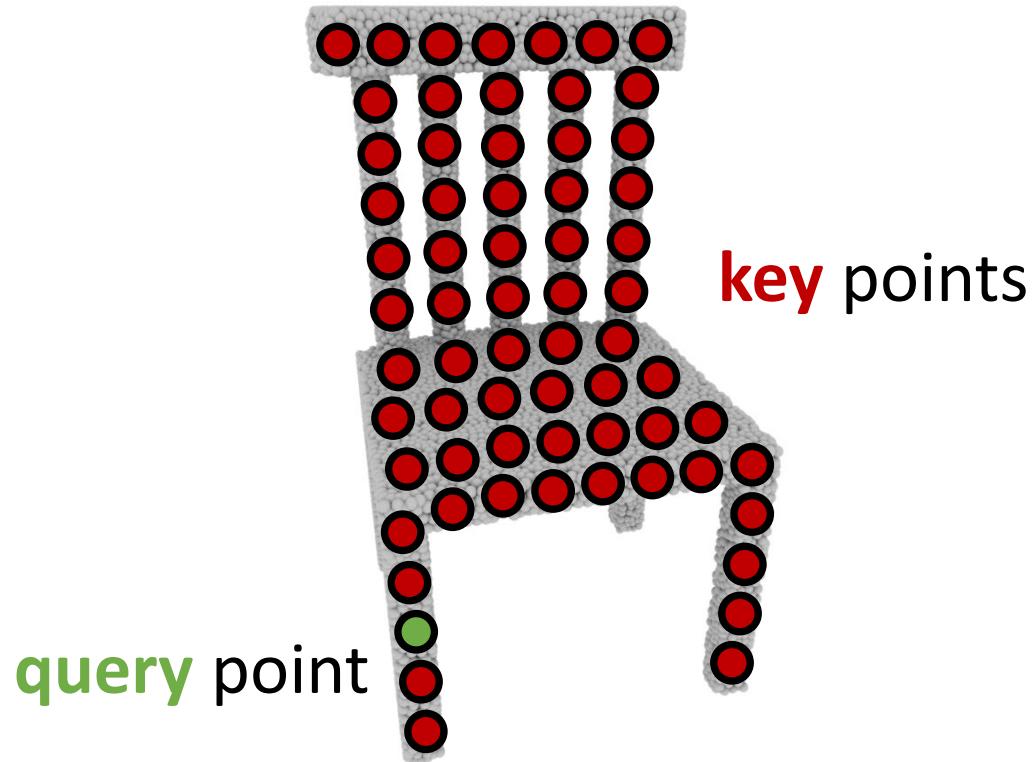
Why use attention for 3D representations?

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

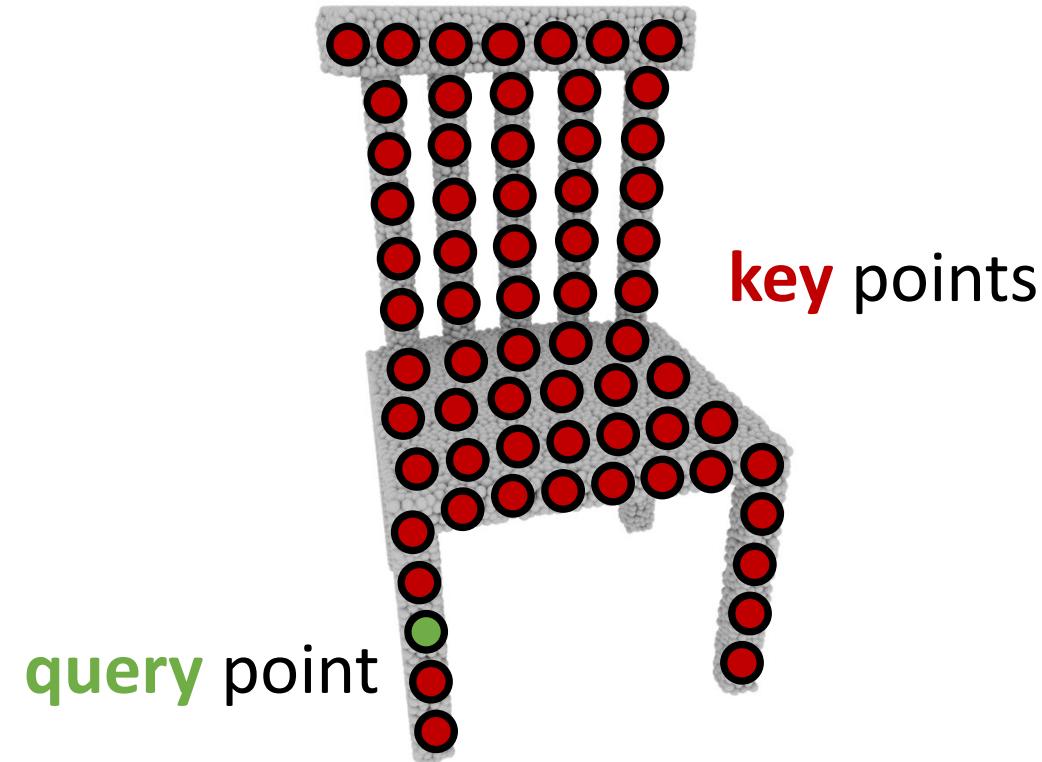


Why use attention for 3D representations?

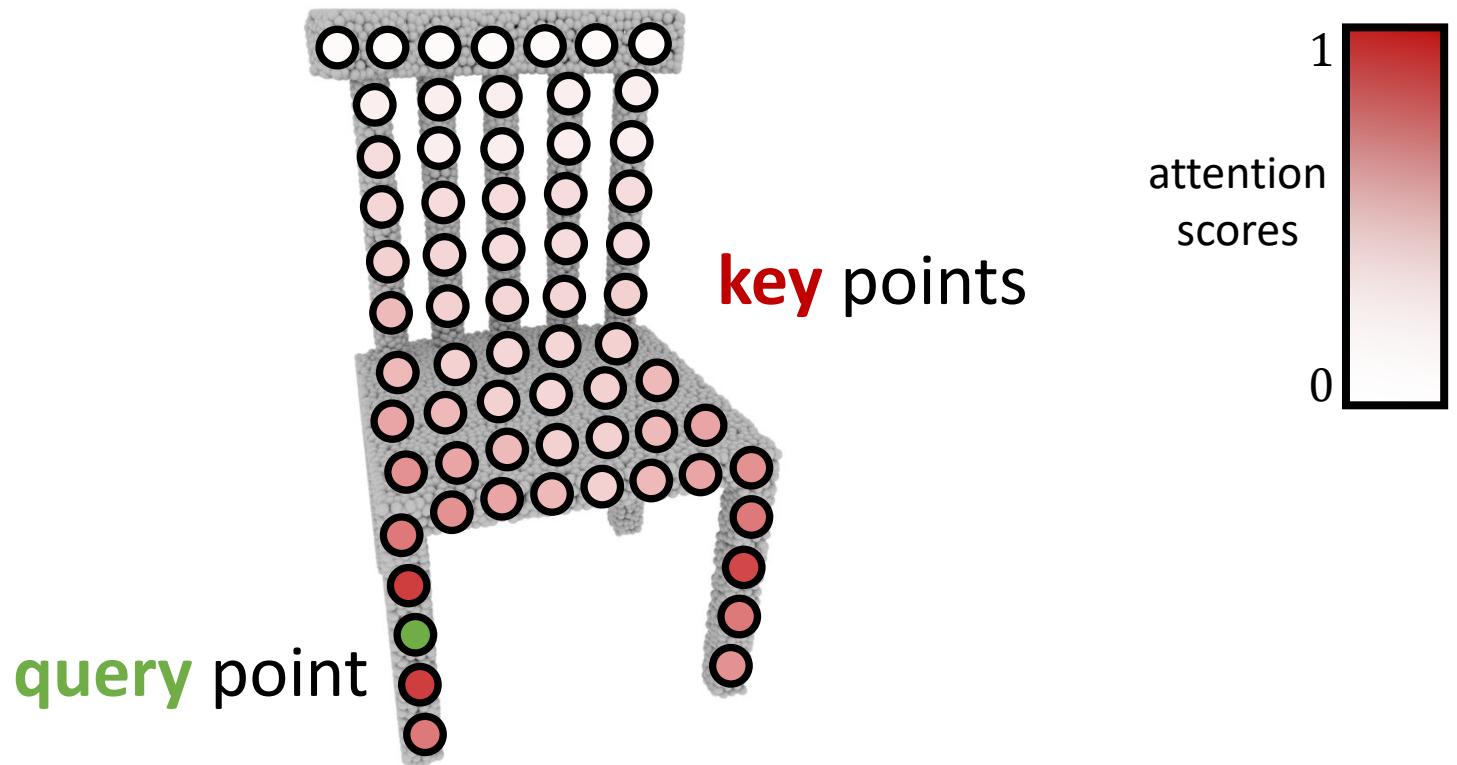
Encode points such that their features capture **relations** wrt the rest of the shape

$$\text{attention scores} = \text{query representation} \cdot \text{key representations}$$

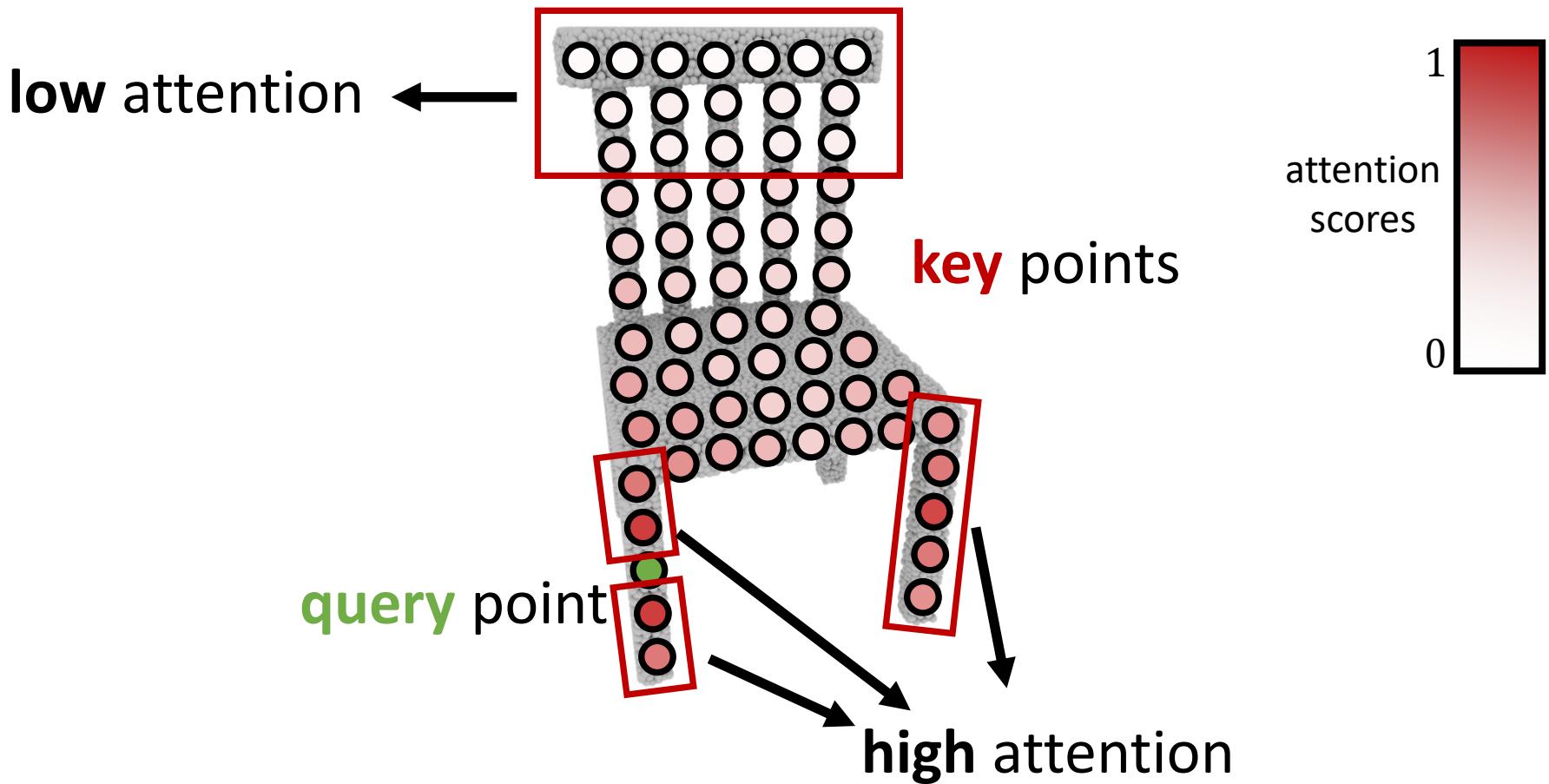
The diagram illustrates the computation of attention scores. On the left, a red box labeled "attention scores" contains five small squares. In the middle, a green box labeled "query representation" contains five squares of different colors (red, yellow, purple, orange, blue). To its right is a multiplication sign. On the far right, a red box labeled "key representations" contains a 5x5 grid of colored squares. A dot between the query and key boxes indicates matrix multiplication.



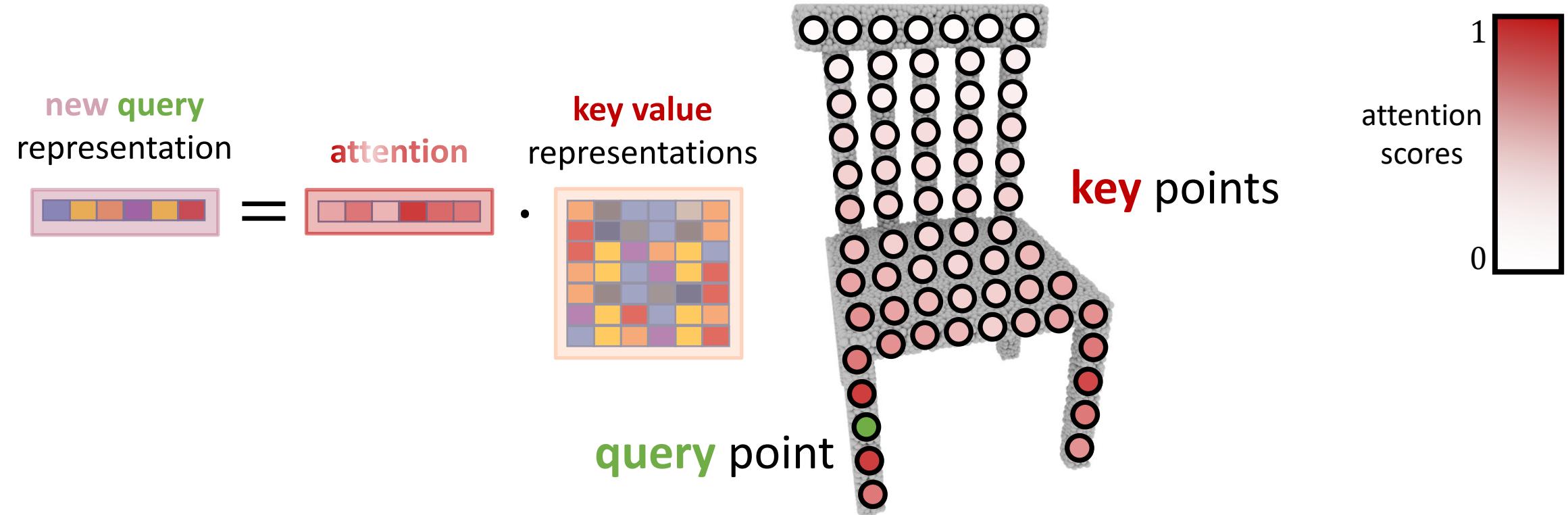
Why use attention for 3D representations?



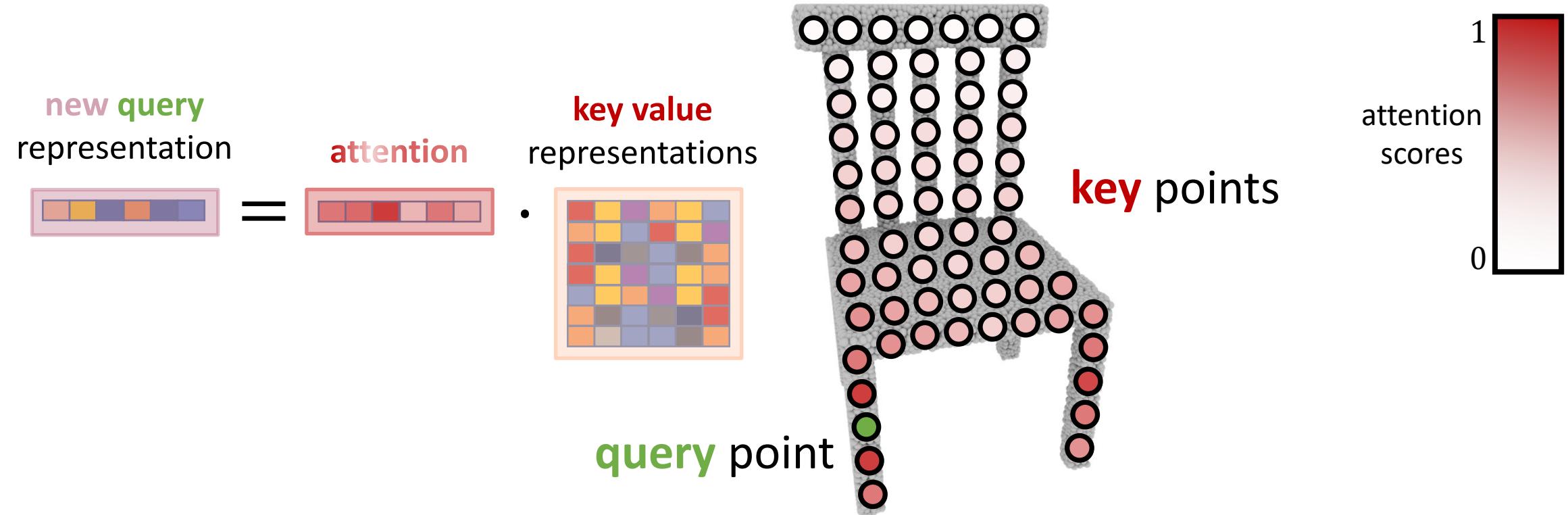
Why use attention for 3D representations?



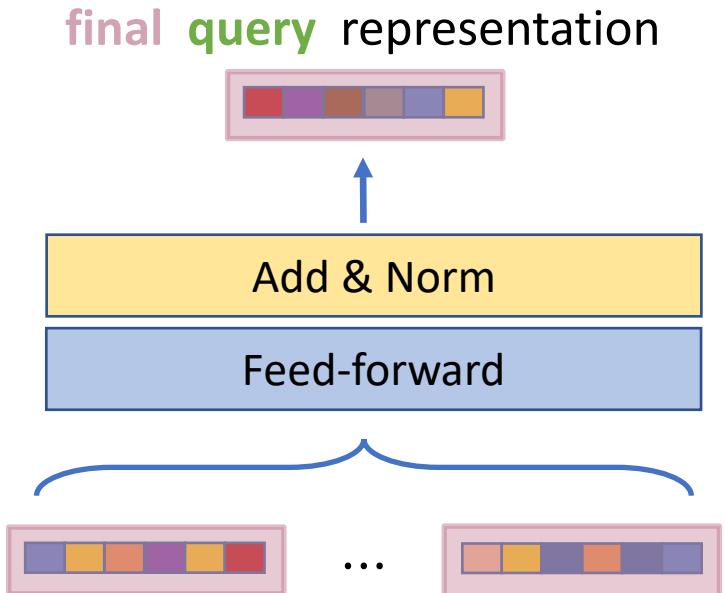
Why use attention for 3D representations?



Why use attention for 3D representations?

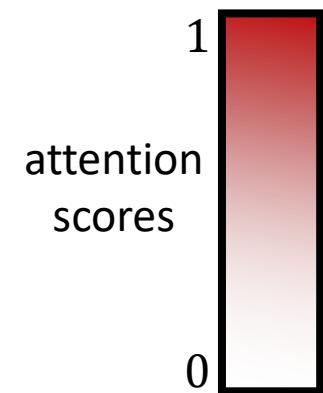
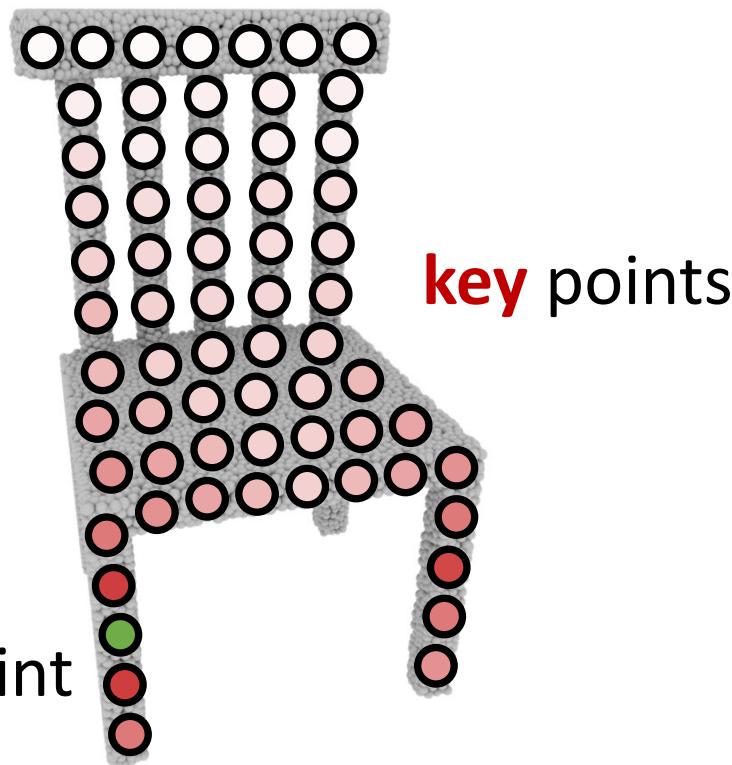


Why use attention for 3D representations?



query representations from
multiple attention heads

query point



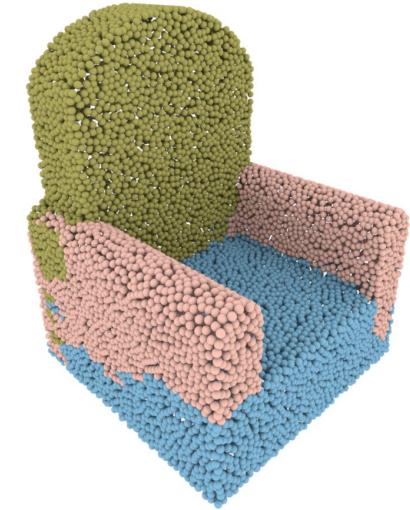
Motivation: Long-range interactions **across** shapes



test shape



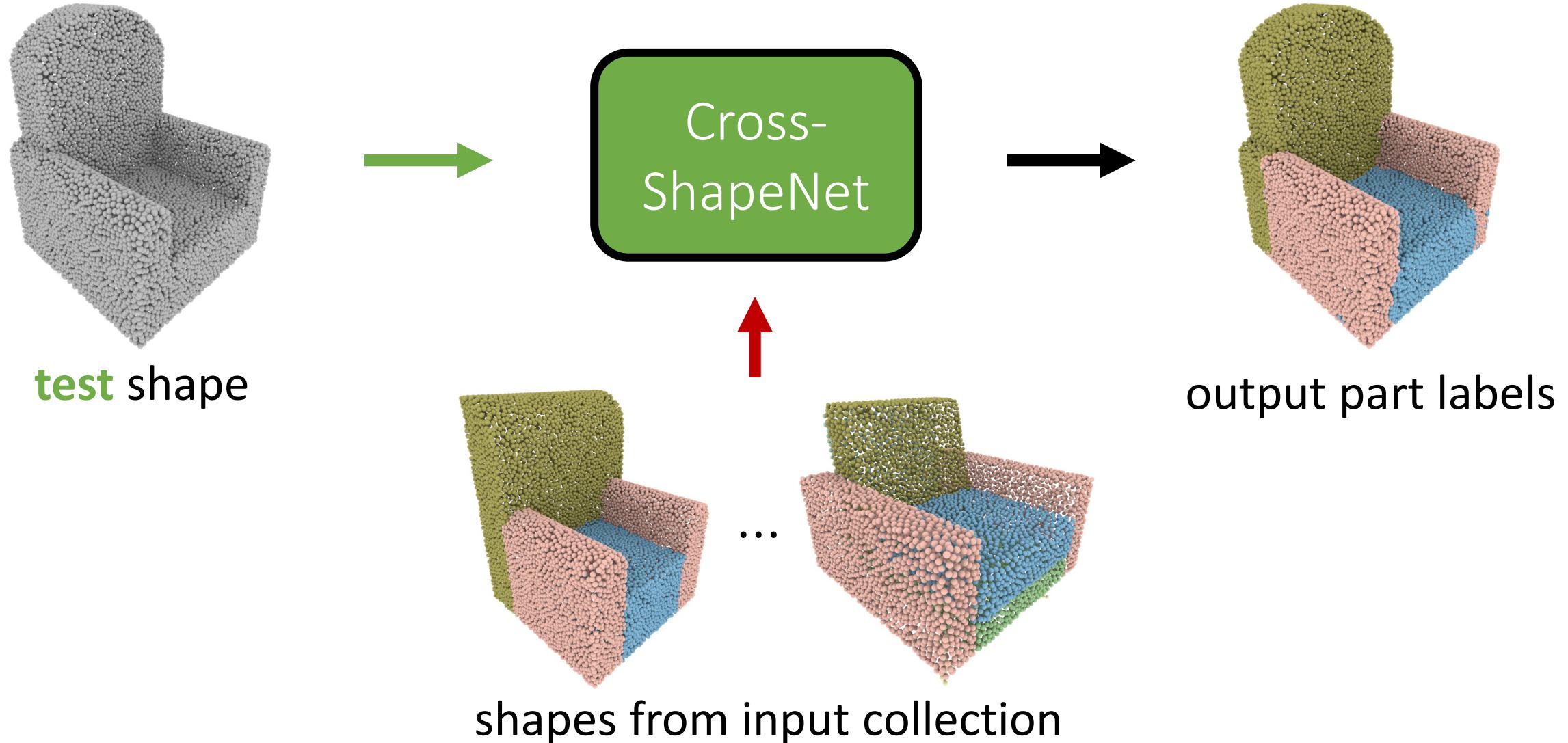
Prior Work



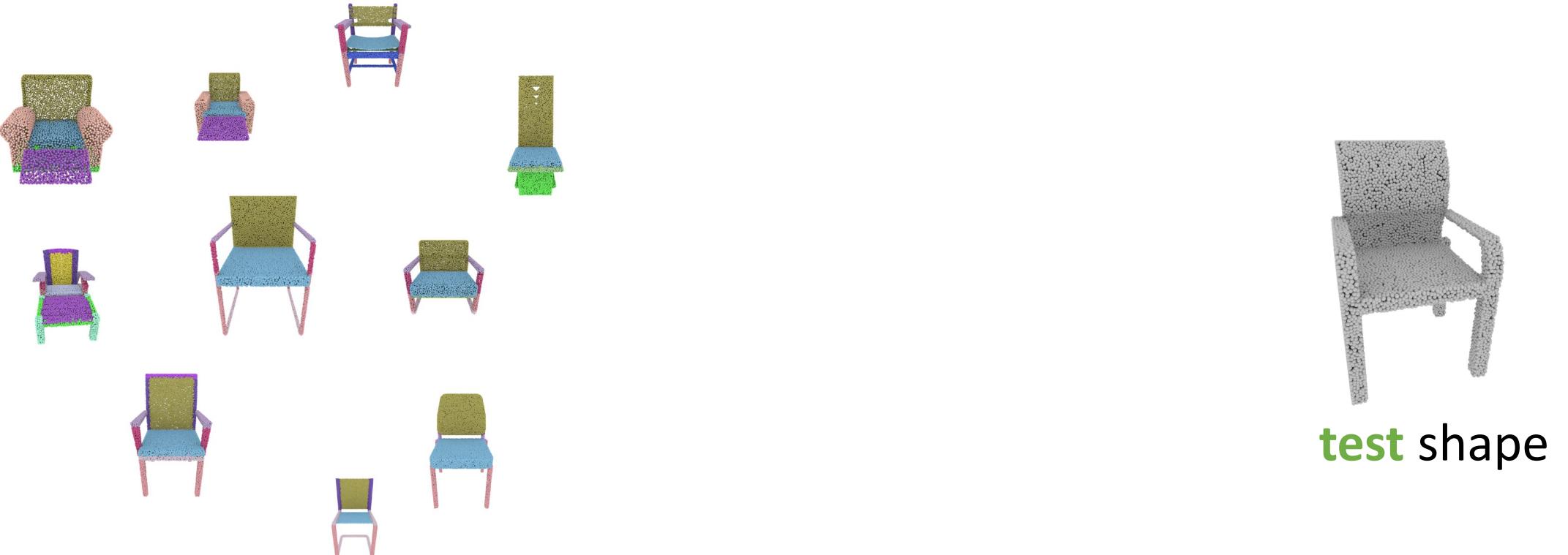
output part labels

No interactions across shapes

Motivation: Long-range interactions **across** shapes



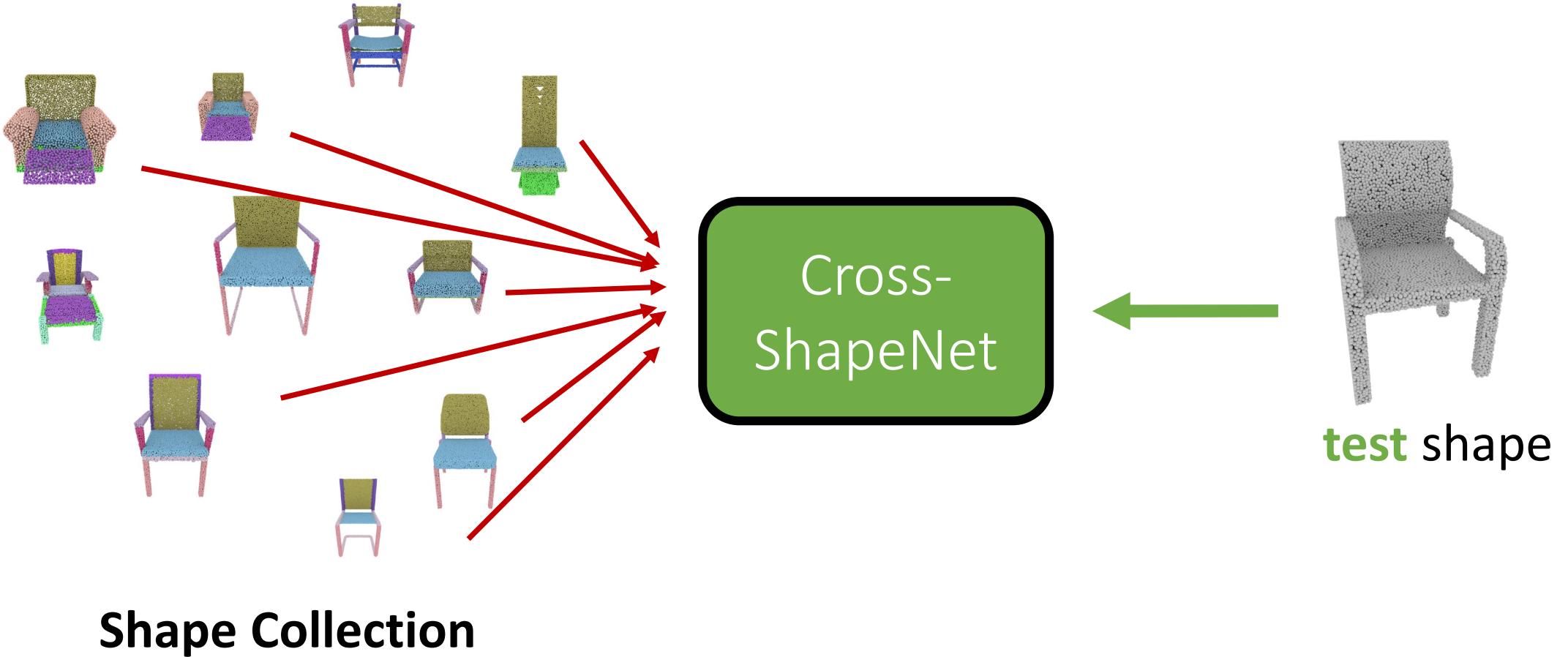
Key challenge: Retrieve compatible shapes



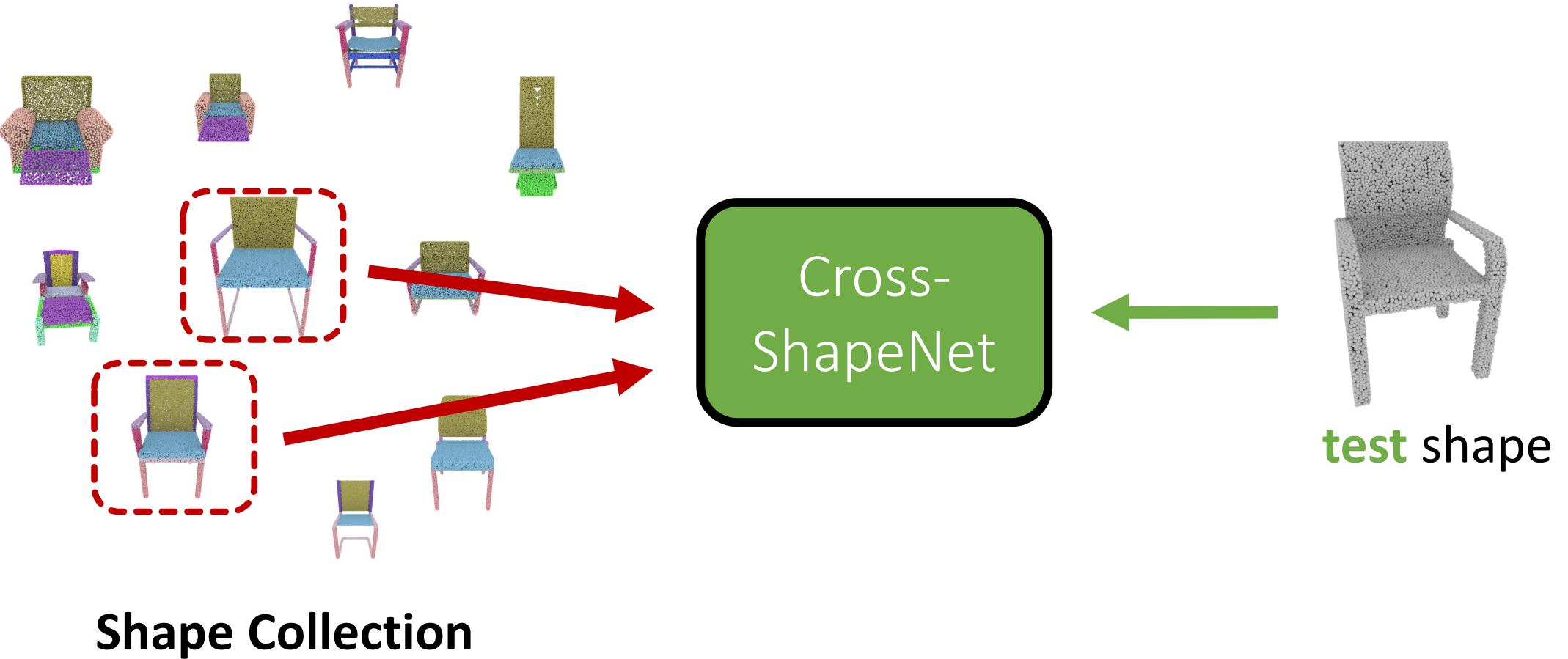
Shape Collection

test shape

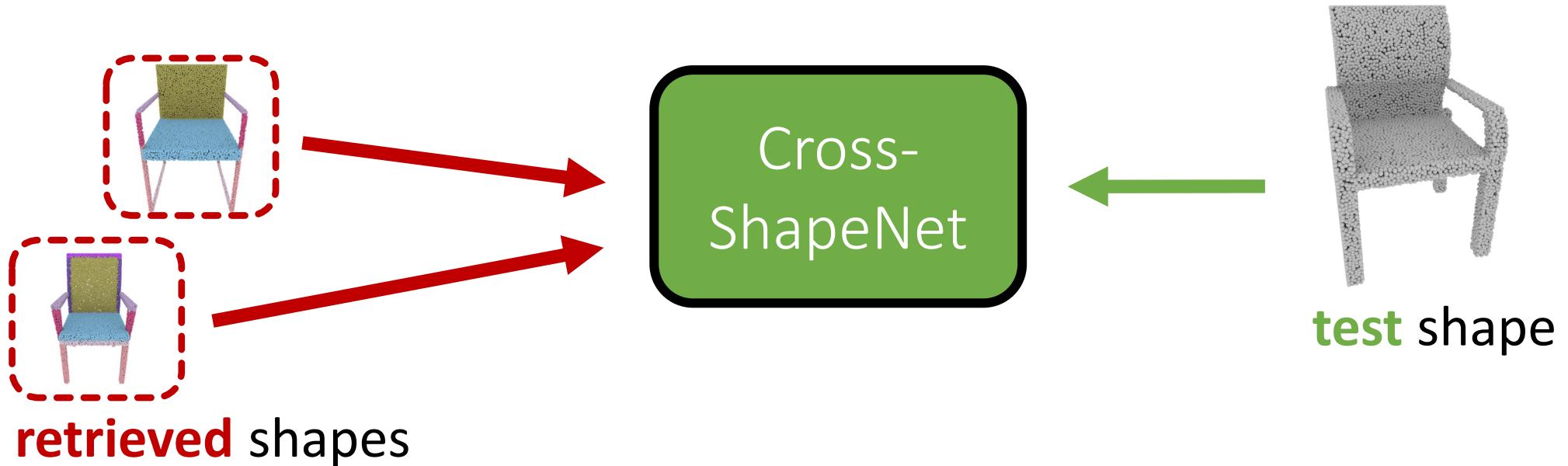
Key challenge: Retrieve compatible shapes



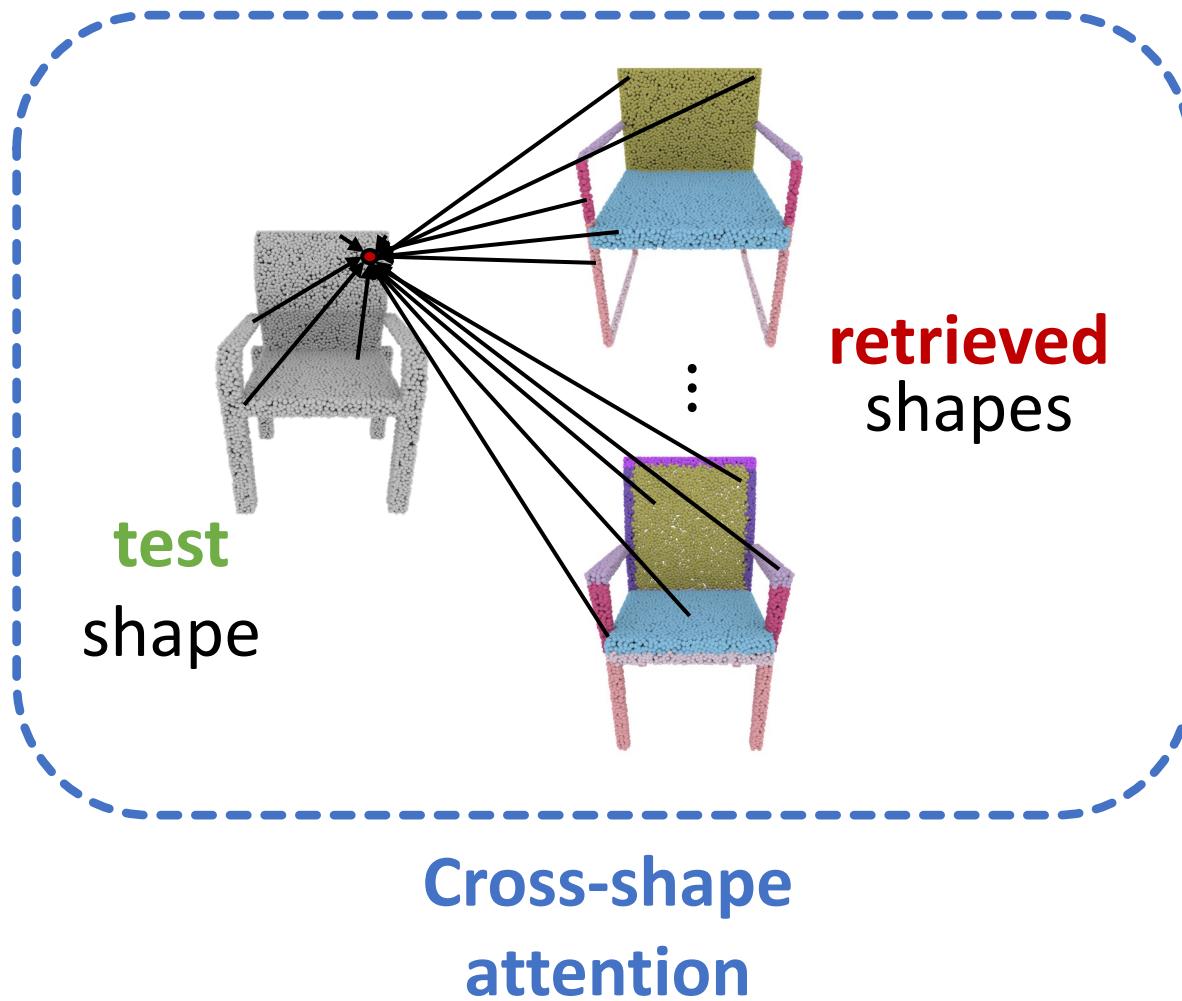
Key challenge: Retrieve compatible shapes



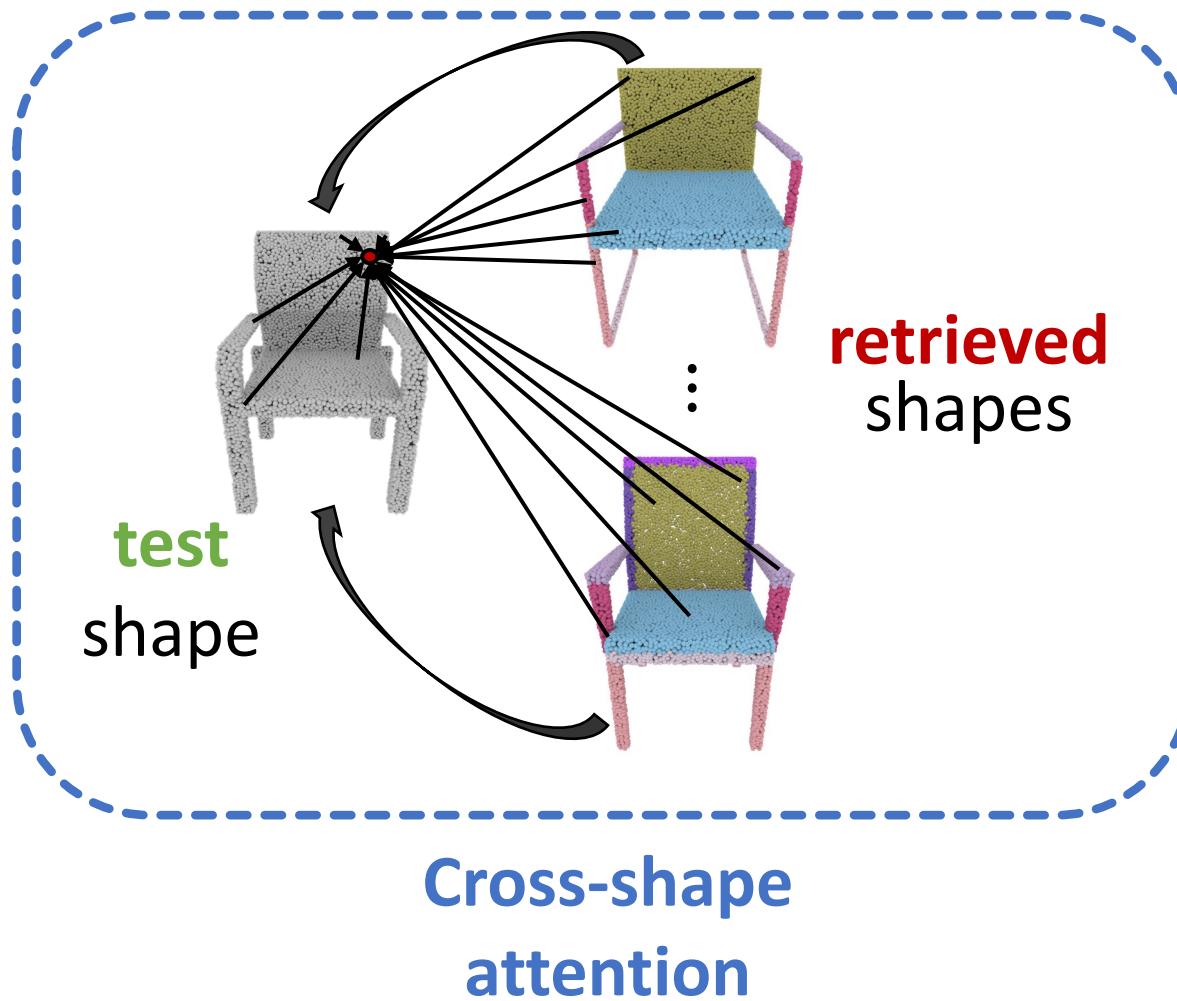
Key challenge: Combine multiple shapes



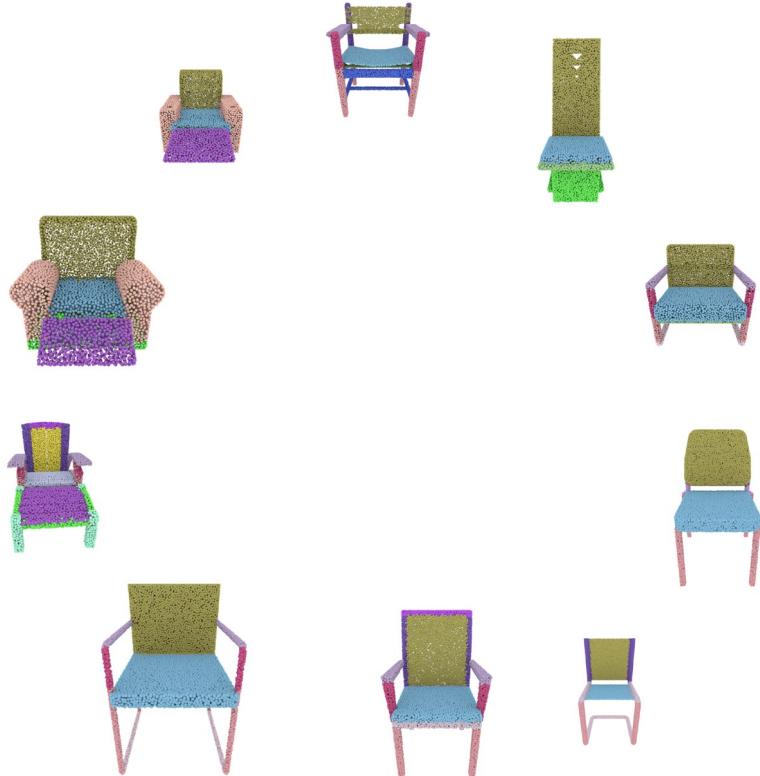
Key challenge: Combine multiple shapes



Key challenge: Combine multiple shapes

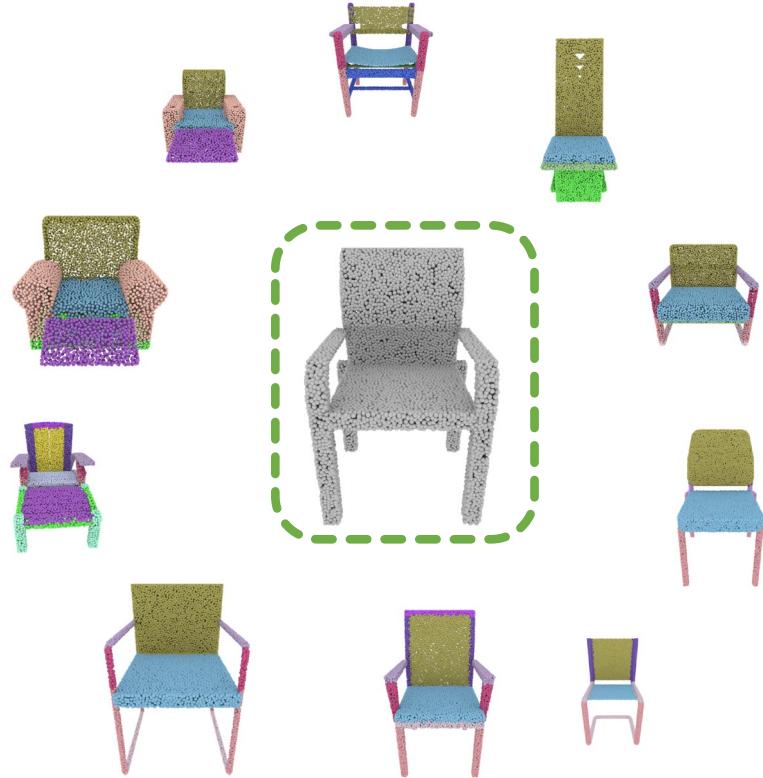


Pipeline



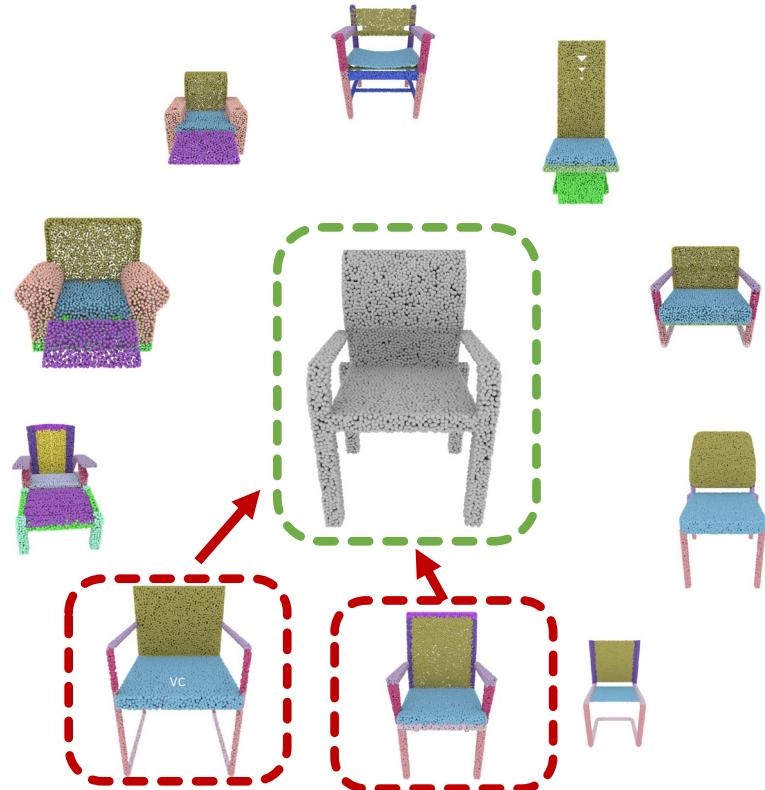
Shape Collection

Pipeline



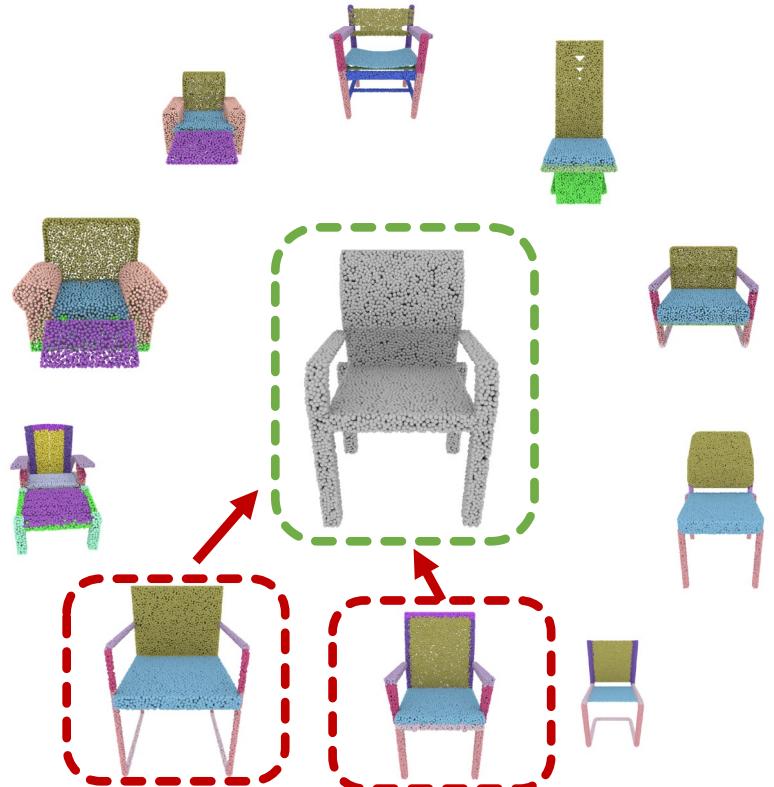
Shape Collection

Pipeline

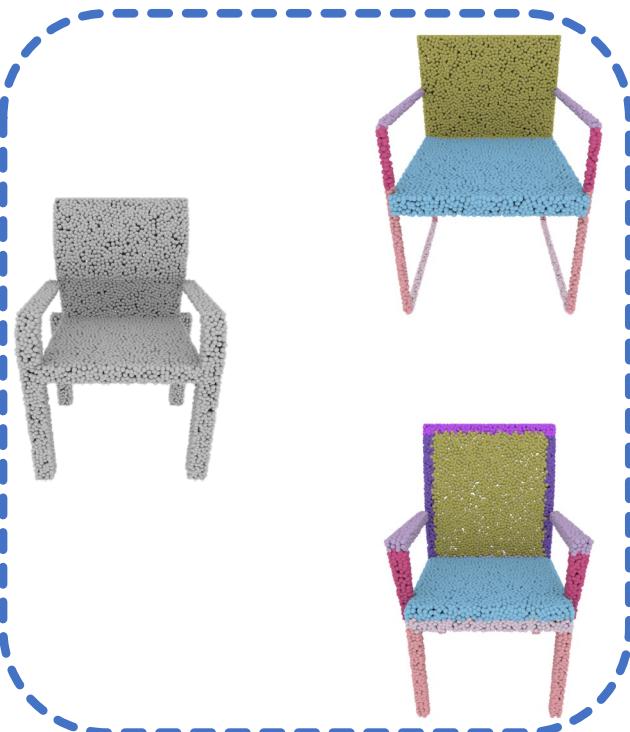


Shape Collection

Pipeline

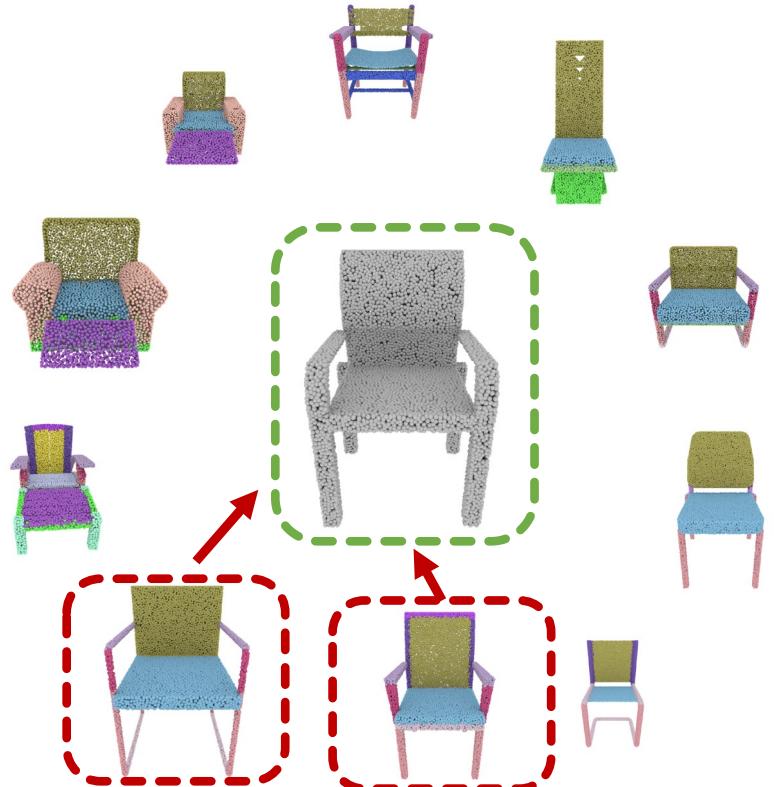


Shape Collection

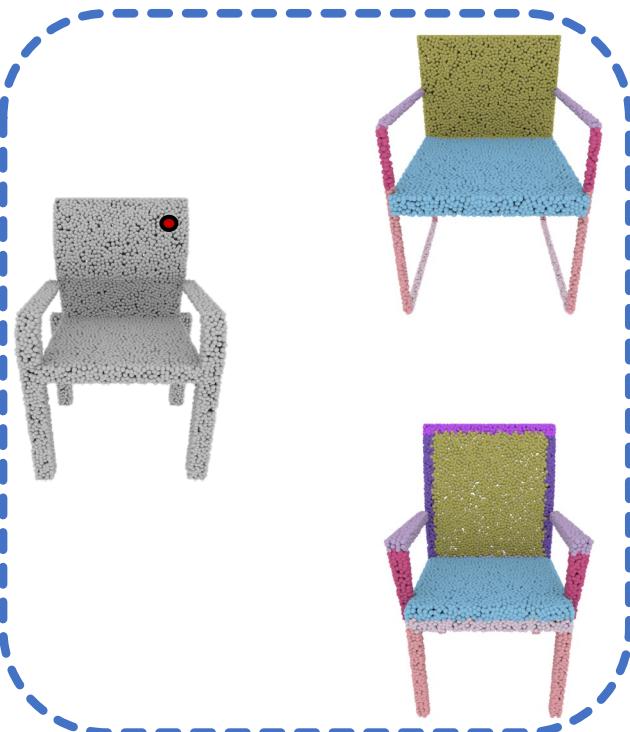


**Cross-shape
attention**

Pipeline

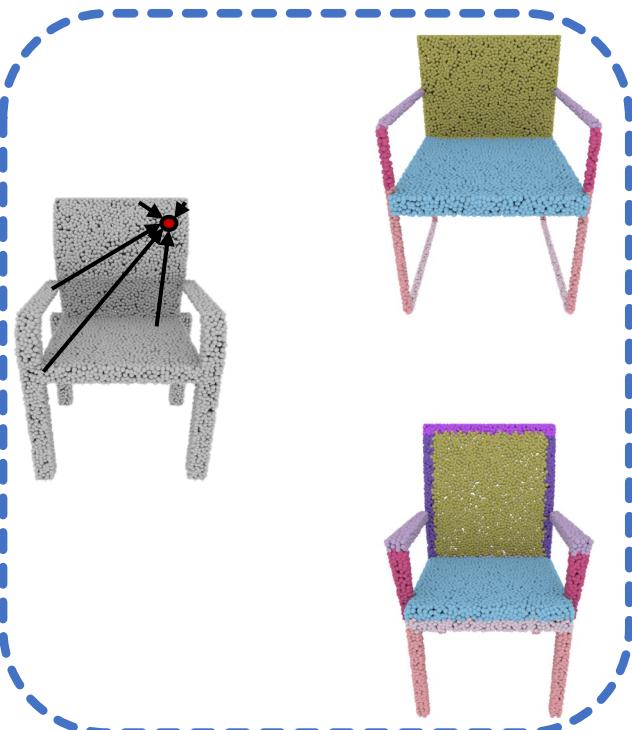
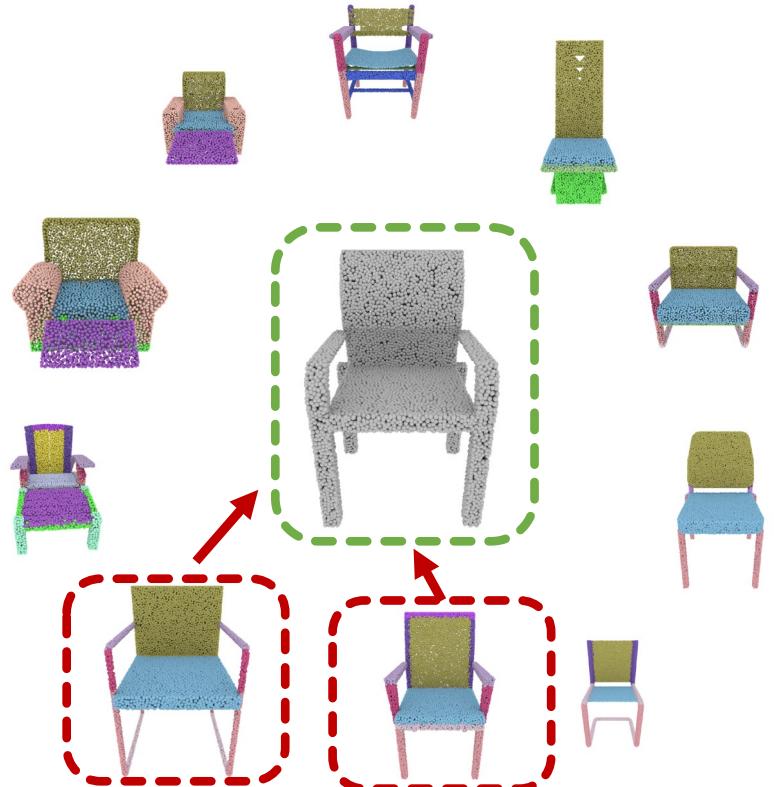


Shape Collection

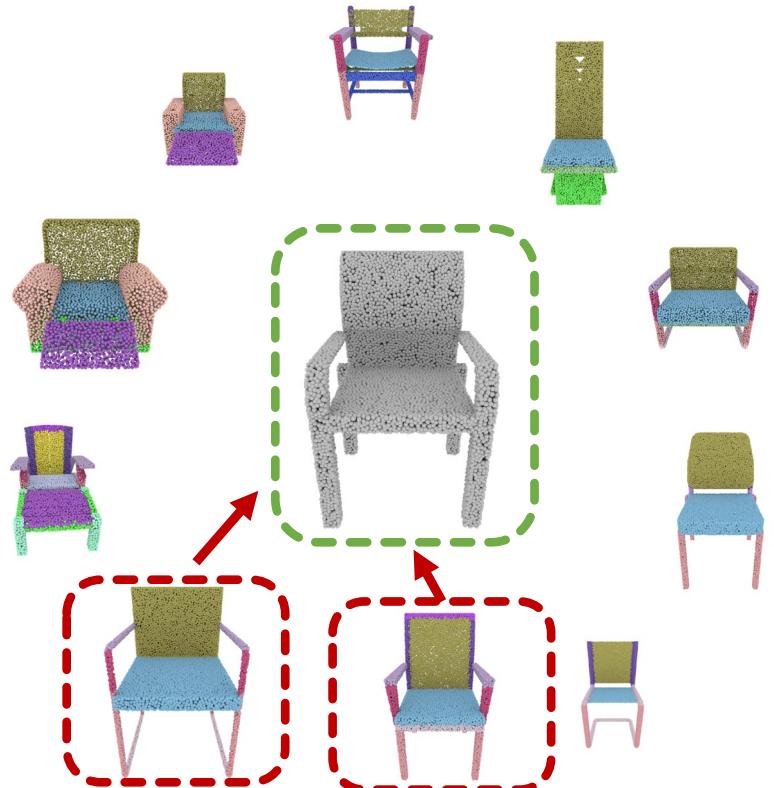


**Cross-shape
attention**

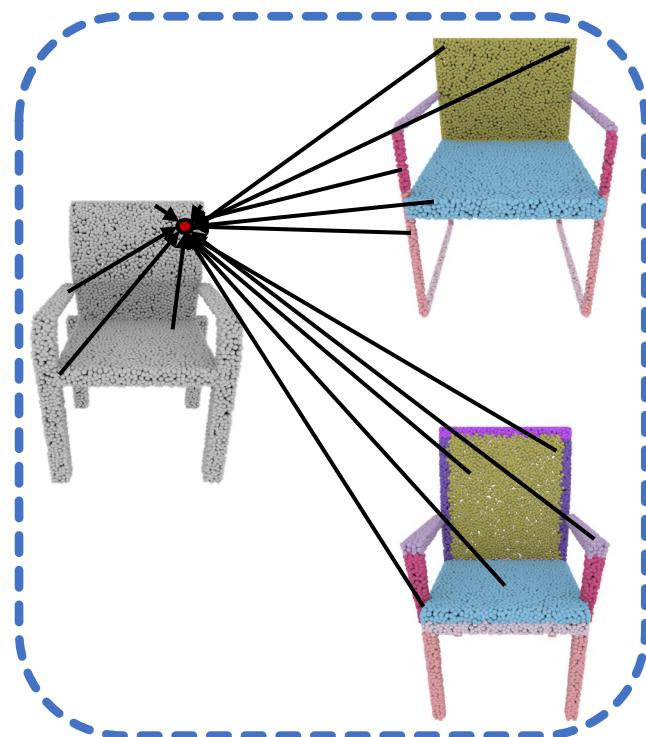
Pipeline



Pipeline

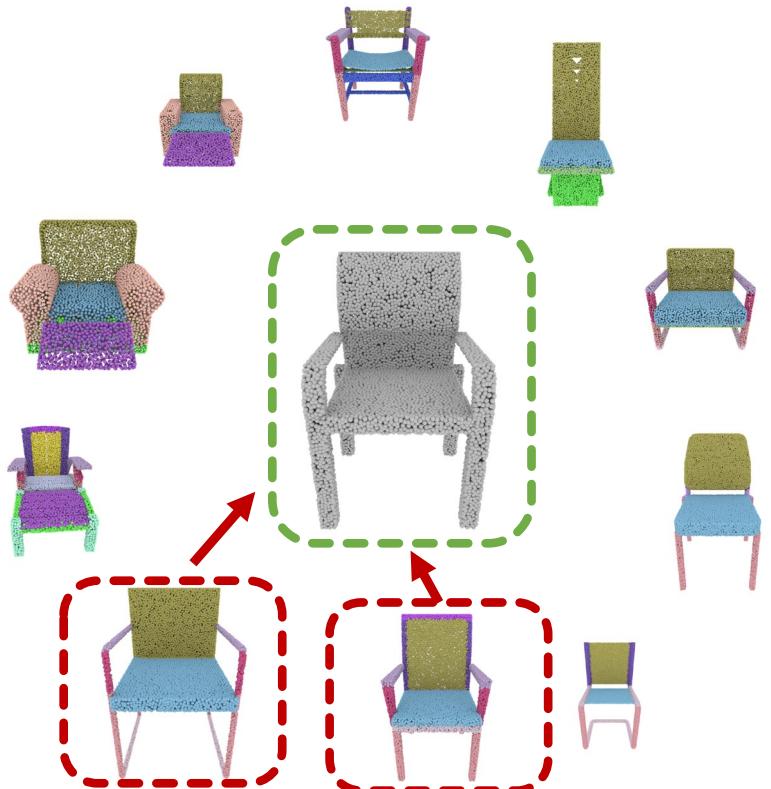


Shape Collection

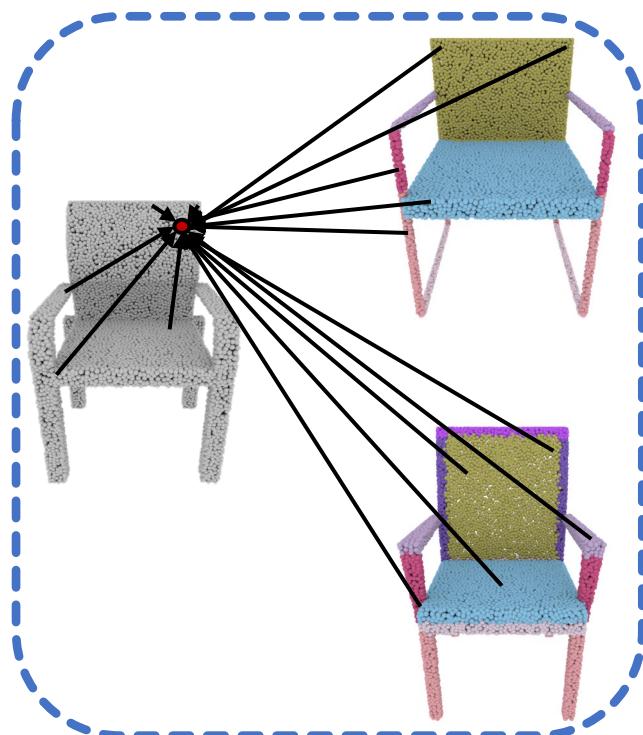


**Cross-shape
attention**

Pipeline



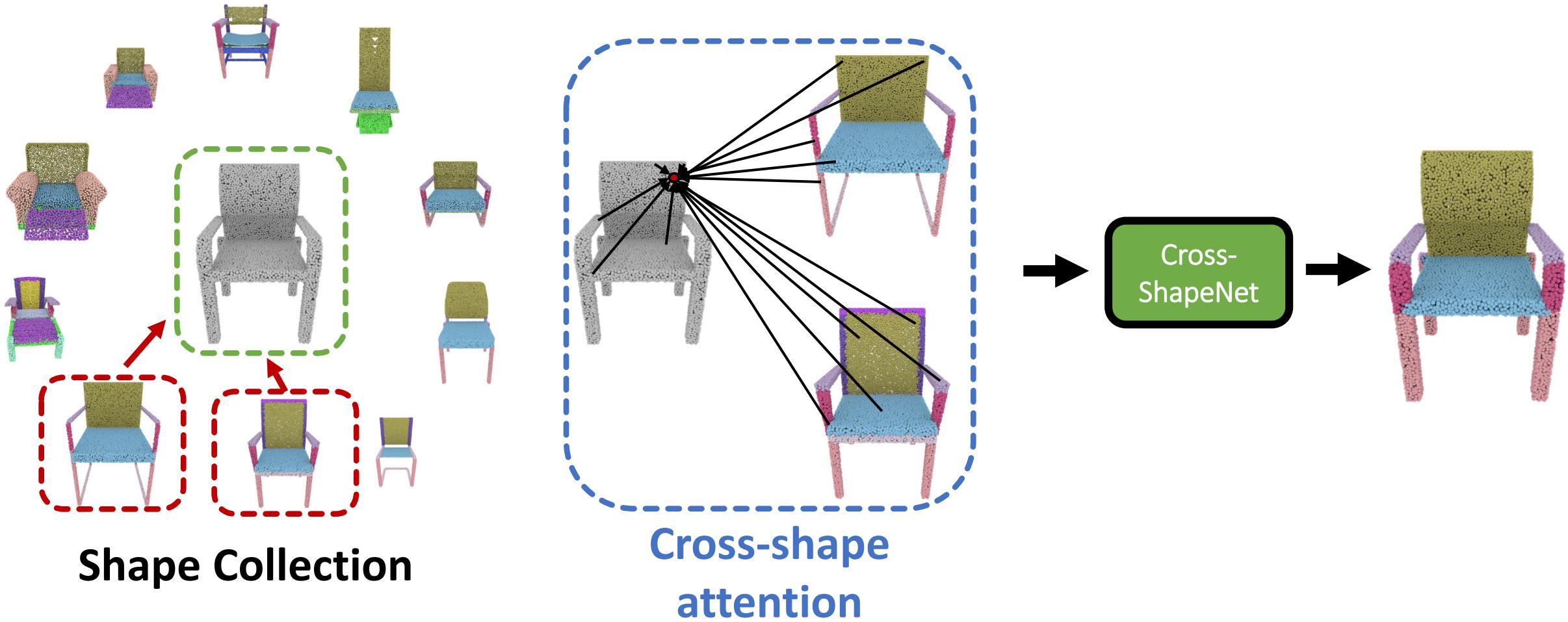
Shape Collection



**Cross-shape
attention**



Pipeline

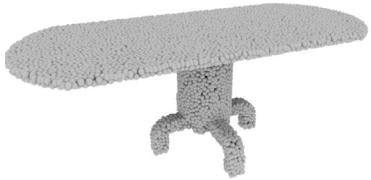


Cross-Shape Attention

query shape $\mathcal{S}_m = \{\mathbf{p}_i\}_{i=1}^M$



key shape $\mathcal{S}_n = \{\mathbf{p}_j\}_{j=1}^N$

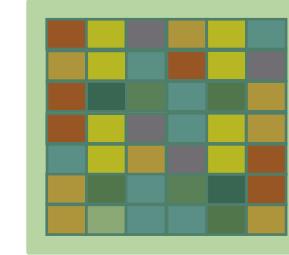


Cross-Shape Attention

query shape $\mathcal{S}_m = \{\mathbf{p}_i\}_{i=1}^M$



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

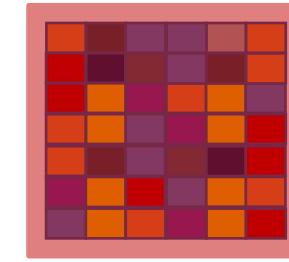


Backbone point
representations

key shape $\mathcal{S}_n = \{\mathbf{p}_j\}_{j=1}^N$

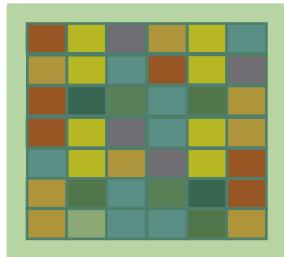


$$\mathbf{X}_n \in R^{N \times D}$$

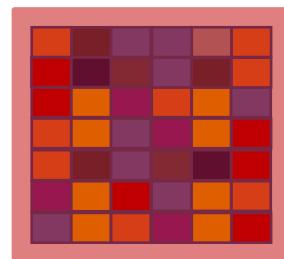


Cross-Shape Attention

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

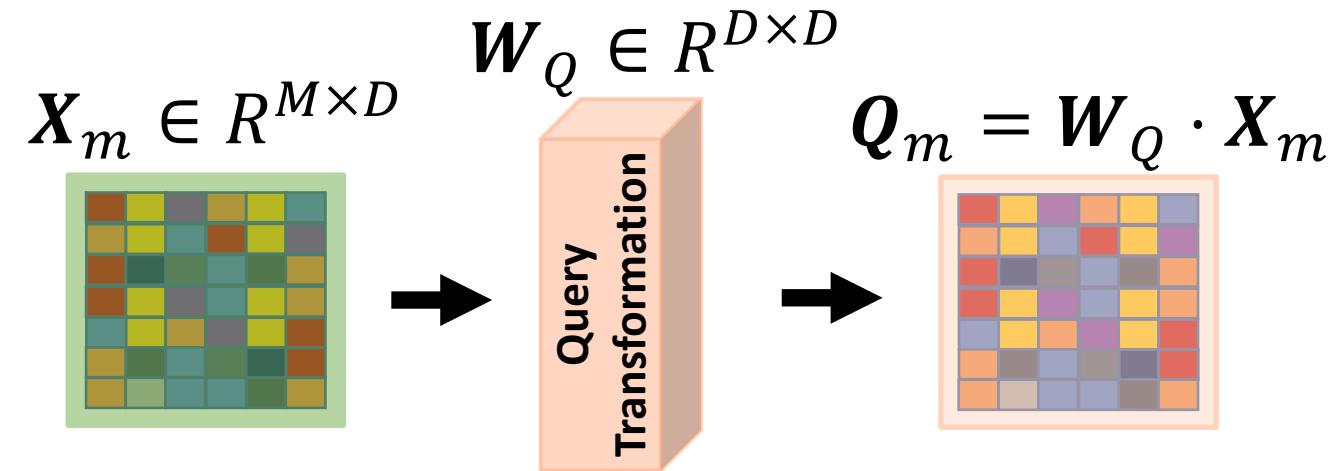


Backbone point
representations



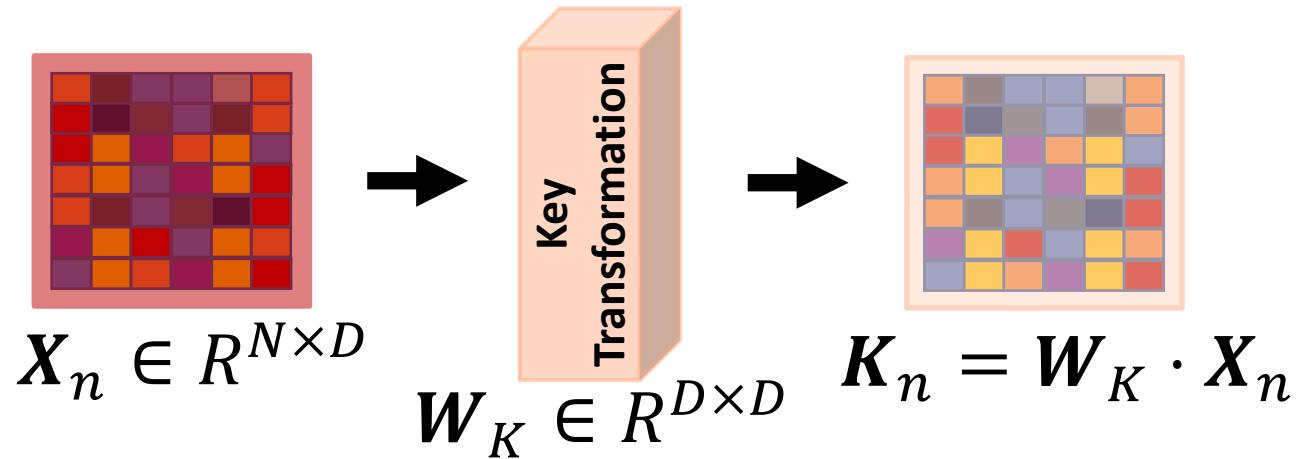
$$X_n \in R^{N \times D}$$

Cross-Shape Attention

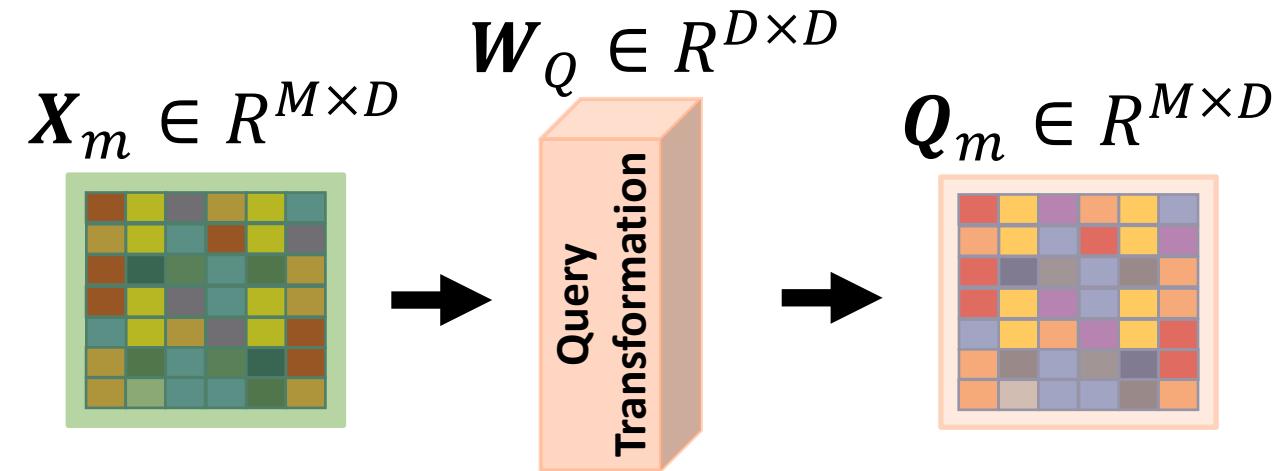


Backbone point
representations

Intermediate
representations

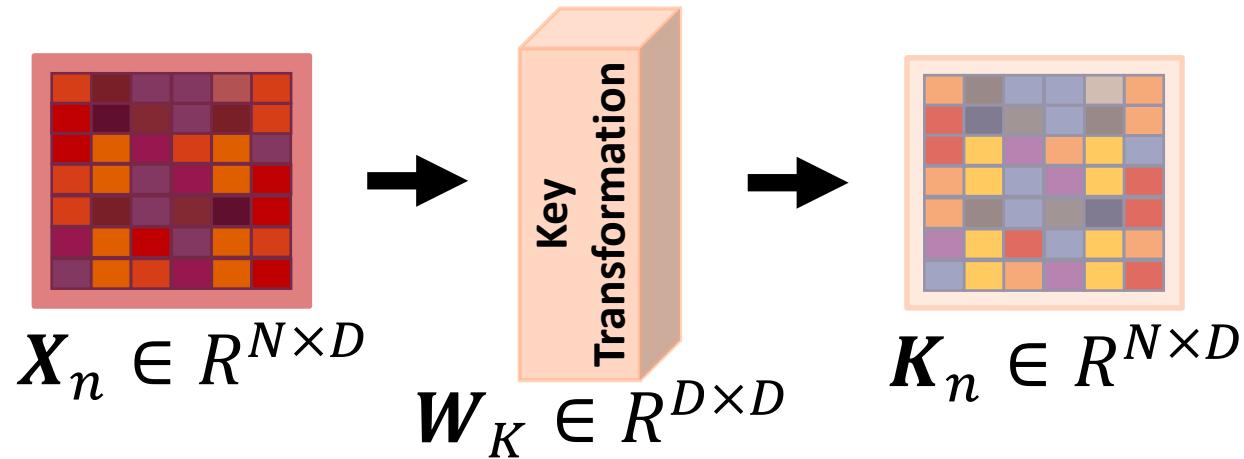


Cross-Shape Attention

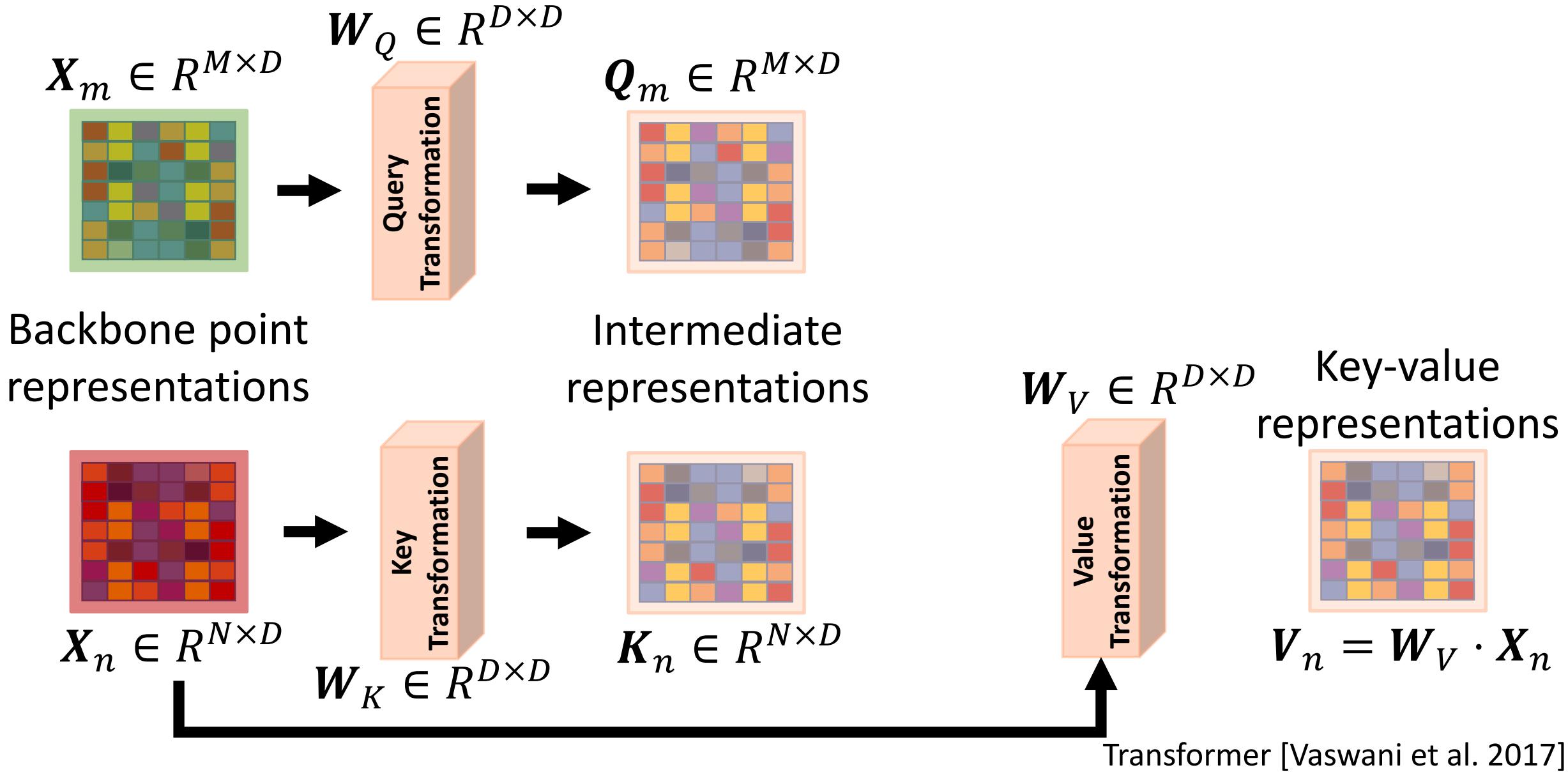


Backbone point
representations

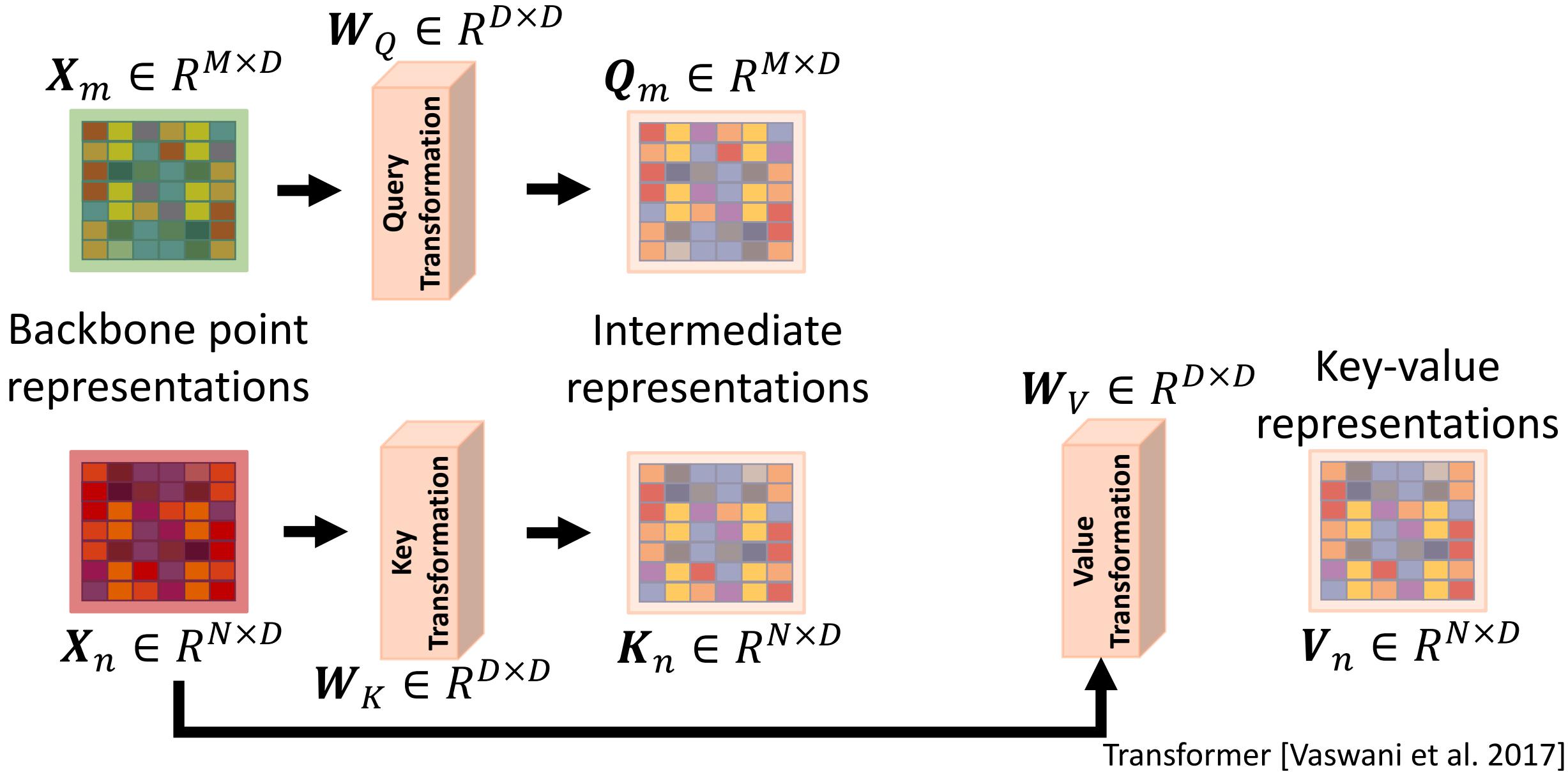
Intermediate
representations



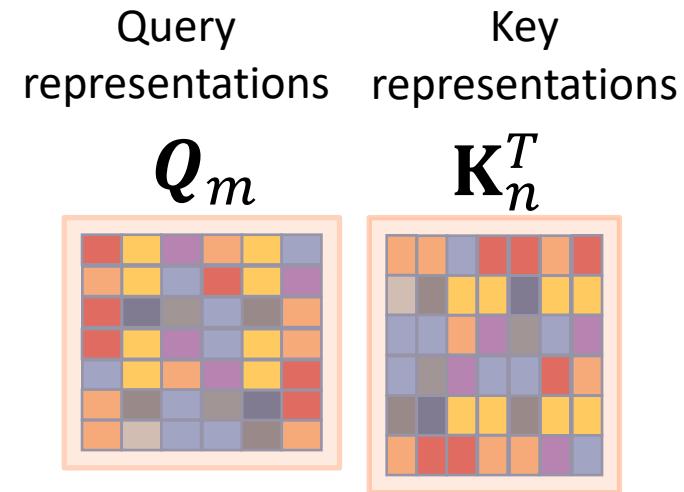
Cross-Shape Attention



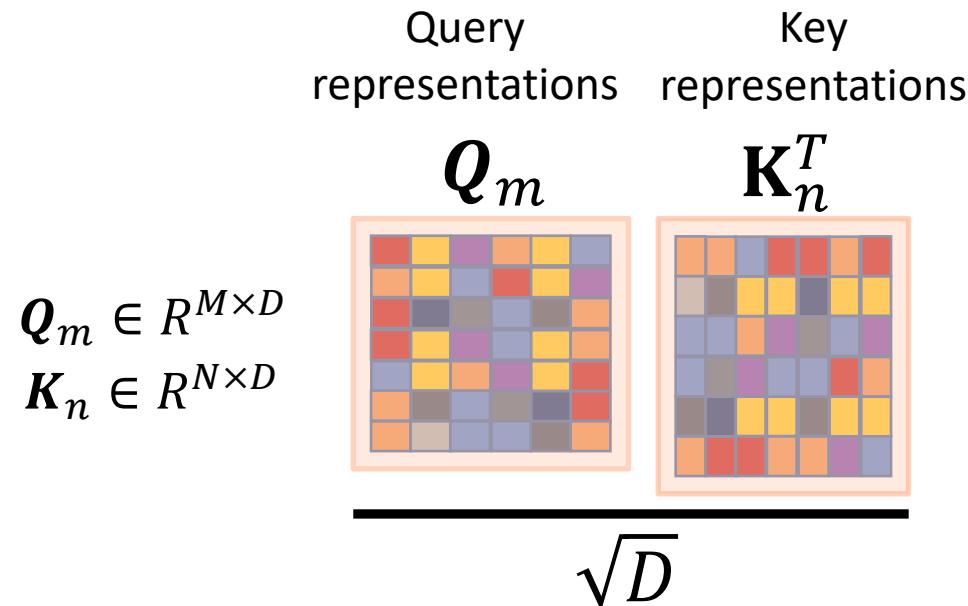
Cross-Shape Attention



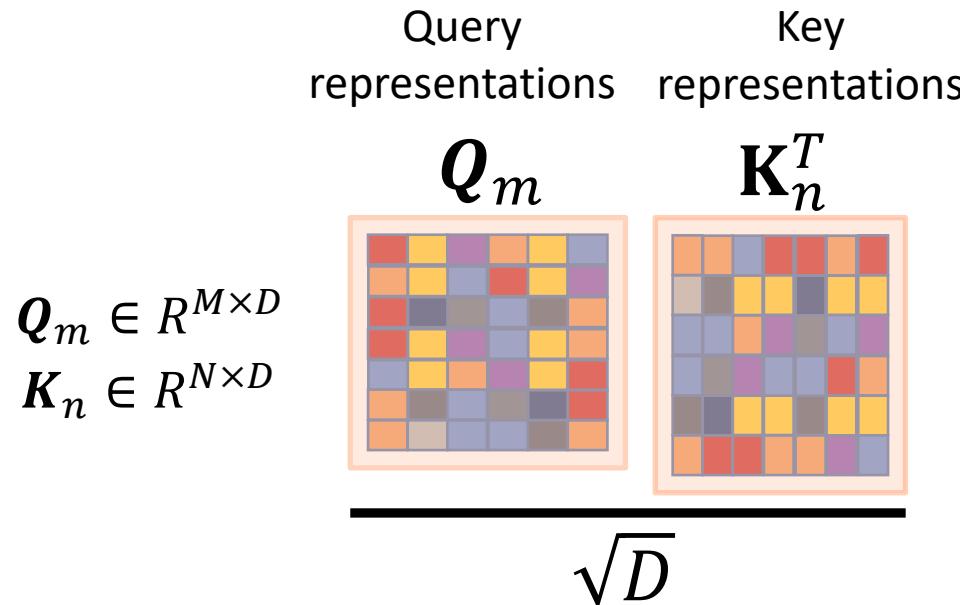
Cross-Shape Attention



Cross-Shape Attention



Cross-Shape Attention



$$Var\left(\frac{\mathbf{Q}_{i,:}\mathbf{K}_{:,j}^T}{\sqrt{D}}\right) = 1,$$
$$\forall i = 1, \dots, M$$
$$\forall j = 1, \dots, N$$

Cross-Shape Attention

Query representations Key representations

$$softmax \left(\frac{\begin{matrix} Q_m \\ K_n^T \end{matrix}}{\sqrt{D}} \right) = A_{m,n} \in R^{M \times N}$$

Attention matrix

The diagram illustrates the computation of a cross-shape attention matrix. It starts with two input matrices: Q_m (Query representations) and K_n^T (Key representations). These are combined into a single column vector. This vector is then scaled by \sqrt{D} . Finally, the softmax function is applied to this scaled vector to produce the attention matrix $A_{m,n}$, which is a $R^{M \times N}$ matrix.

Cross-Shape Attention

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

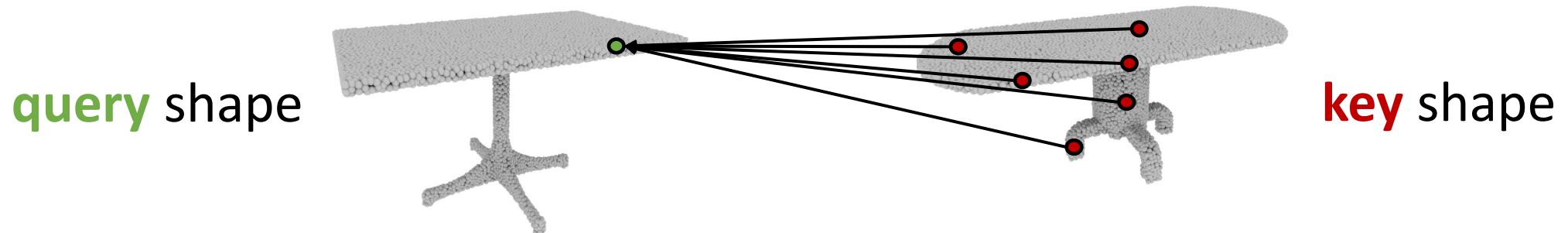
Diagram illustrating the computation of Cross-Shape Attention:

- Cross-shape attention matrix**: A square matrix $A_{m,n}$ with a blue border, containing a grid of gray and white squares.
- Key shape value representations**: A square matrix V_n with an orange border, containing a grid of colored squares (orange, red, yellow, purple).
- Product**: The multiplication of $A_{m,n}$ and V_n is indicated by a dot between them.
- Equal sign**: An equals sign ($=$) followed by a pink-bordered square matrix $X_m'^{(CSA)}$, which contains a grid of colored squares (orange, red, yellow, purple), representing the result of the cross-shape attention operation.
- Cross-shape attention representations**: The resulting matrix $X_m'^{(CSA)}$.

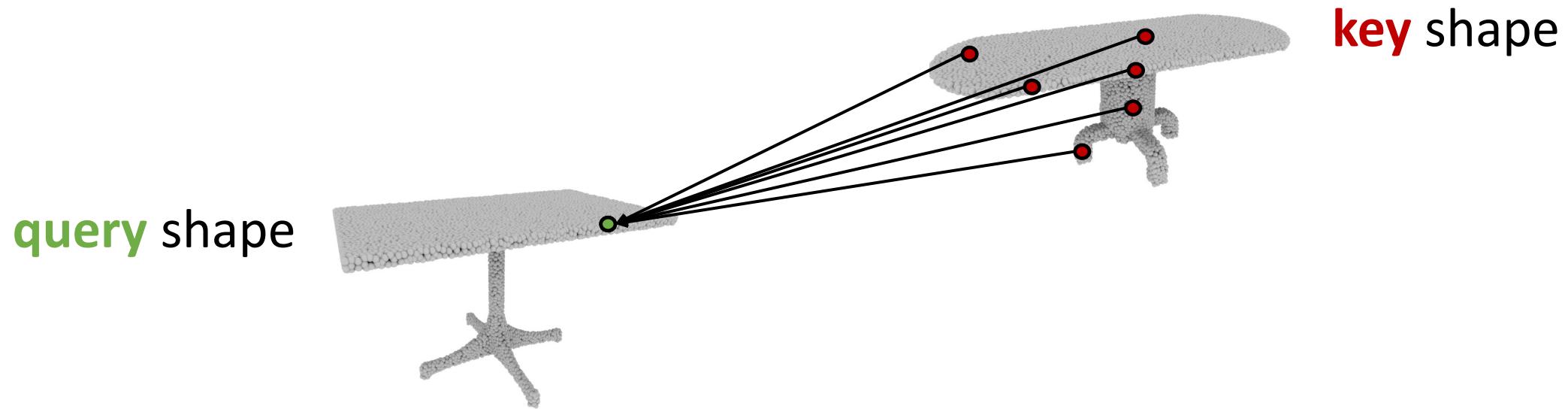
Cross-Shape Attention

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

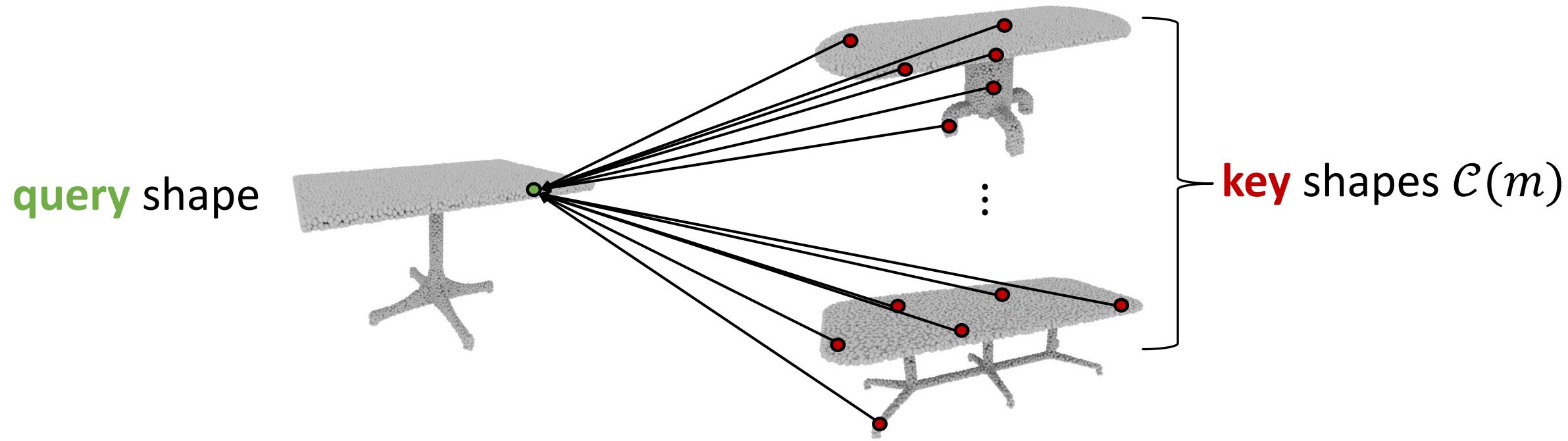
Cross-shape attention matrix **Key shape** value representations **Cross-shape** attention representations



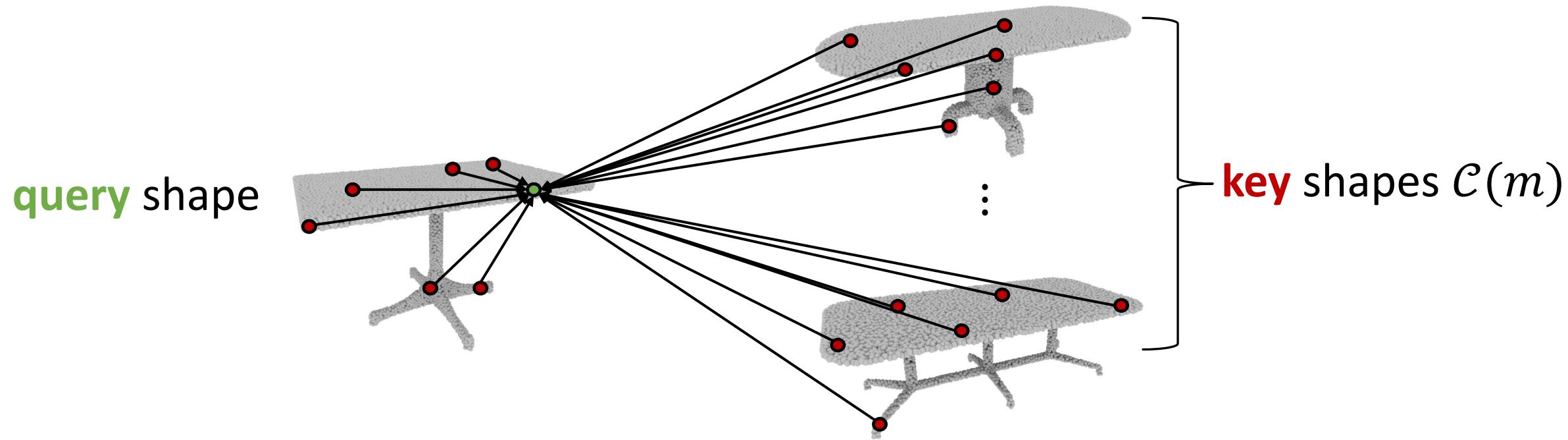
Cross-Shape Attention for multiple shapes



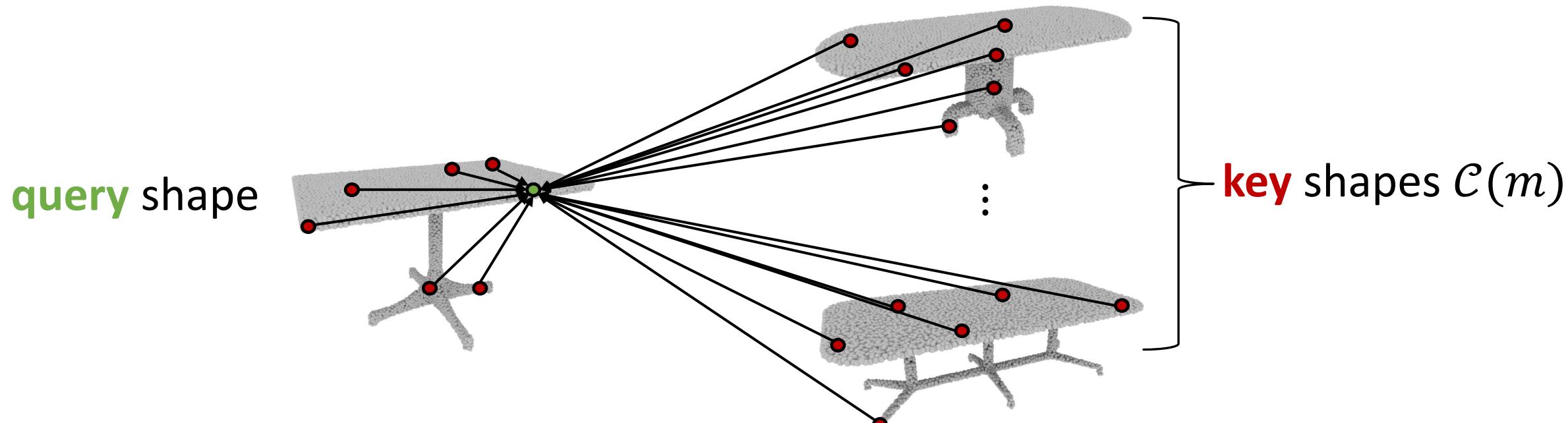
Cross-Shape Attention for multiple shapes



Cross-Shape Attention for multiple shapes



Cross-Shape Attention for multiple shapes



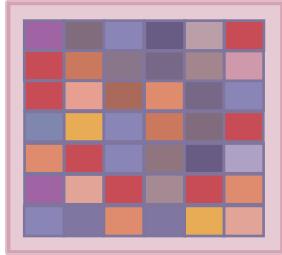
- $\mathcal{C}(m)$: set of compatible key shapes
- $c(m, n)$: compatibility function between query shape S_m and key shape S_n

Cross-shape
attention output

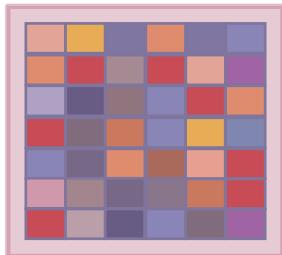
$$X'_m = \sum_{n \in \{\mathcal{C}(m), m\}} c(m, n) A_{m,n} V_n$$

Compatibility function

$$X'_m^{(SSA)} \in R^{M \times D}$$

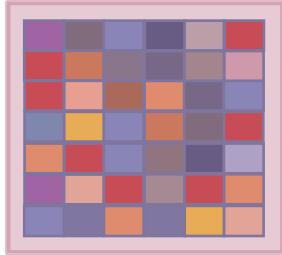


$$X'_n^{(SSA)} \in R^{N \times D}$$



Compatibility function

$$X'_m \in R^{M \times D}$$

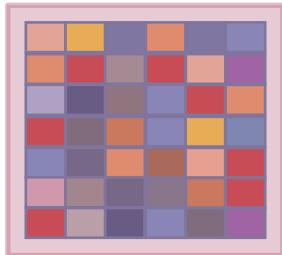


$$\underset{i}{\text{avg}} X'_{m,i}$$

$$y_m^{(SSA)} \in R^D$$



$$X'_n \in R^{N \times D}$$



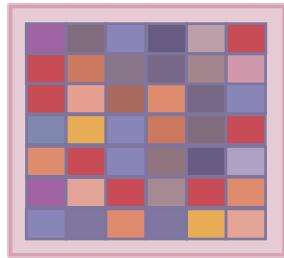
$$\underset{i}{\text{avg}} X'_{n,i}$$

$$y_n^{(SSA)} \in R^D$$



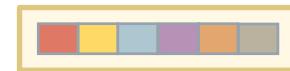
Compatibility function

$$X'_m \in R^{M \times D}$$



$$\text{avg}_{\bar{i}} X'_{m,i} \rightarrow$$

$$y_m^{(SSA)} \in R^D$$



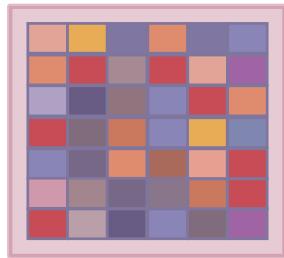
$$U_Q \in R^{D \times D}$$

Query Transformation

$$u_m = U_Q y_m^{(SSA)}$$



$$X'_n \in R^{N \times D}$$



$$\text{avg}_{\bar{i}} X'_{n,i} \rightarrow$$

$$y_n^{(SSA)} \in R^D$$



$$U_K \in R^{D \times D}$$

Key Transformation

$$u_n = U_K y_n^{(SSA)}$$



Compatibility function

$$u_m \in R^D$$



$$u_n \in R^D$$



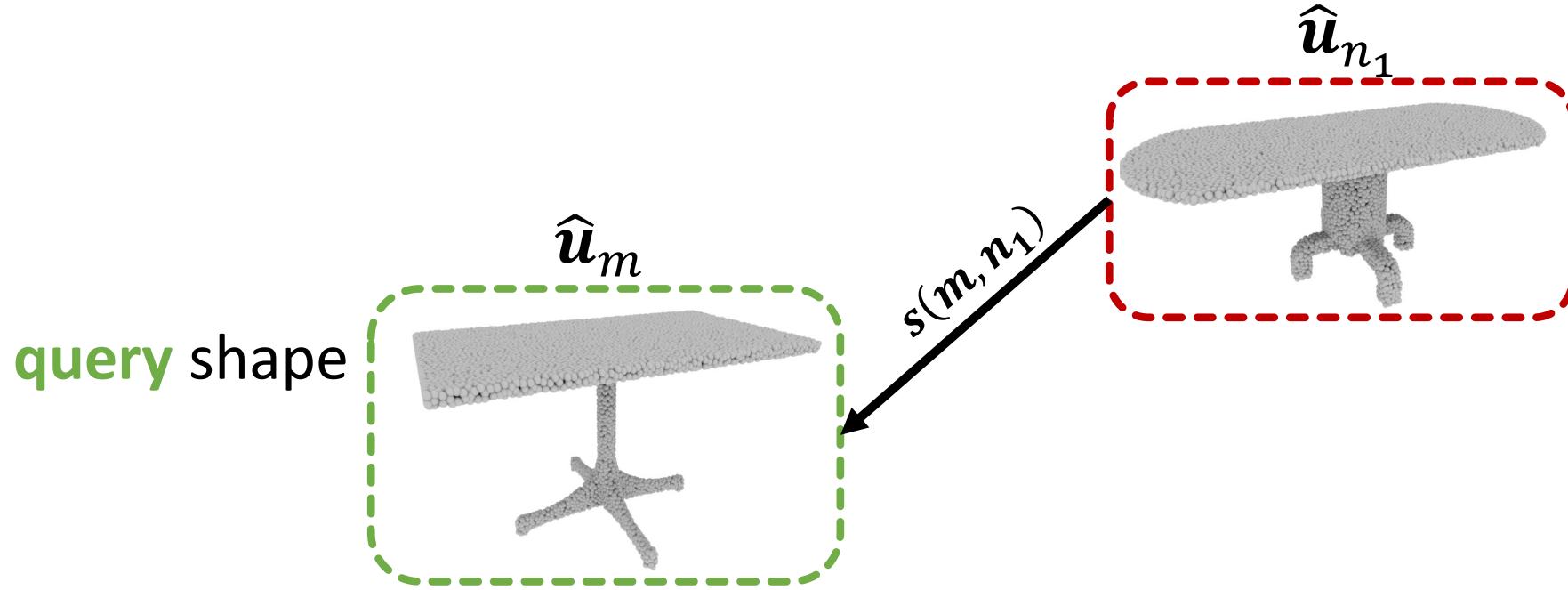
$$\hat{u}_m = u_m / \|u_m\|$$

$$\hat{u}_n = u_n / \|u_n\|$$

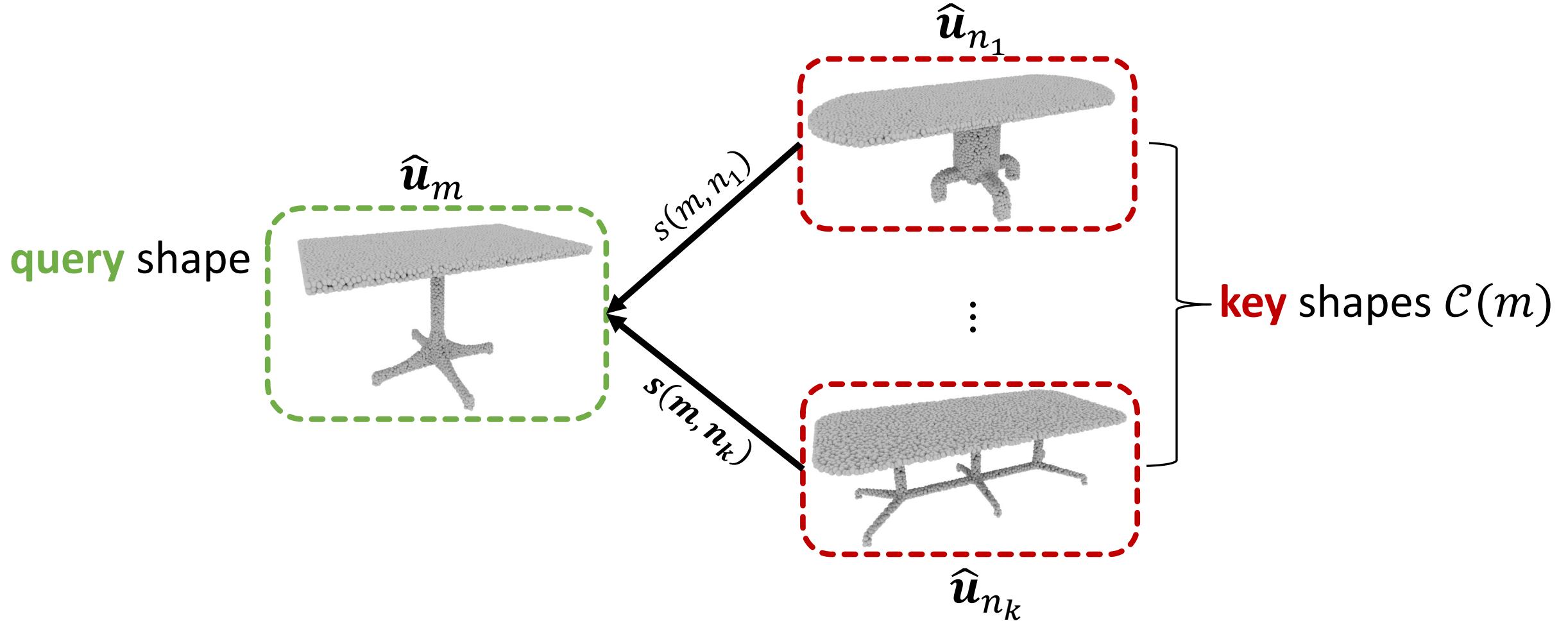
Cosine similarity

$$s(m, n) = \hat{u}_m \cdot \hat{u}_n$$

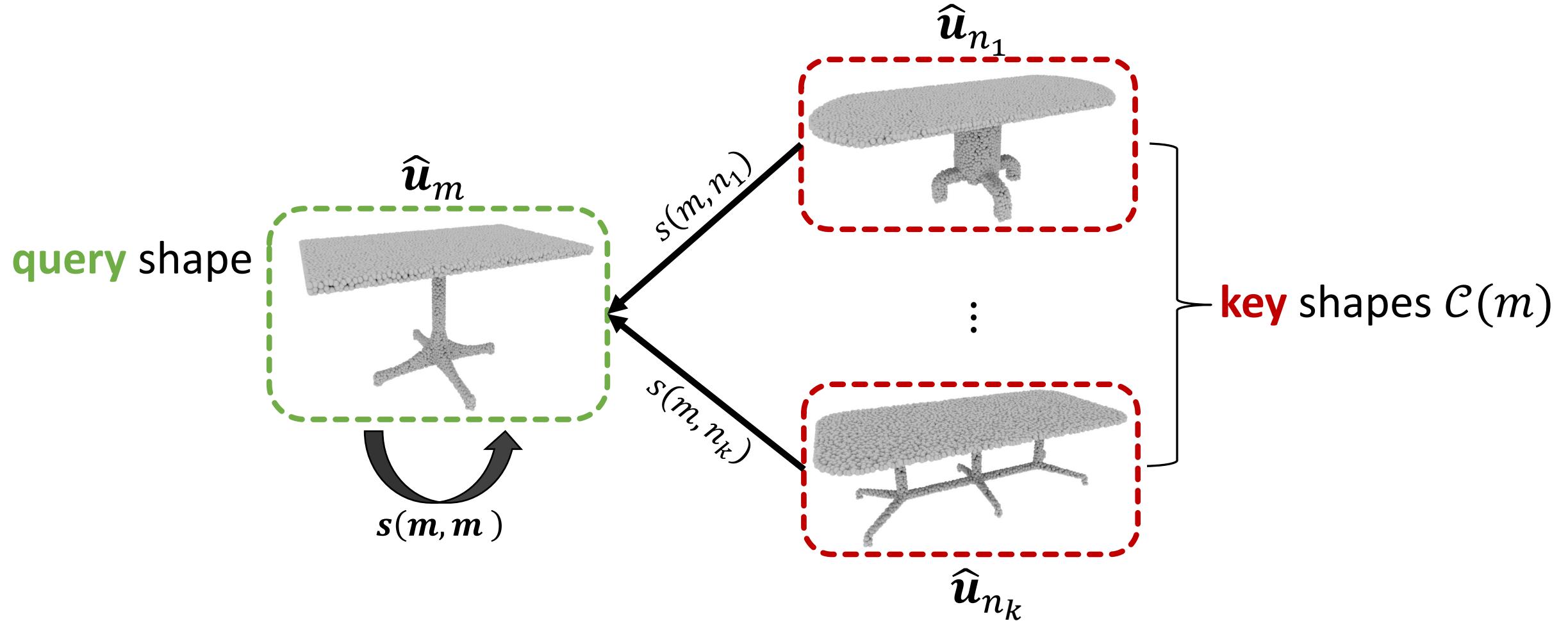
Compatibility function



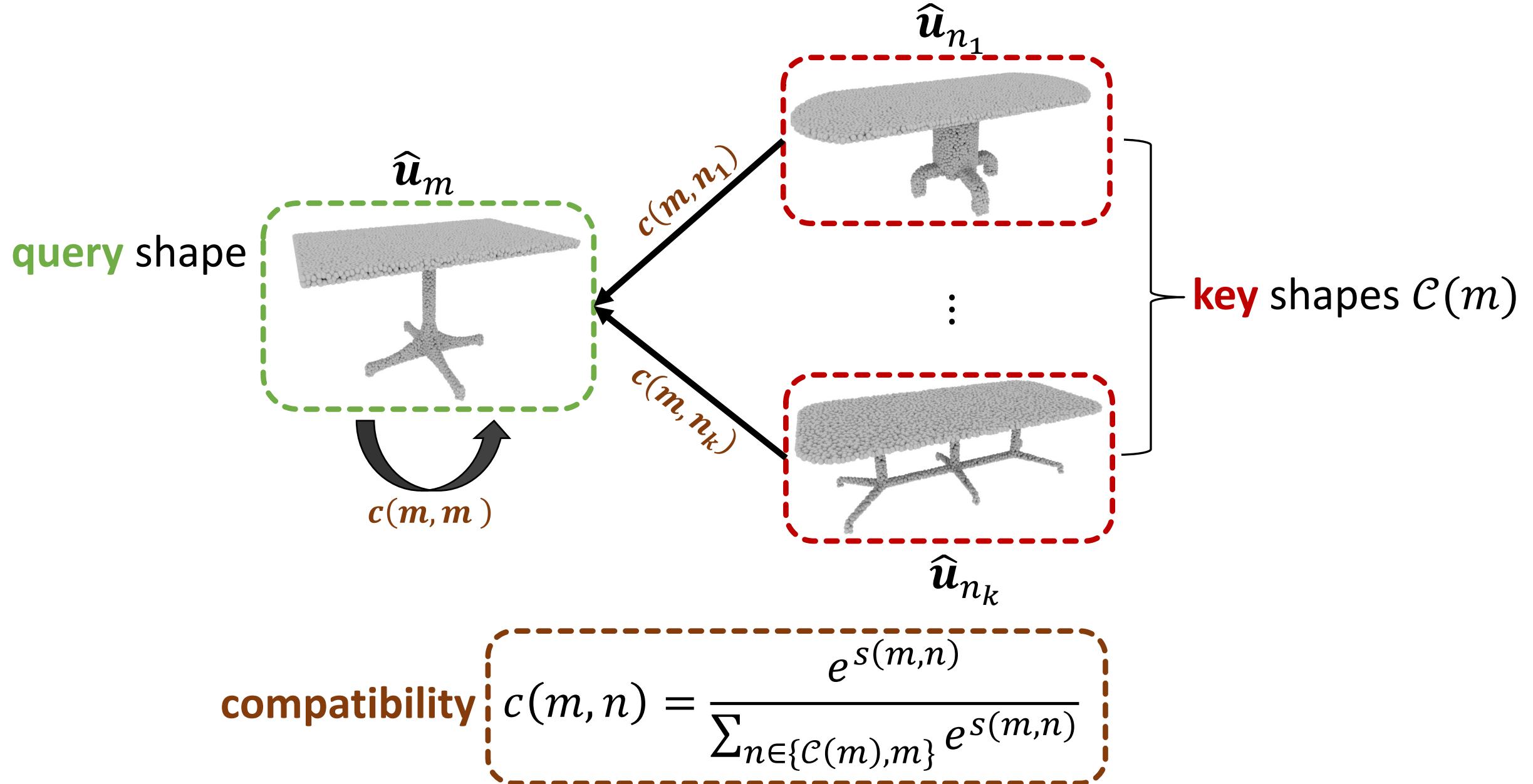
Compatibility function



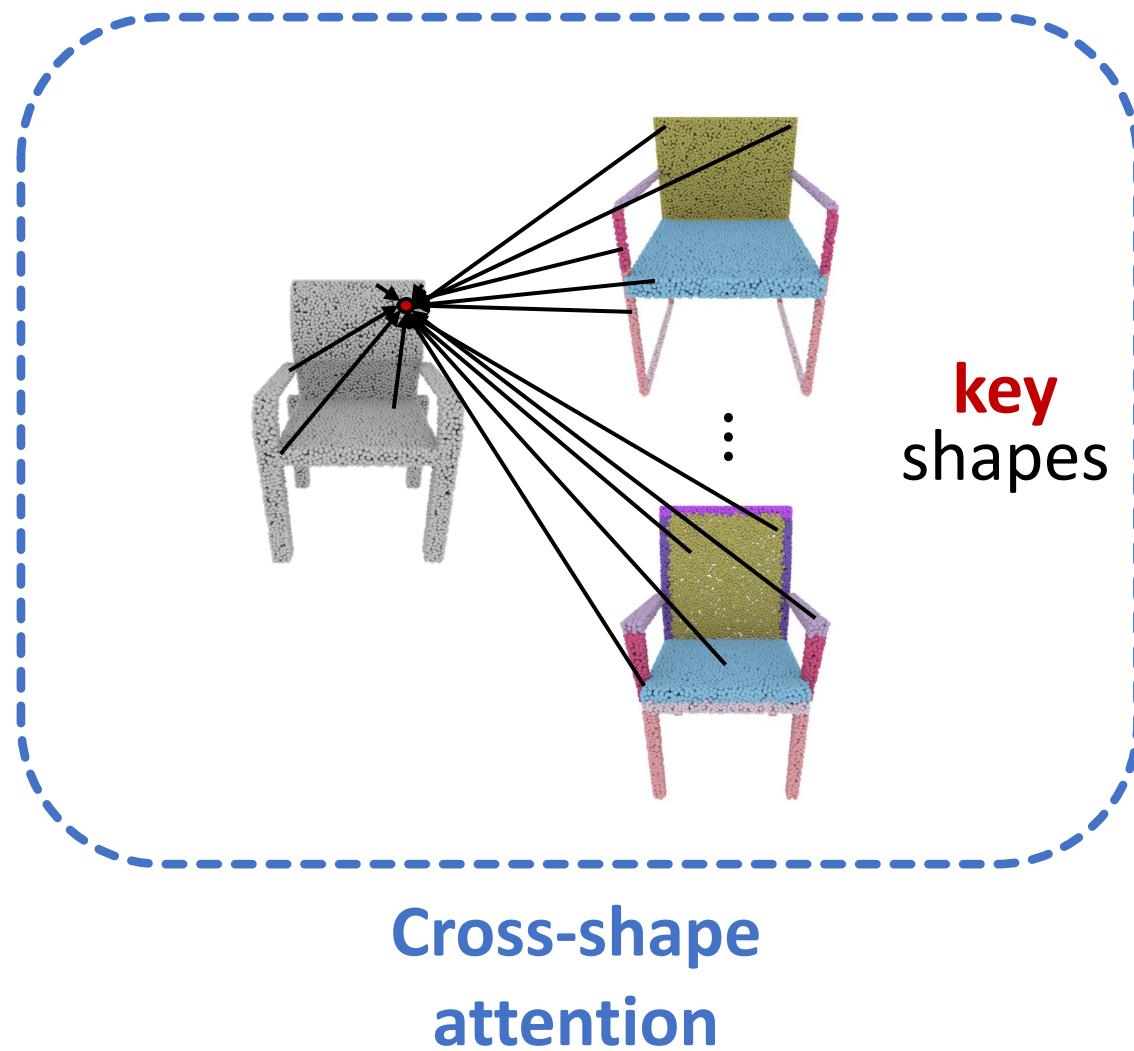
Compatibility function



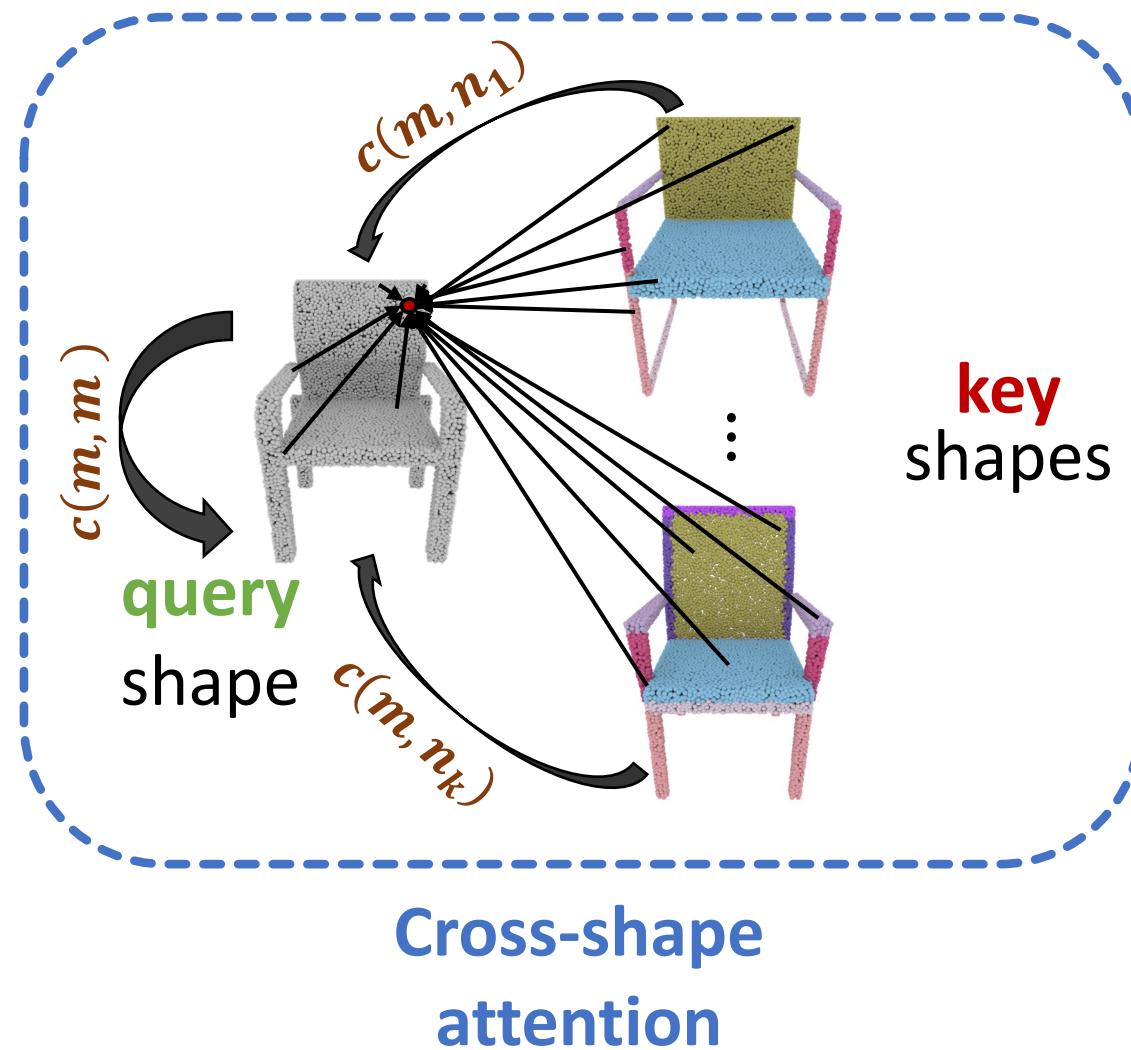
Compatibility function



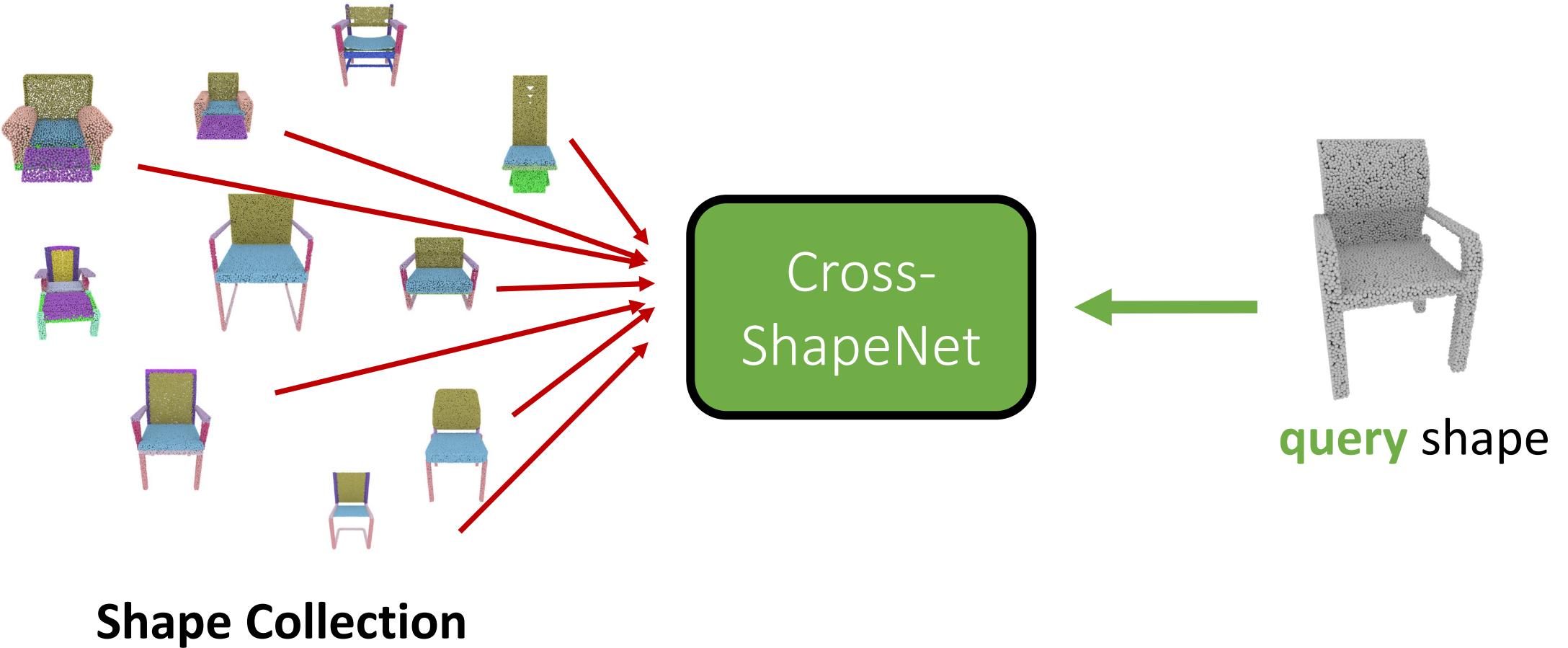
Cross-Shape Attention for multiple shapes



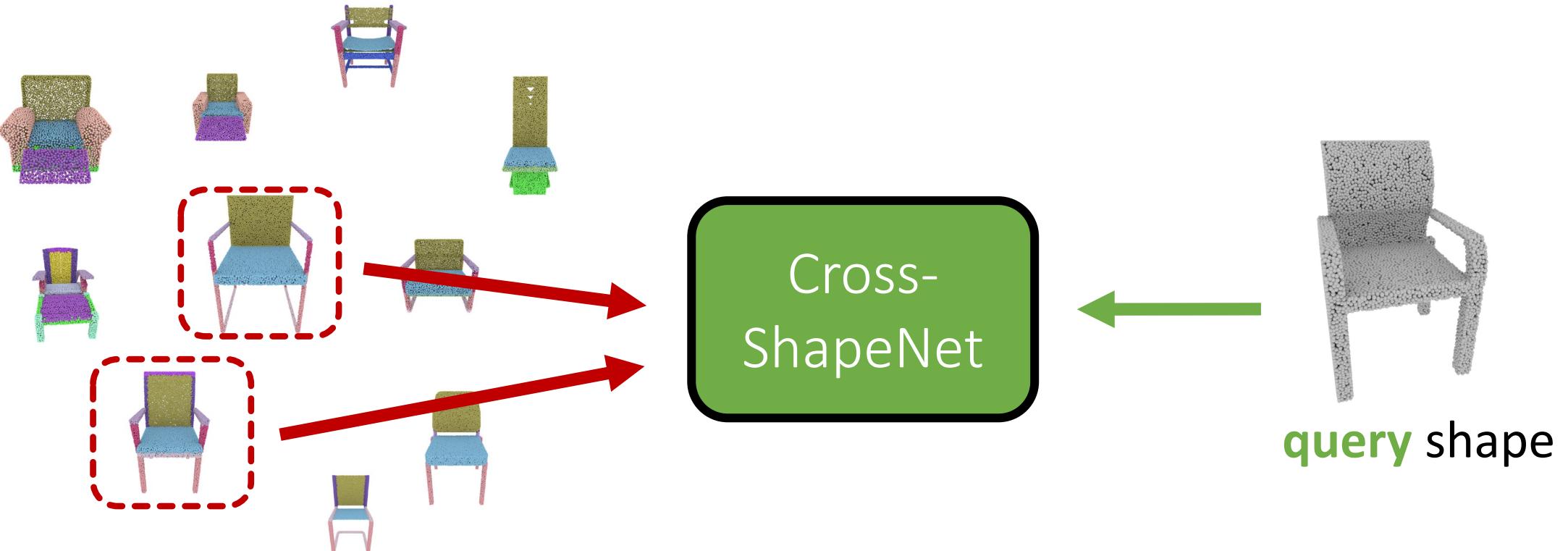
Cross-Shape Attention for multiple shapes



Retrieve compatible shapes

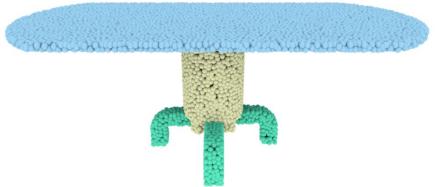
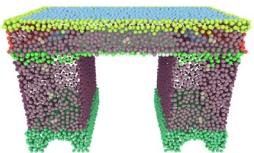
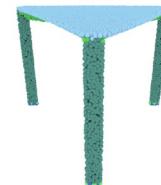
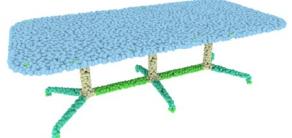
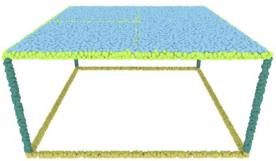
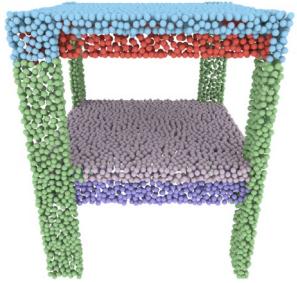


Retrieve compatible shapes

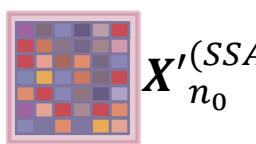
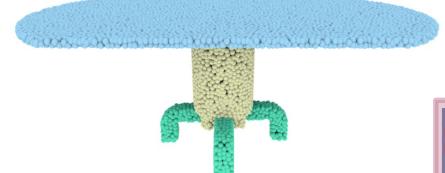
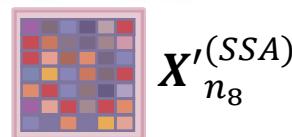
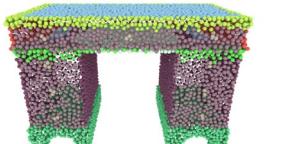
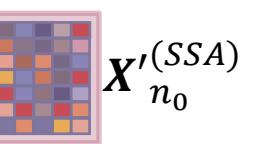
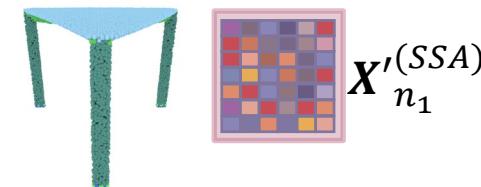
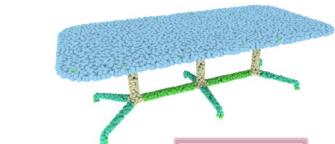
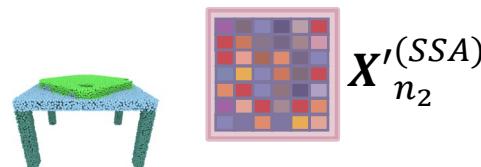
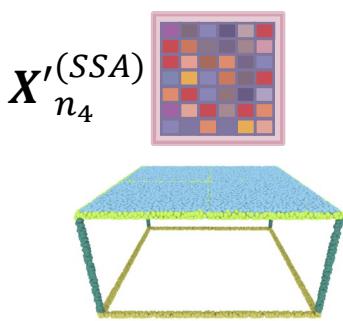
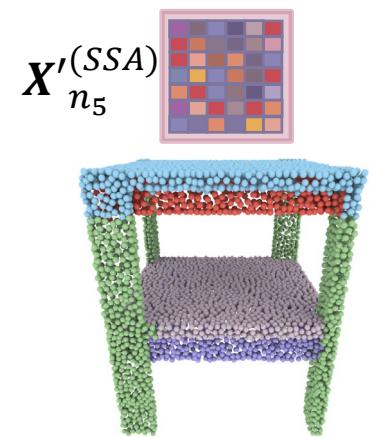


Shape Collection

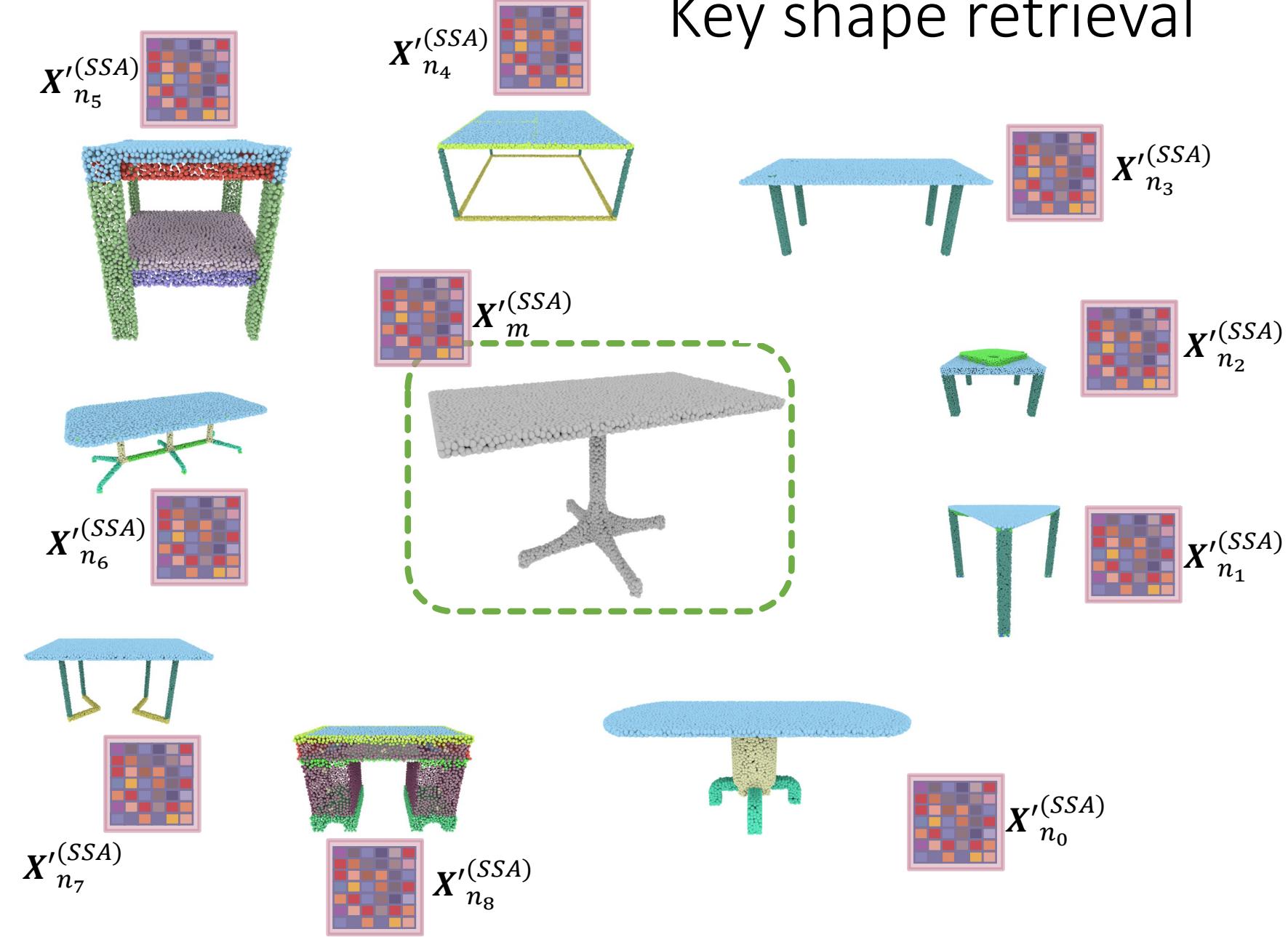
Key shape retrieval



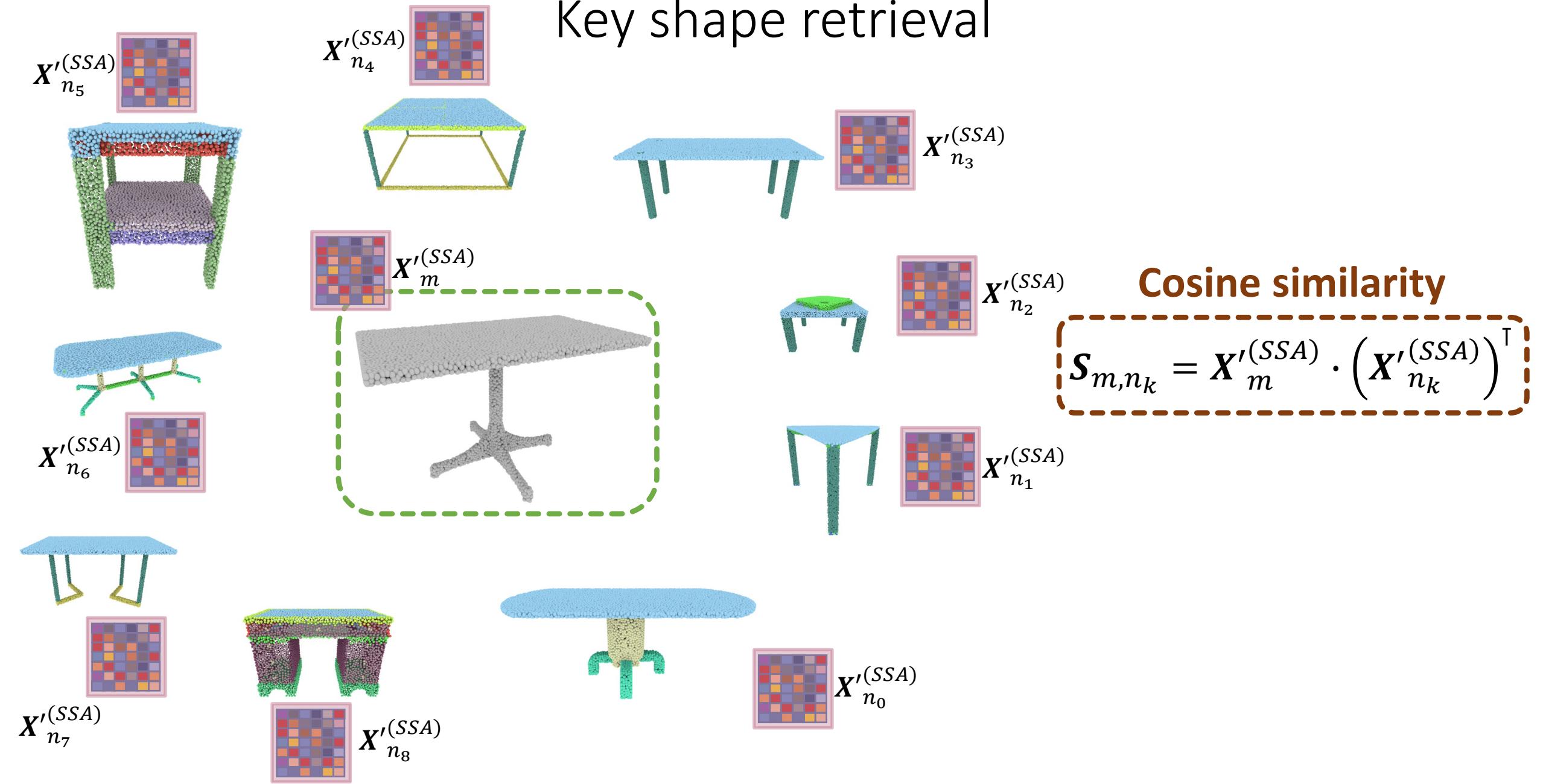
Key shape retrieval



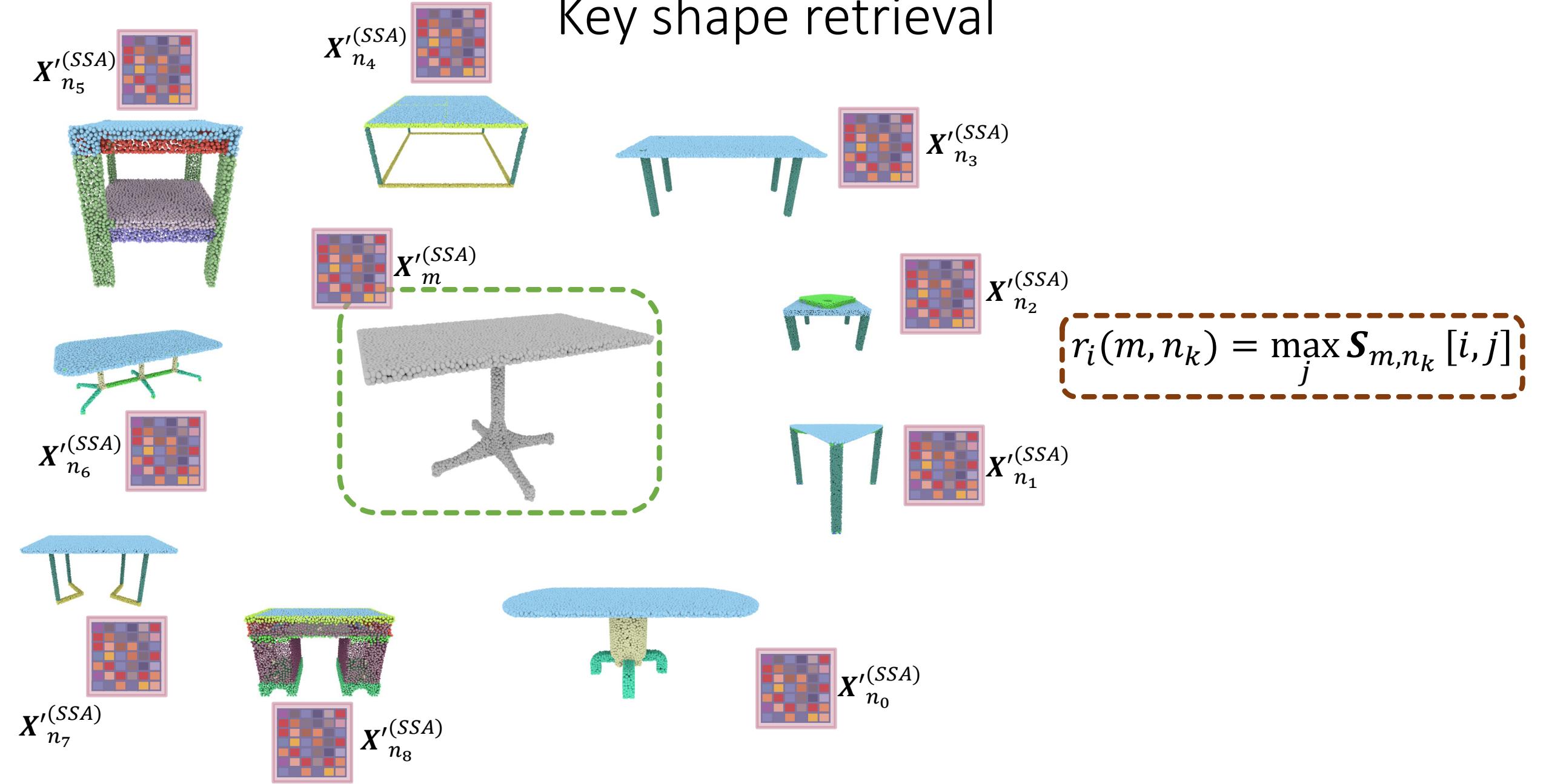
Key shape retrieval



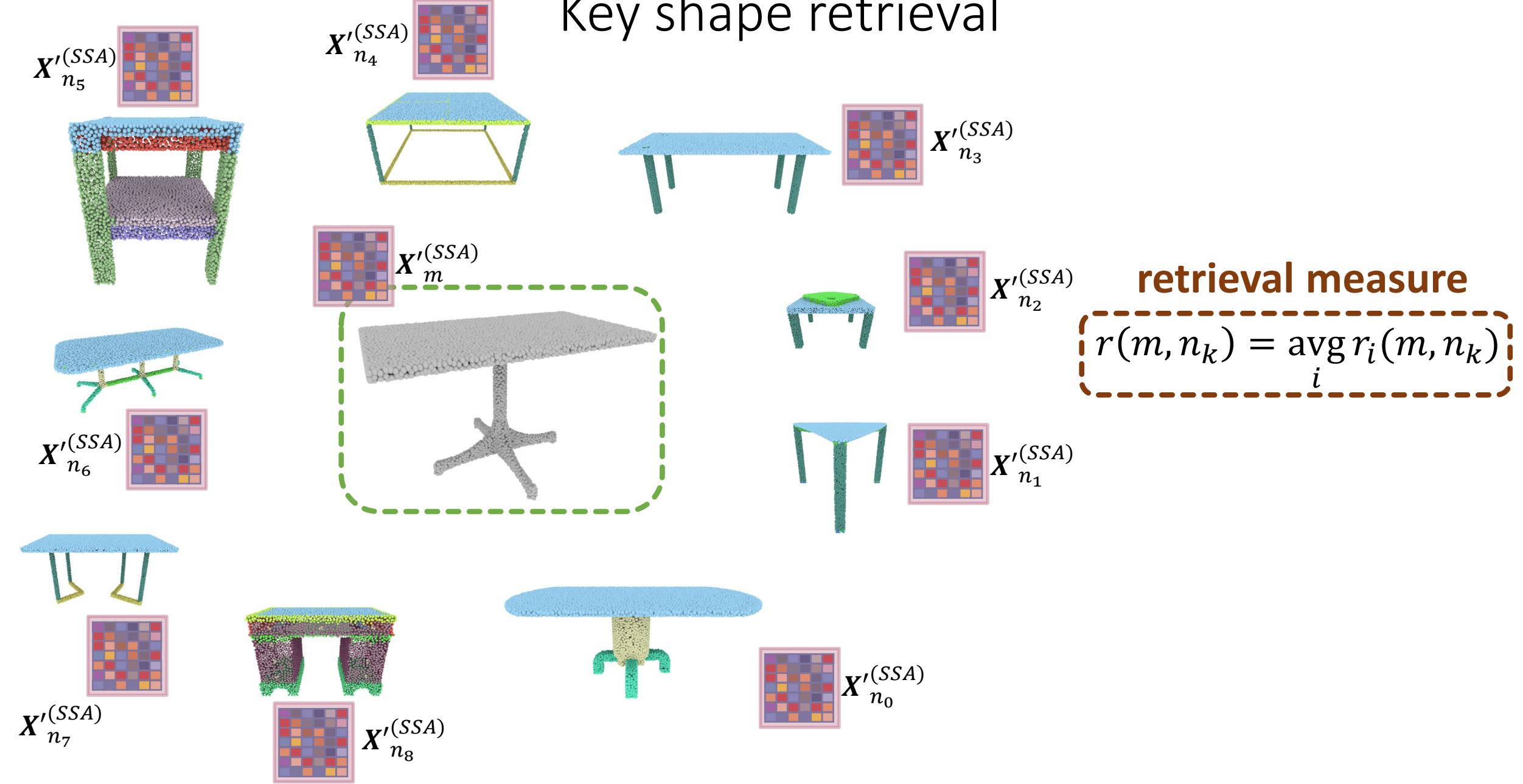
Key shape retrieval



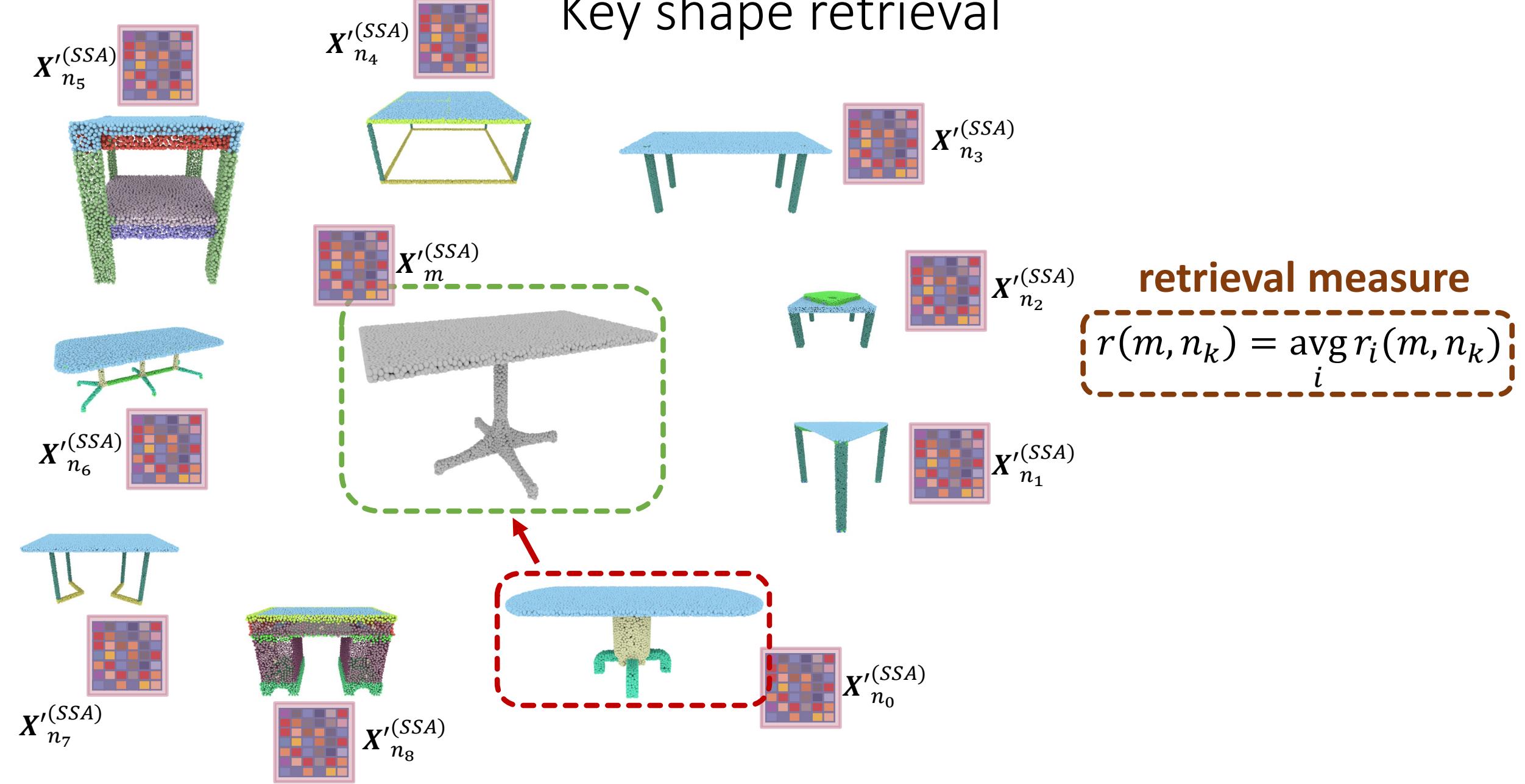
Key shape retrieval



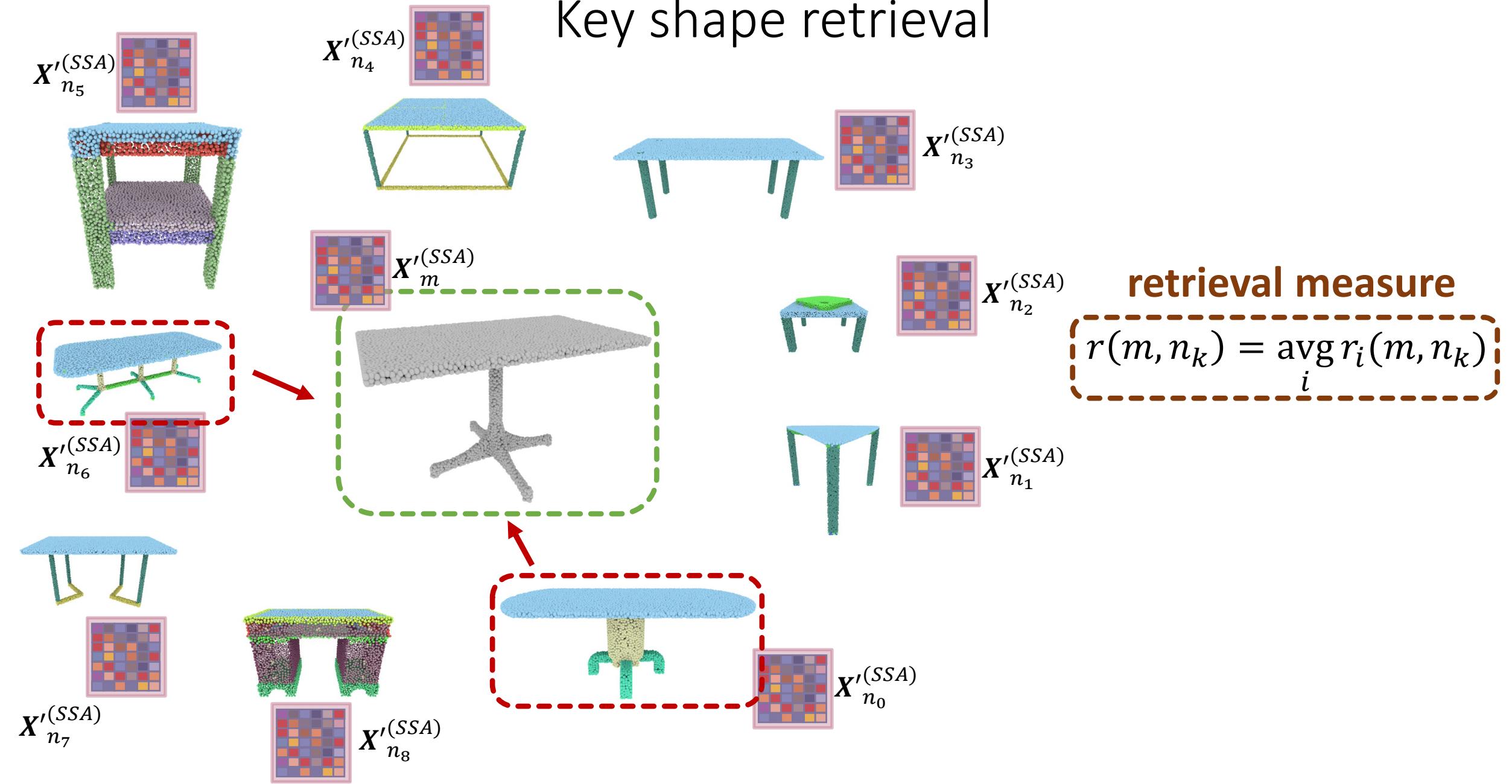
Key shape retrieval



Key shape retrieval

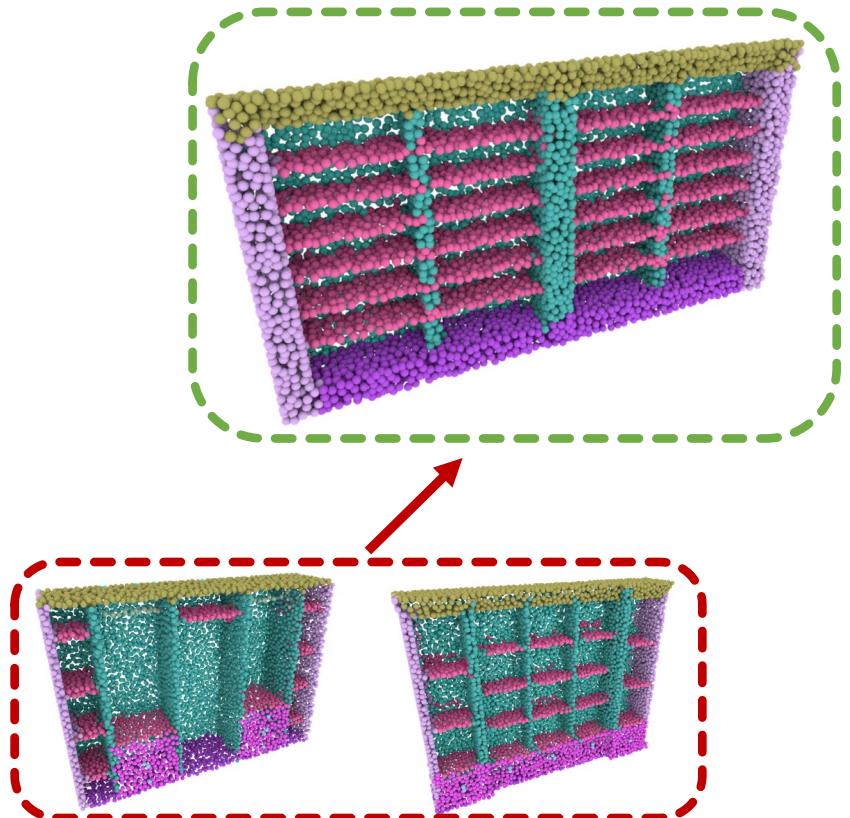


Key shape retrieval

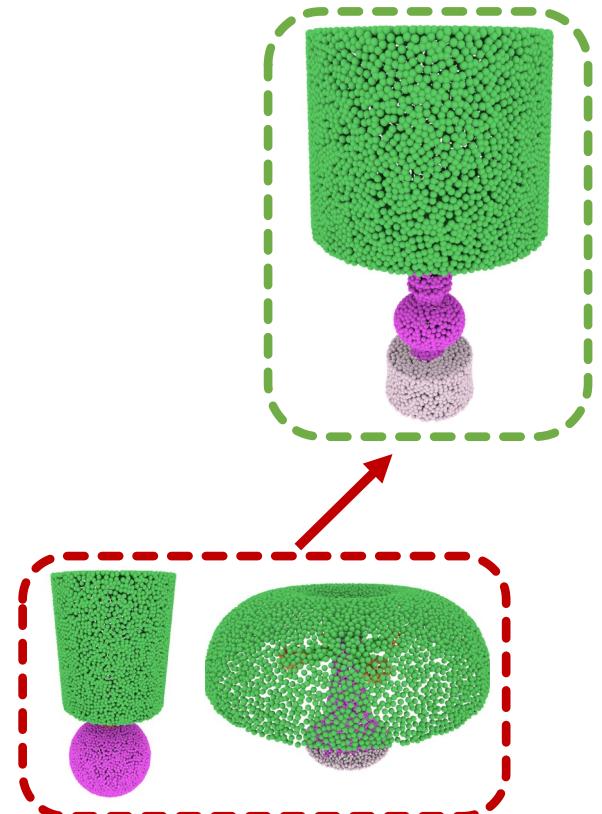
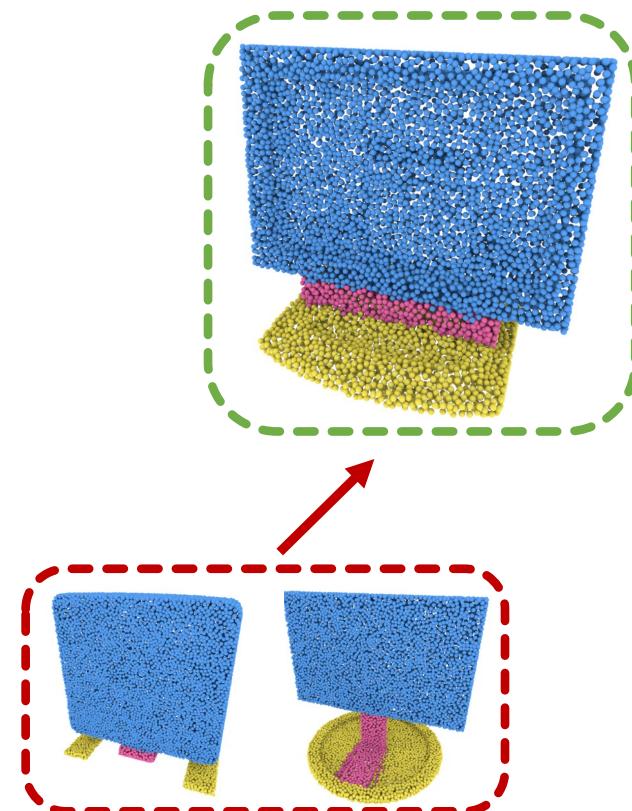


Key shape retrieval: Examples

query shapes

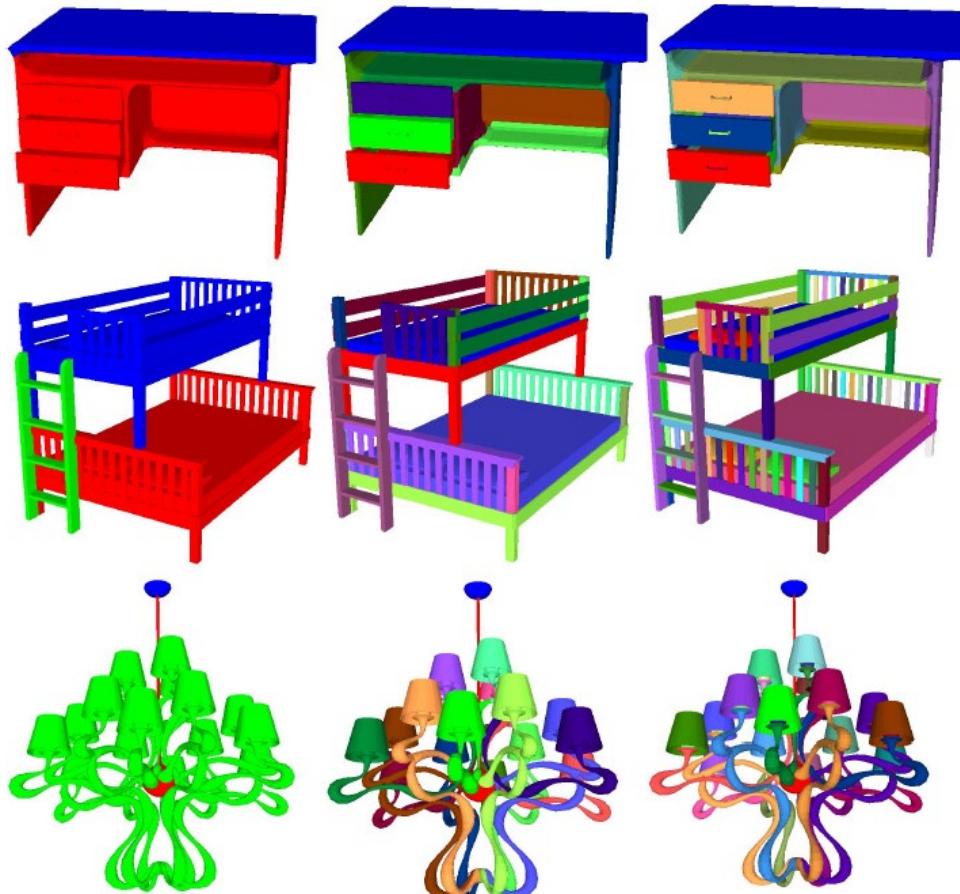


key shapes



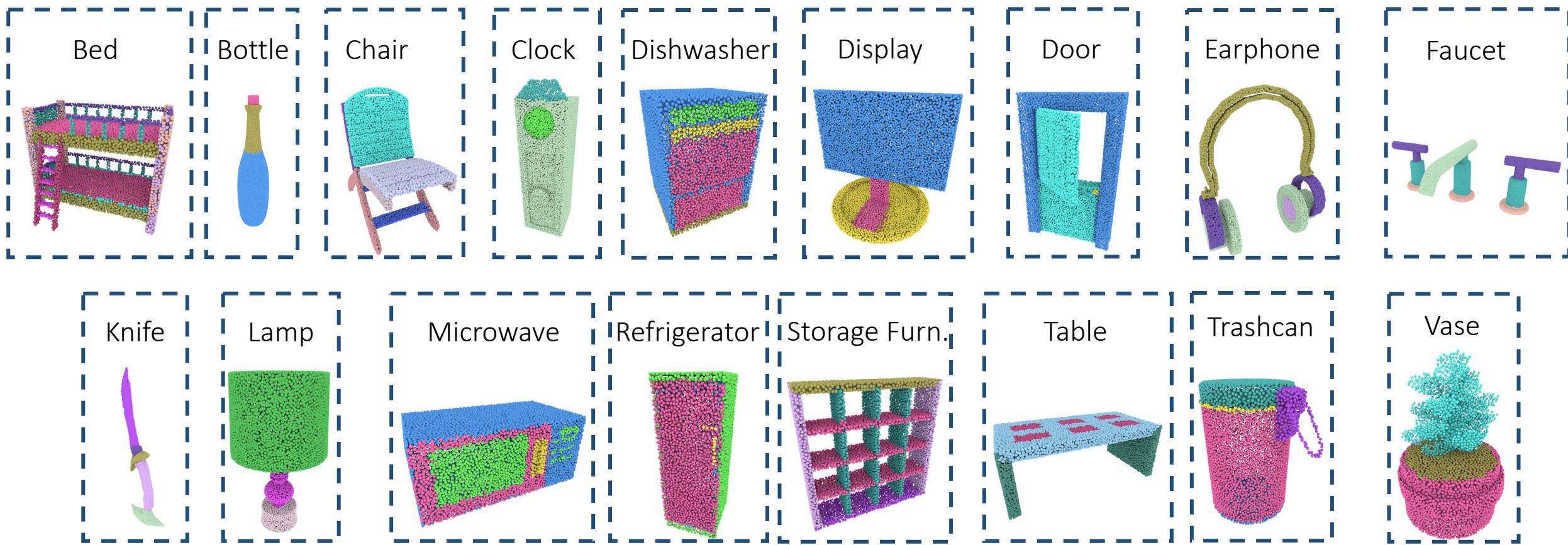
PartNet dataset

Coarse → Fine-grained



[Mo et al. 2019]

PartNet dataset



[Mo et al. 2019]

Examples of shape collections

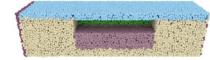
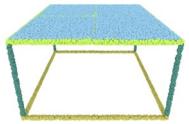
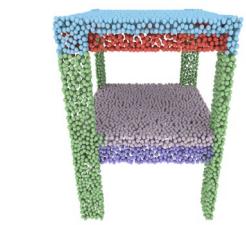
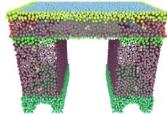
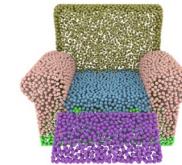


Table
category



5,707 training shapes

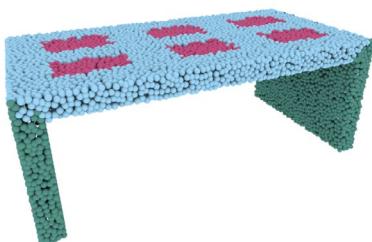
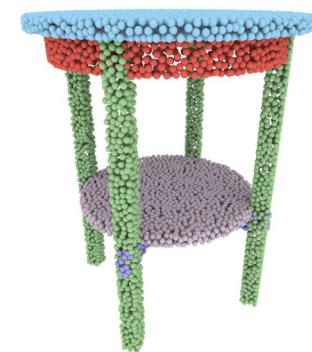


Chair
category



4,489 training shapes

Training details: Loss



training
data

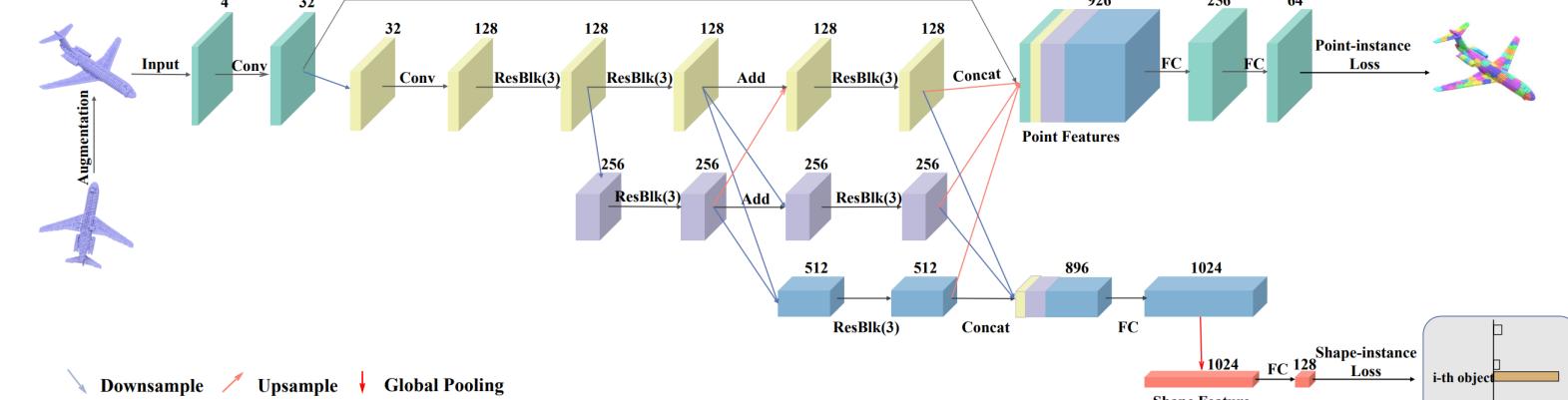
$$L_{CE} = - \sum_{p_i \in \mathcal{S}_k} \hat{q}_i \log q_i$$

\mathcal{S}_k : shape $k = \{\mathbf{p}_i\}_{i=1}^{P_k}$

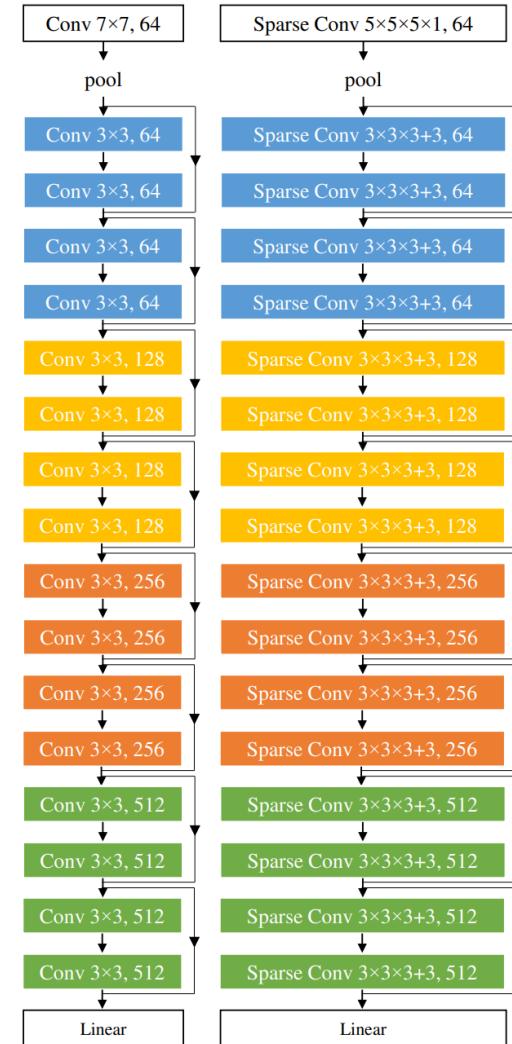
\hat{q}_i : ground-truth one-hot label vector for point \mathbf{p}_i

q_i : predicted label probabilities for point \mathbf{p}_i

Training details: Backbones

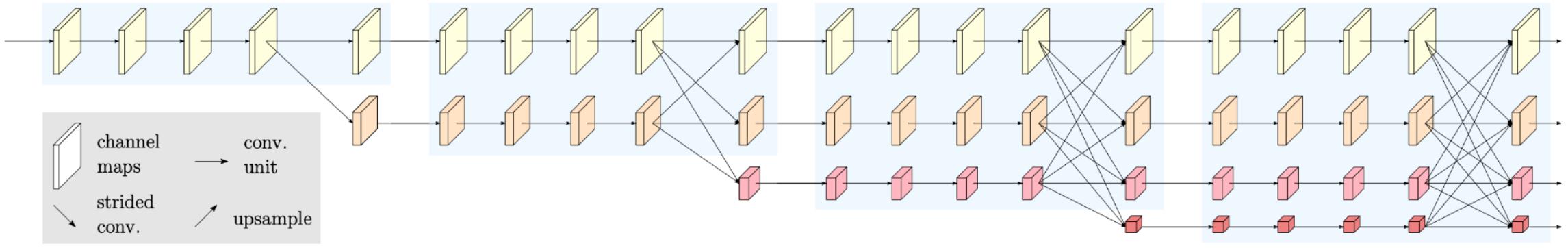


MID-FC [Wang et al. 2021]



MinkowskiNet [Choy et al. 2019]

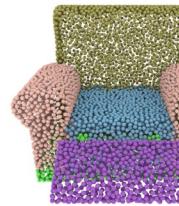
Training details: Backbones



HRNet [Wang et al. 2021]

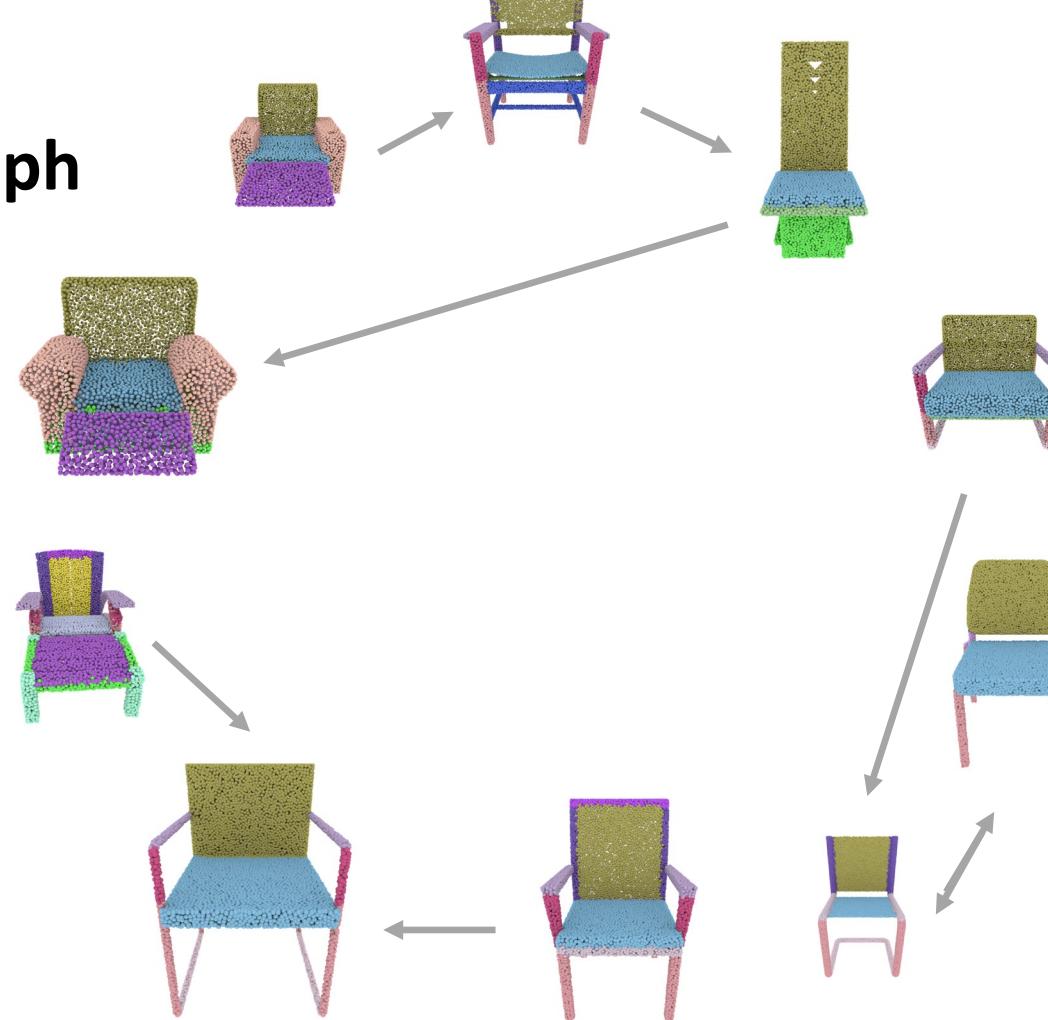
Training details: Collection graph

Shape Collection



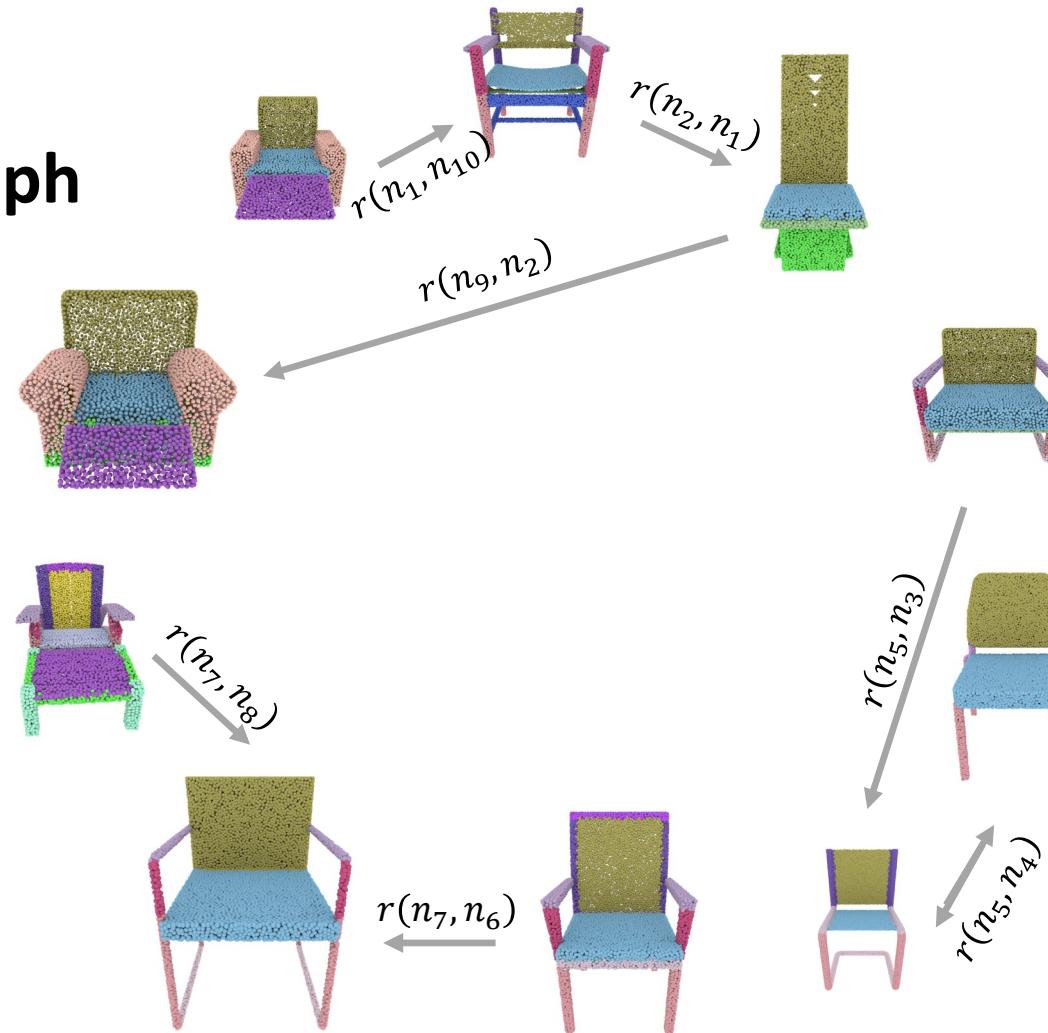
Training details: Collection graph

Collection graph



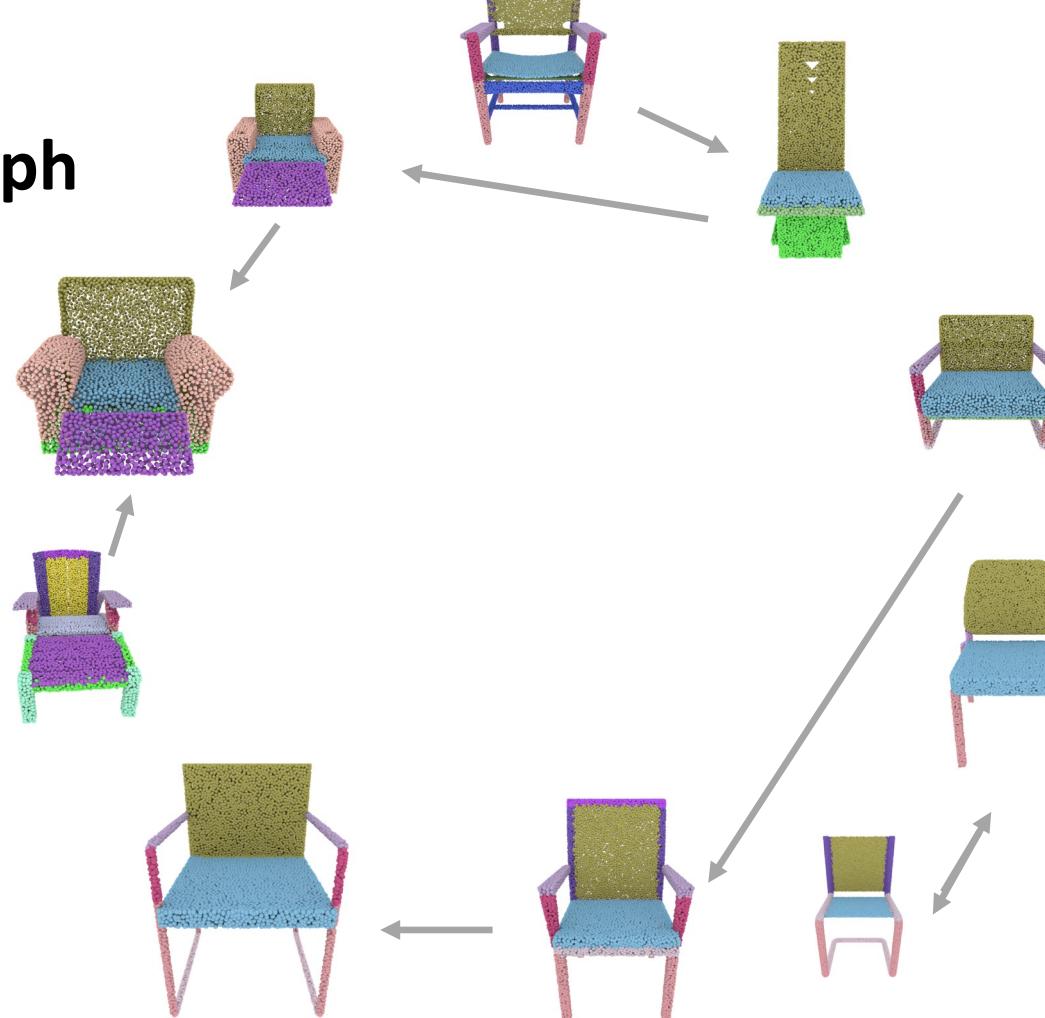
Training details: Collection graph

Collection graph

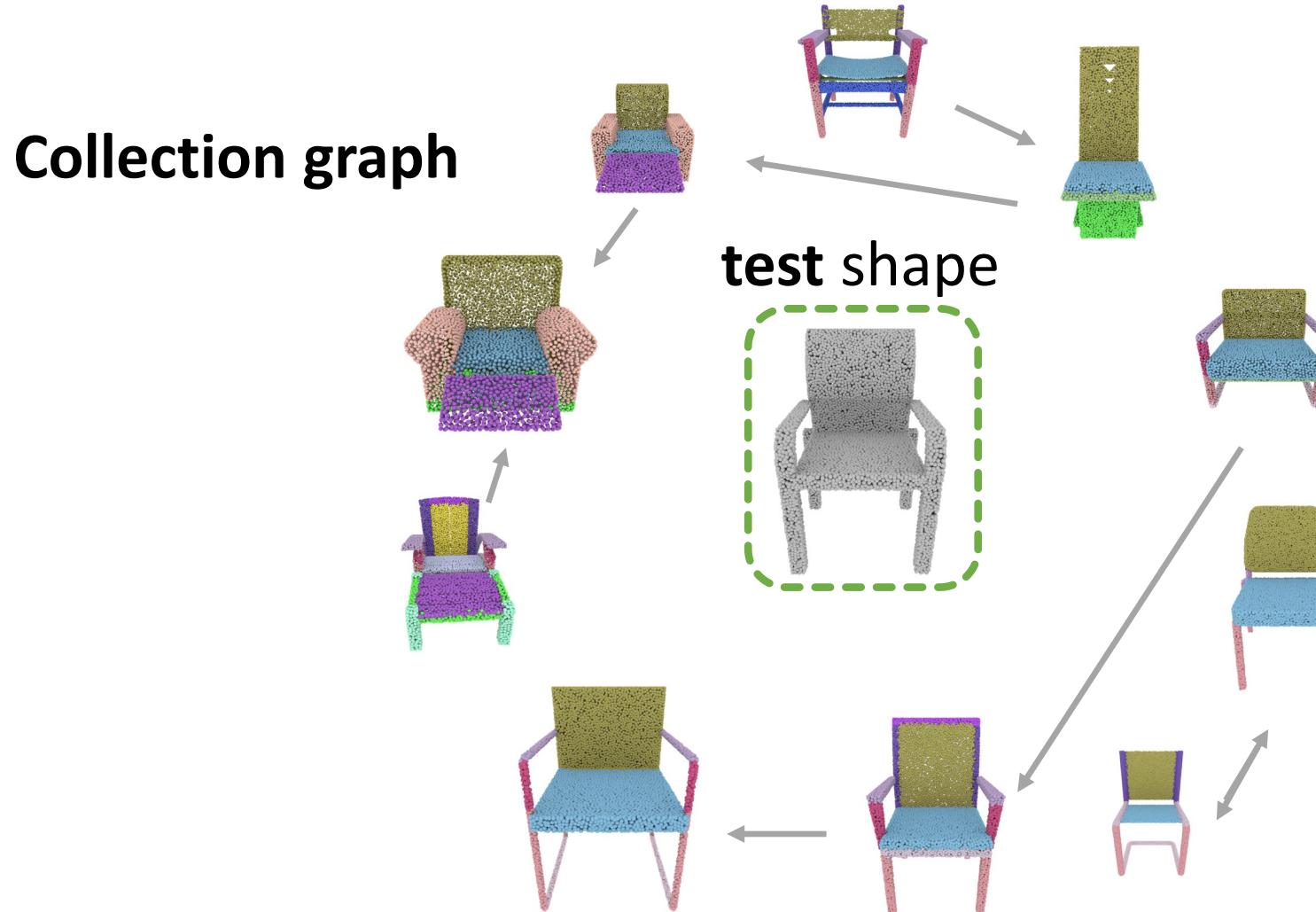


Training details: Collection graph

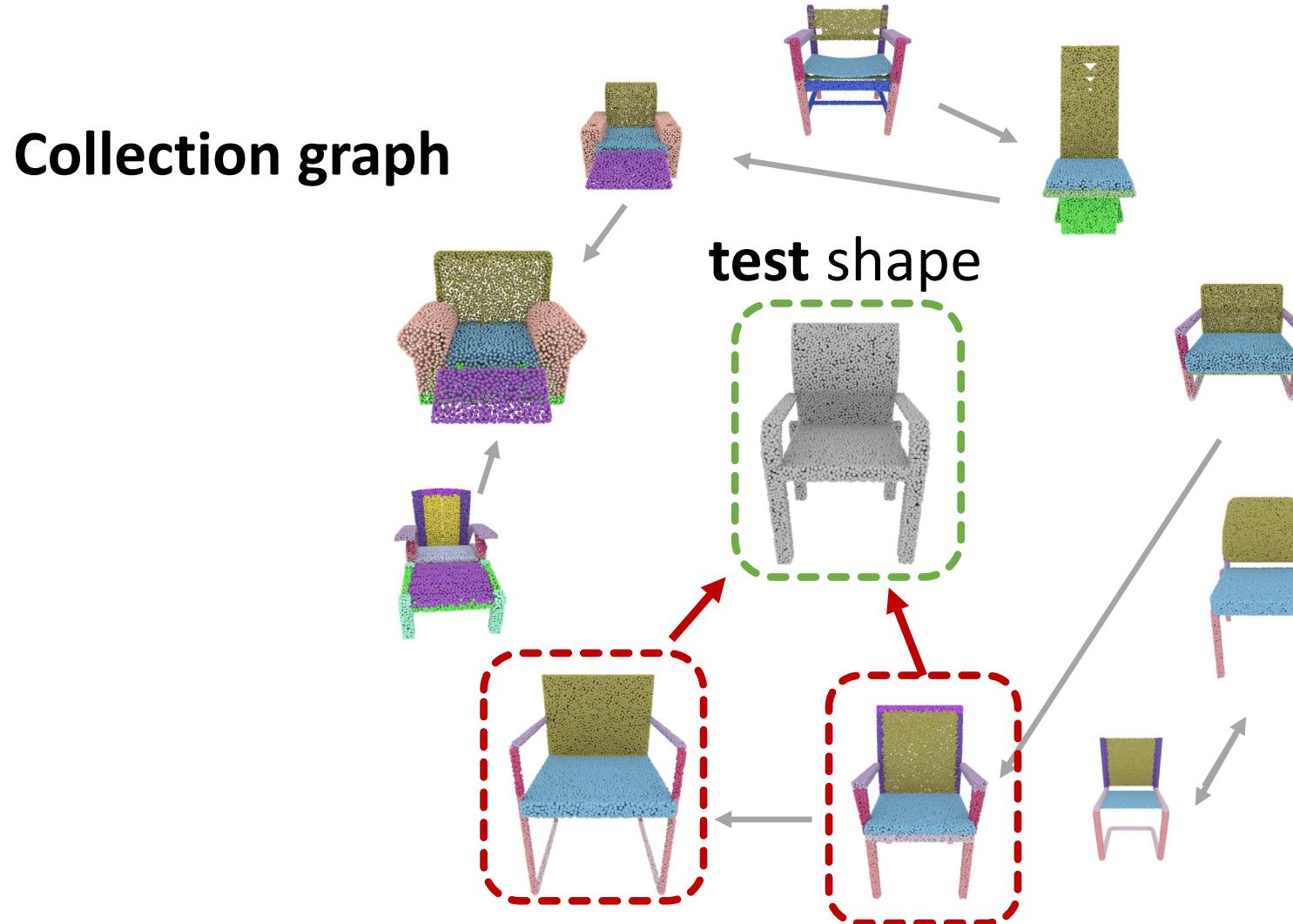
Collection graph



Inference: Collection graph



Inference: Collection graph



Results

Method	Part IoU
--------	----------

Results: MinkowskiNet variants

Method	Part IoU
MinkHRNet	48.0

Results: MinkowskiNet variants

Method	Part IoU
MinkHRNet	48.0
MinkHRNetCSN-SSA	48.7

+0.7%

Results: MinkowskiNet variants

Method	Part IoU
MinkHRNet	48.0
MinkHRNetCSN-SSA	48.7
MinkHRNetCSN-K1	49.9
MinkHRNetCSN-K2	49.7

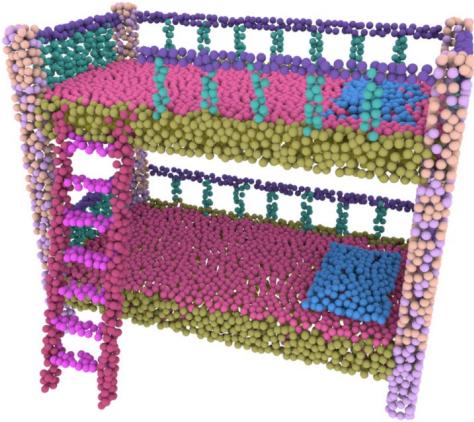
Results: MinkowskiNet variants

Method	Part IoU
MinkHRNet	48.0
MinkHRNetCSN-SSA	48.7
MinkHRNetCSN-K1	49.9
MinkHRNetCSN-K2	49.7

+1.2%

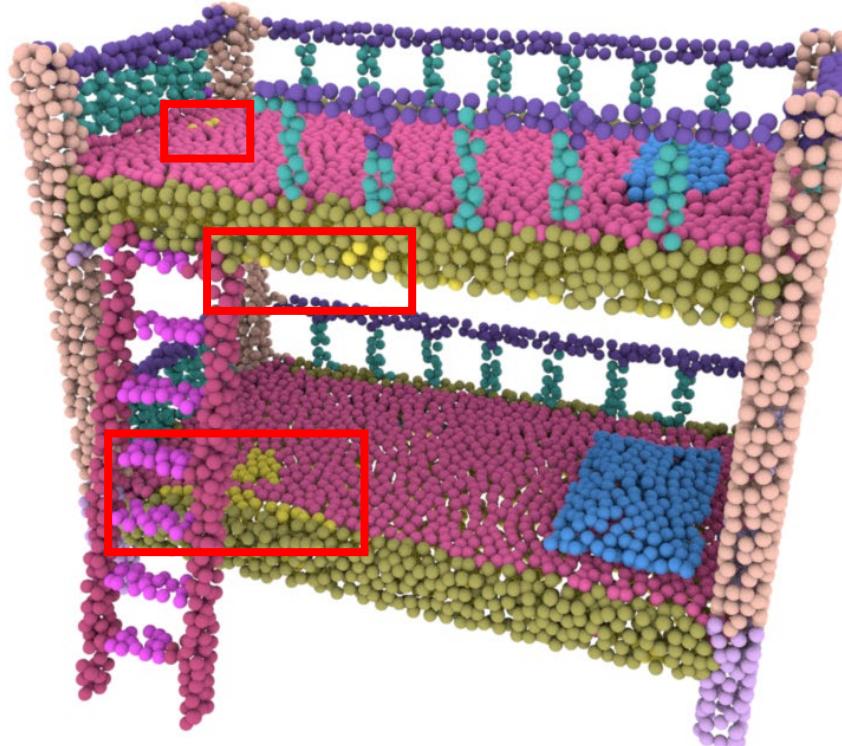
Results: MinkowskiNet variants

Ground truth



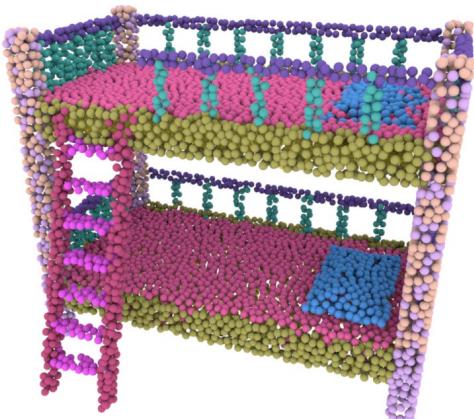
- Pillow
- Mattress
- Stretcher
- Leg
- Horizontal bar
- Vertical bar
- Bed post
- Ladder vertical bar
- Rung

MinkHRNet



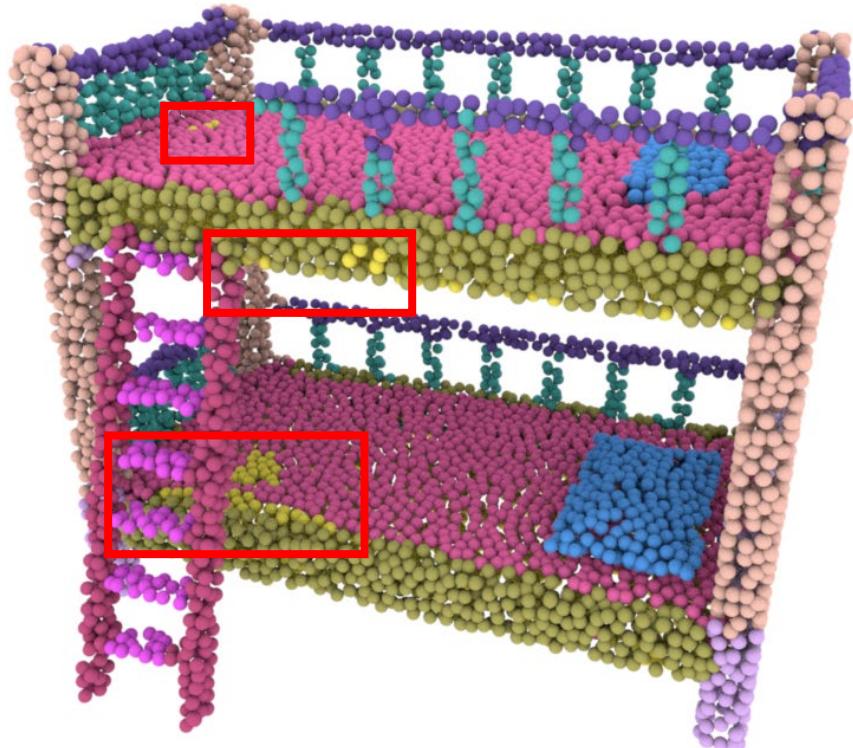
Results: MinkowskiNet variants

Ground truth

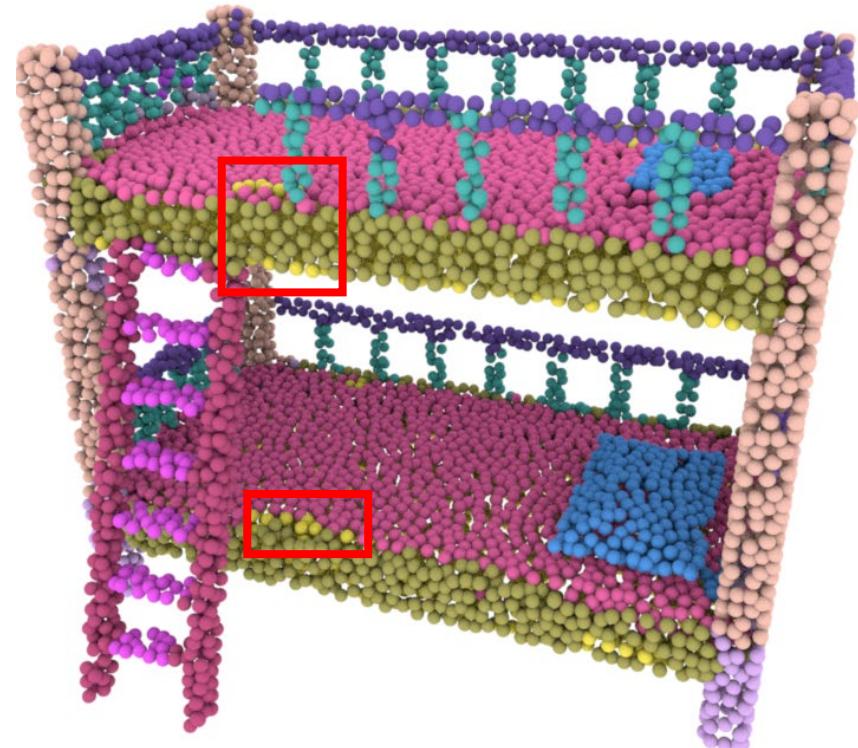


- Pillow
- Mattress
- Stretcher
- Leg
- Horizontal bar
- Vertical bar
- Bed post
- Ladder vertical bar
- Rung

MinkHRNet

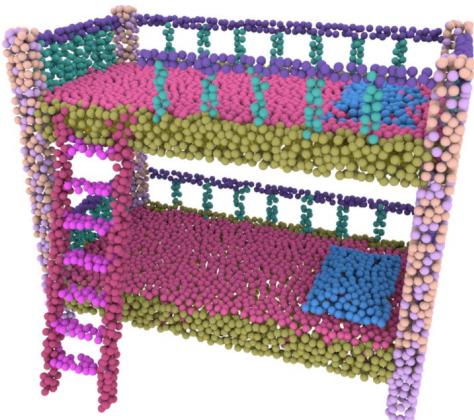


MinkHRNetCSN-SSA



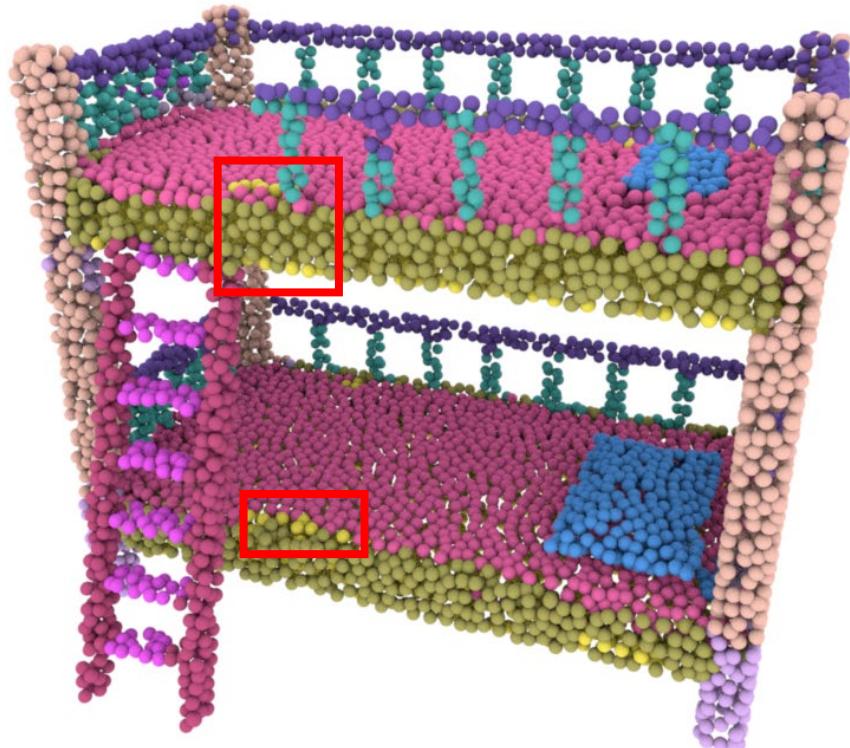
Results: MinkowskiNet variants

Ground truth

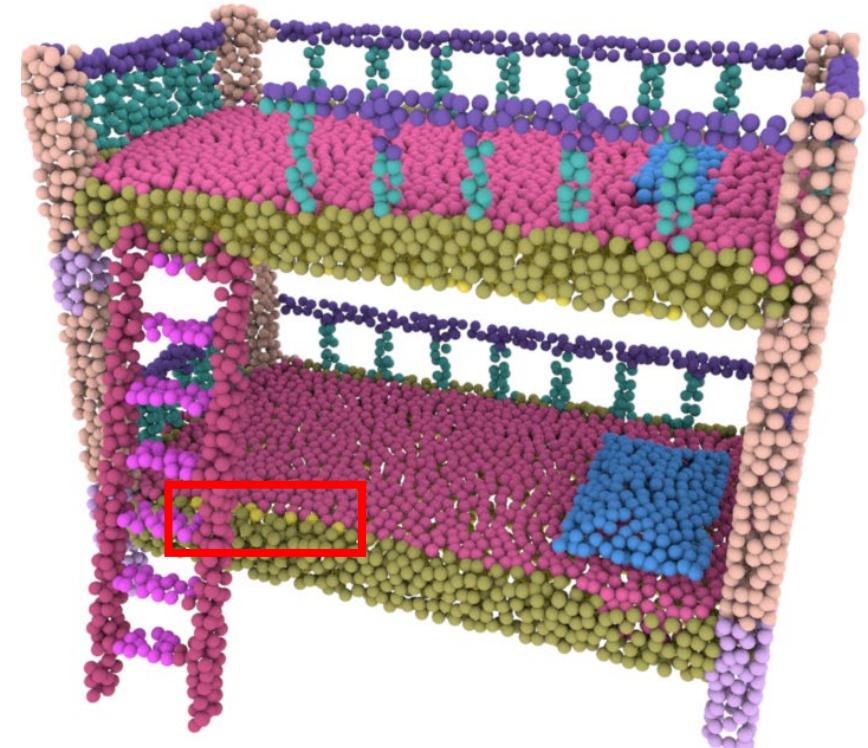


- Pillow
- Mattress
- Stretcher
- Leg
- Horizontal bar
- Vertical bar
- Bed post
- Ladder vertical bar
- Rung

MinkHRNetCSN-SSA



MinkHRNetCSN-K1



Results: MID-FC variants

Method	Part IoU
MID-FC	60.8

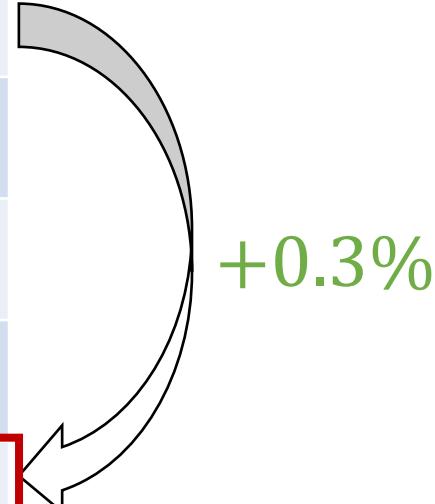
Results: MID-FC variants

Method	Part IoU
MID-FC	60.8
MID-FC-CSN-SSA	61.8

+1.0%

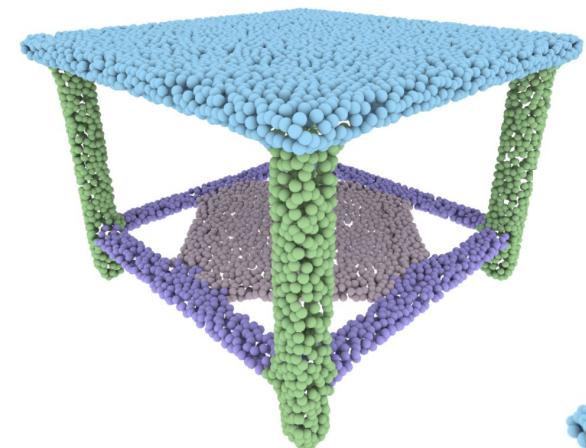
Results: MID-FC variants

Method	Part IoU
MID-FC	60.8
MID-FC-CSN-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



+0.3%

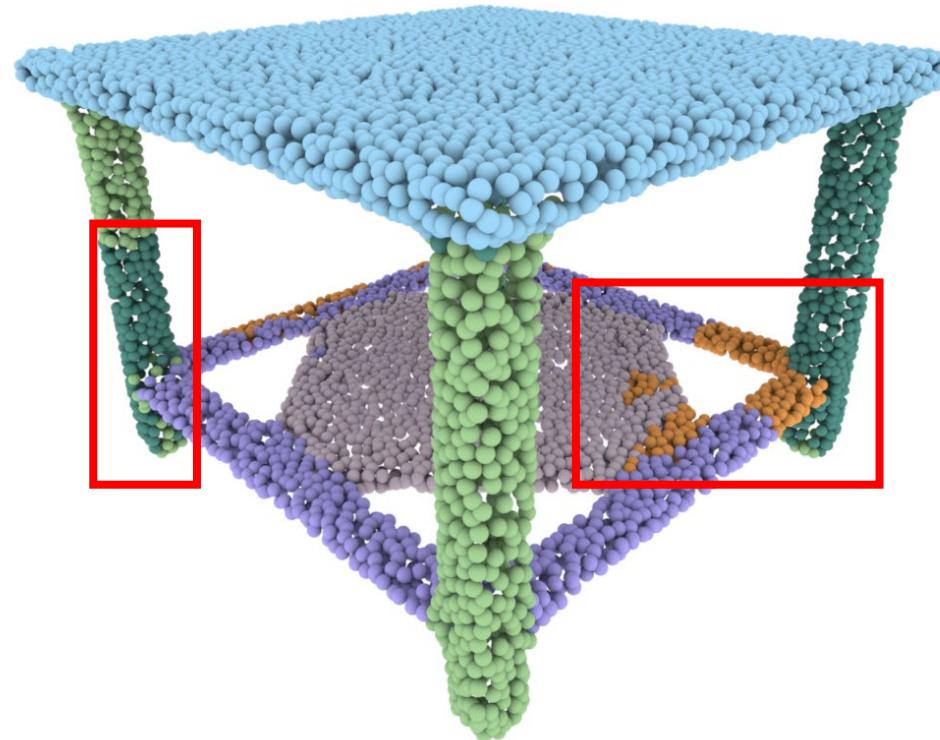
Ground truth



- Bar
- Leg
- Board
- Shelf

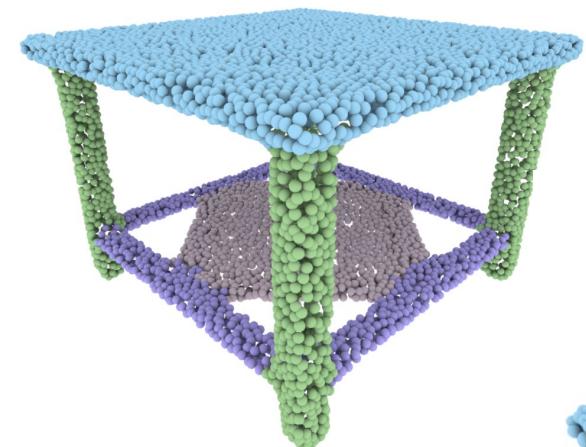
Results: MID-FC variants

MID-FC



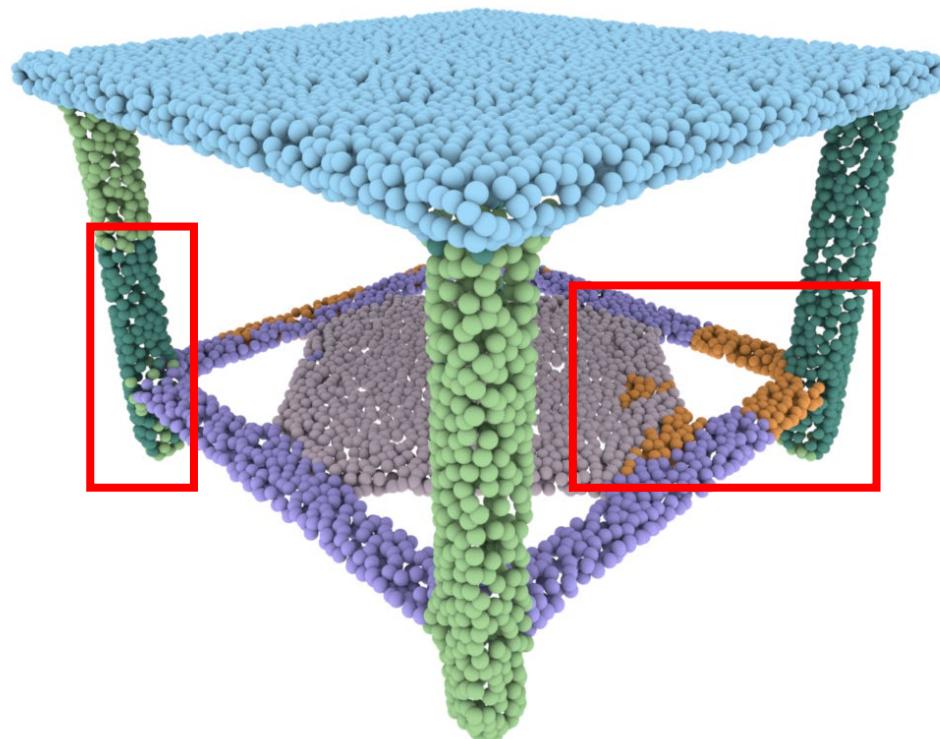
Ground truth

Results: MID-FC variants

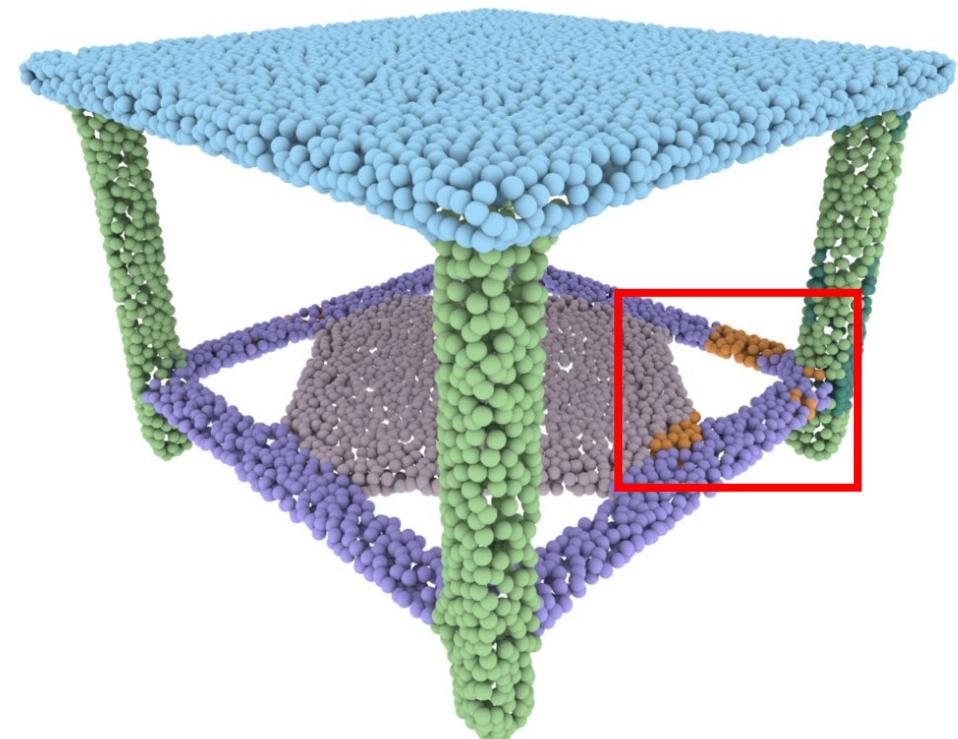


- Bar
- Leg
- Board
- Shelf

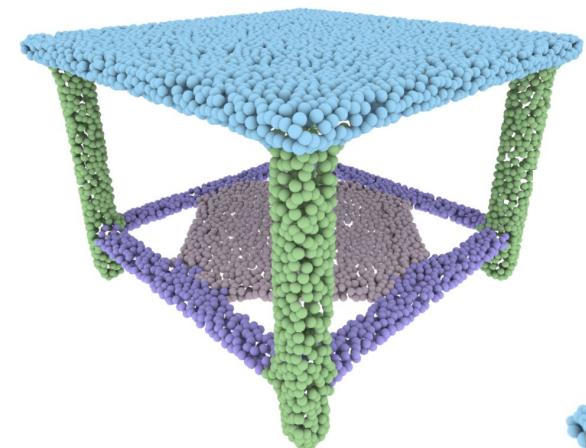
MID-FC



MID-FC-CSN-SSA

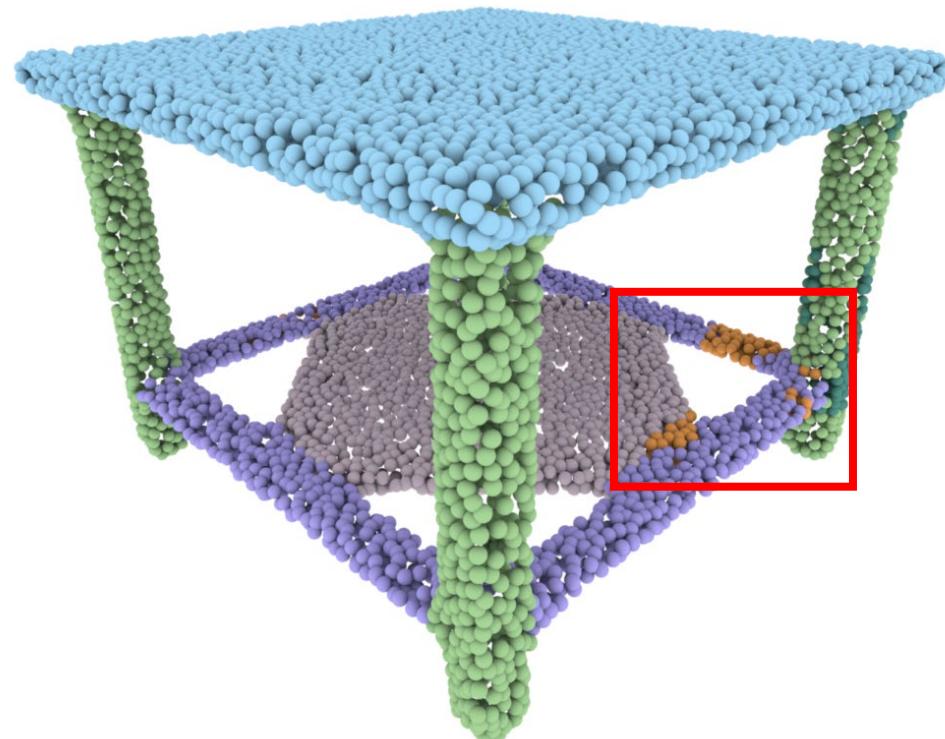


Ground truth



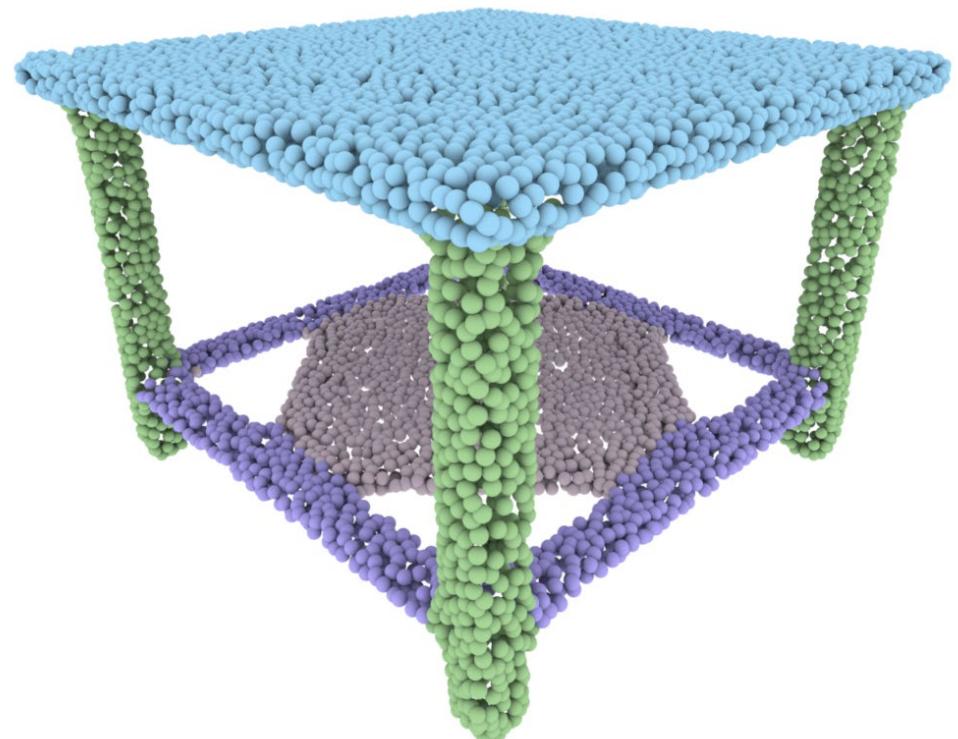
Results: MID-FC variants

MID-FC-CSN-SSA



- [purple square] Bar
- [green square] Leg
- [light blue square] Board
- [grey square] Shelf

MID-FC-CSN-K4



Results: Comparison with other methods

Method	Part IoU
ResGCN-28 (Li et al. 2023)	45.1
CloserLook3D (Liu et al. 2020)	53.8
MinkResUNet (Choy et al. 2019)	46.8
MinkHRNetCSN-K1 (ours)	49.9
MID-FC (Wang et al. 2021)	60.8
MID-FC-CSN-K4 (ours)	62.1

Results: Comparison with other methods

Method	Part IoU
ResGCN-28 (Li et al. 2023)	45.1
CloserLook3D (Liu et al. 2020)	53.8
MinkResUNet (Choy et al. 2019)	46.8
MinkHRNetCSN-K1 (ours)	49.9
MID-FC (Wang et al. 2021)	60.8
MID-FC-CSN-K4 (ours)	62.1

+3.1%

Results: Comparison with other methods

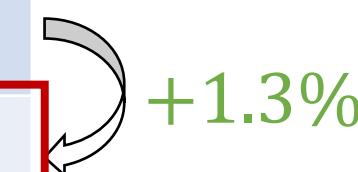
Method	Part IoU
ResGCN-28 (Li et al. 2023)	45.1
CloserLook3D (Liu et al. 2020)	53.8
MinkResUNet (Choy et al. 2019)	46.8
MinkHRNetCSN-K1 (ours)	49.9
MID-FC (Wang et al. 2021)	60.8
MID-FC-CSN-K4 (ours)	62.1

+1.3%

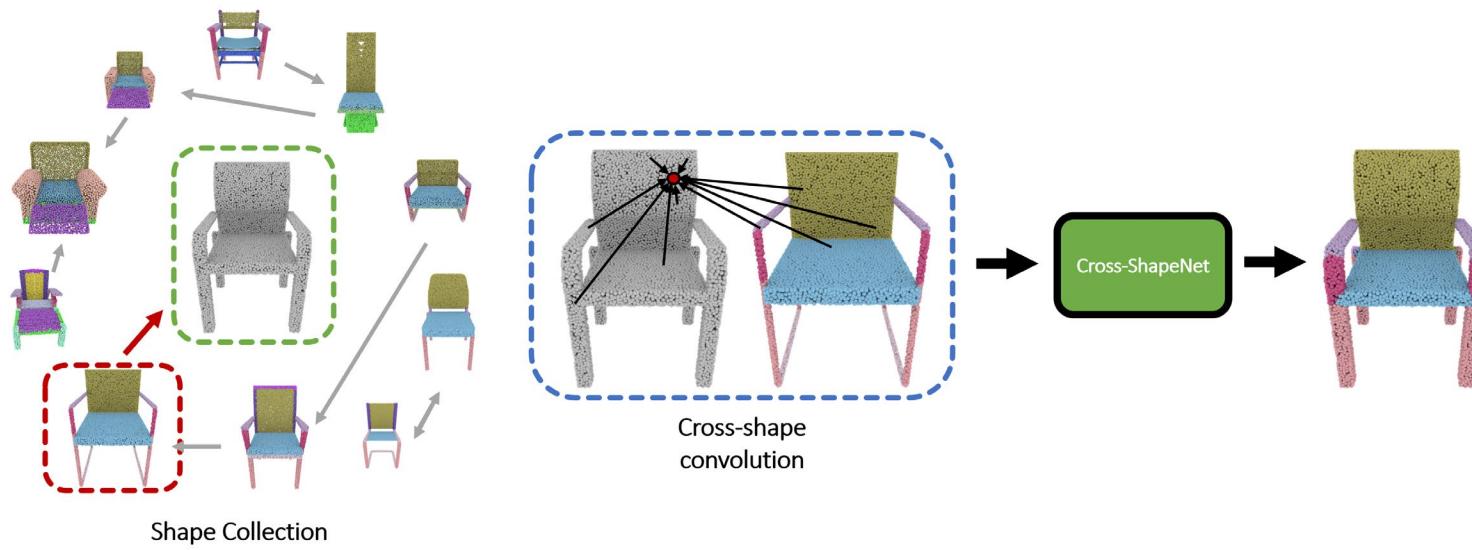
Results: Comparison with other methods

Method	Part IoU
ResGCN-28 (Li et al. 2023)	45.1
CloserLook3D (Liu et al. 2020)	53.8
MinkResUNet (Choy et al. 2019)	46.8
MinkHRNetCSN-K1 (ours)	49.9
MID-FC (Wang et al. 2021)	60.8
MID-FC-CSN-K4 (ours)	62.1

SOTA performance on the PartNet dataset

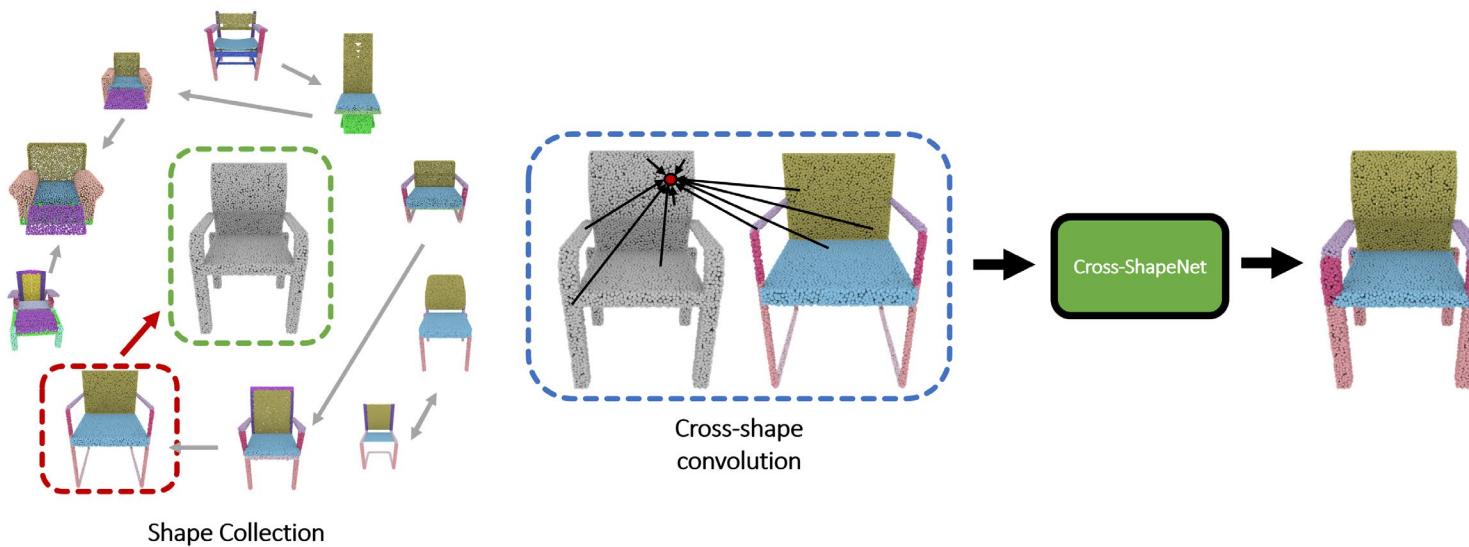


Summary



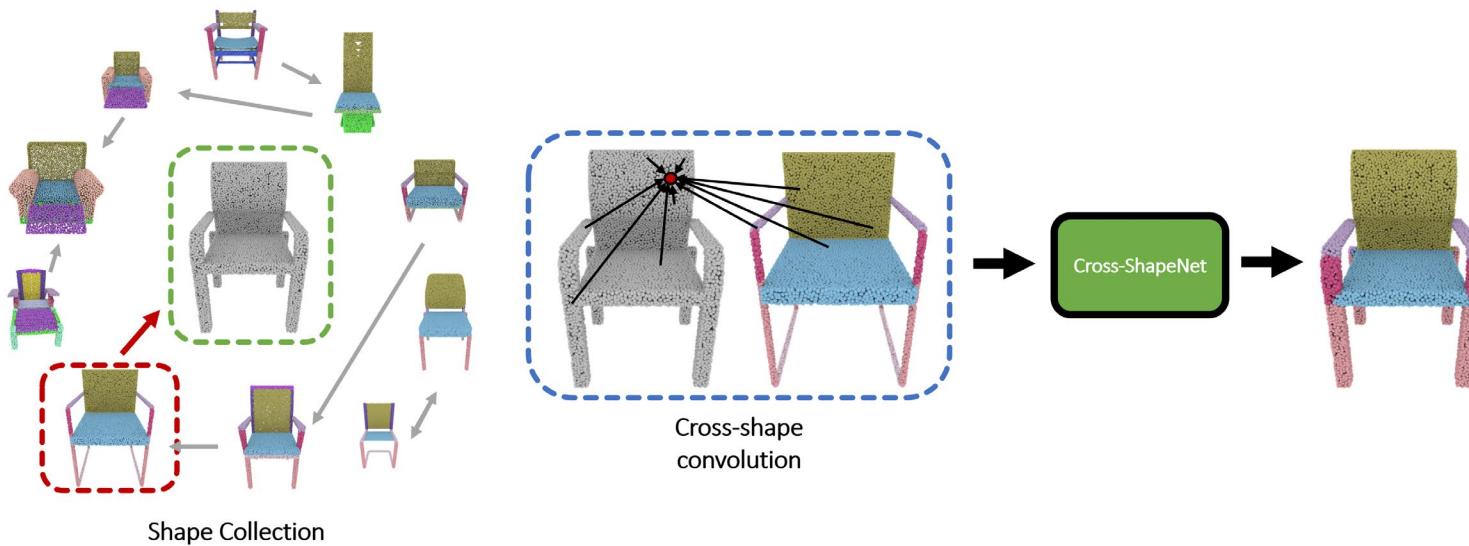
- Enable long range point feature interactions **across shapes**

Summary



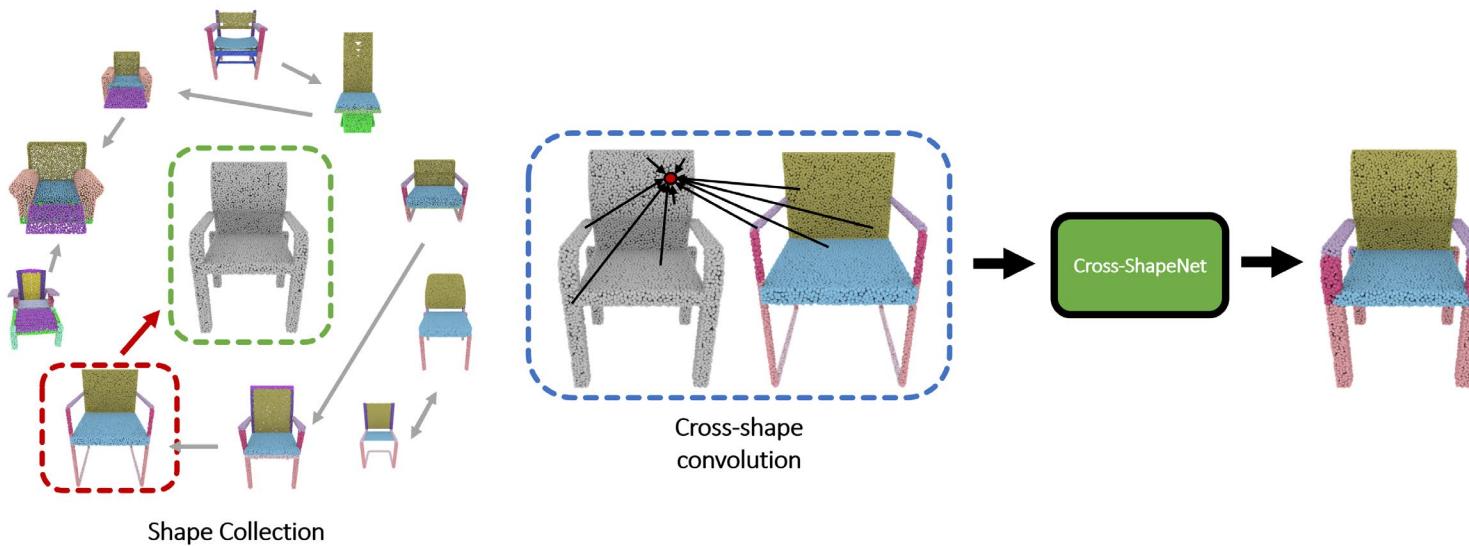
- Enable long range point feature interactions **across shapes**
- Introduce a **novel cross-shape attention** mechanism

Summary



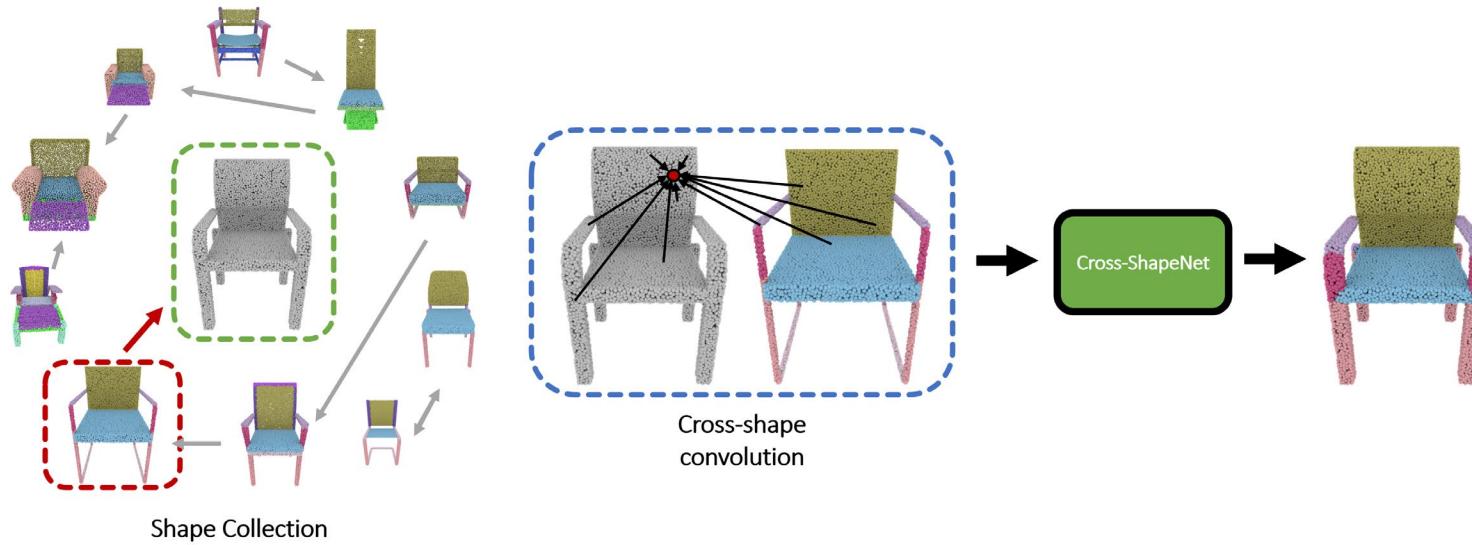
- Enable long range point feature interactions **across shapes**
- Introduce a **novel cross-shape attention** mechanism
- Retrieve **compatible shapes** for cross-shape attention

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

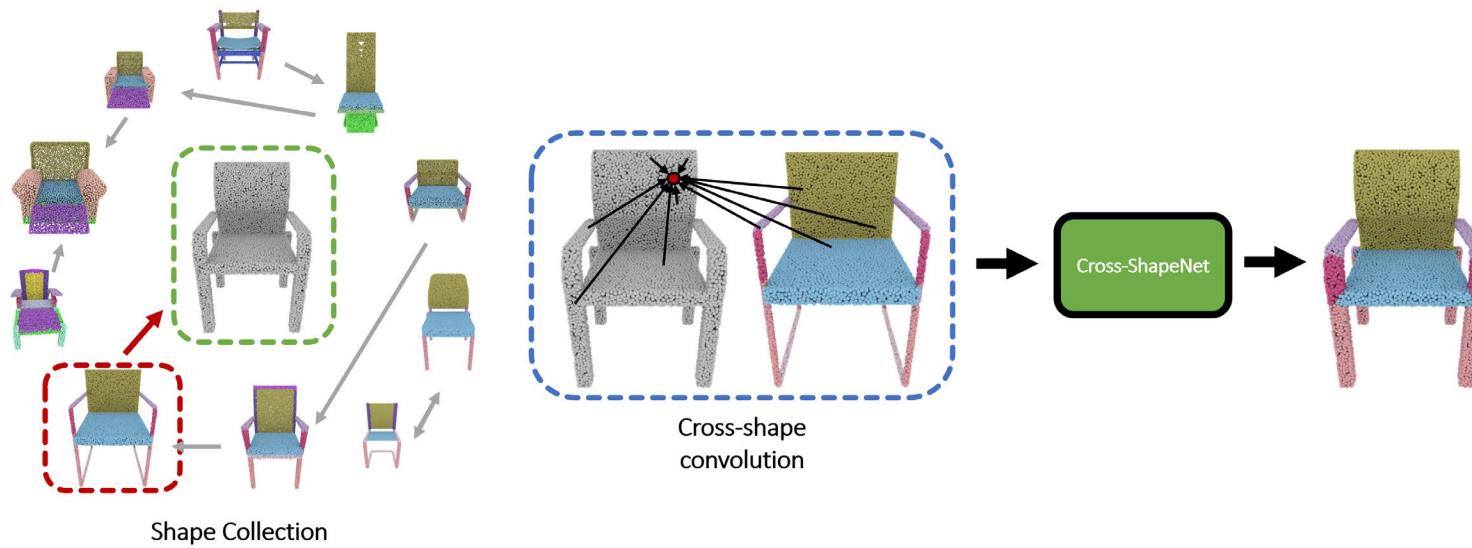
Summary



Limitations:

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

Summary



Limitations:

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

Thank you!

Acknowledgements:



Horizon2020
European Union Funding
for Research & Innovation



DEPUTY MINISTRY OF
RESEARCH, INNOVATION
AND DIGITAL POLICY
REPUBLIC OF CYPRUS



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

