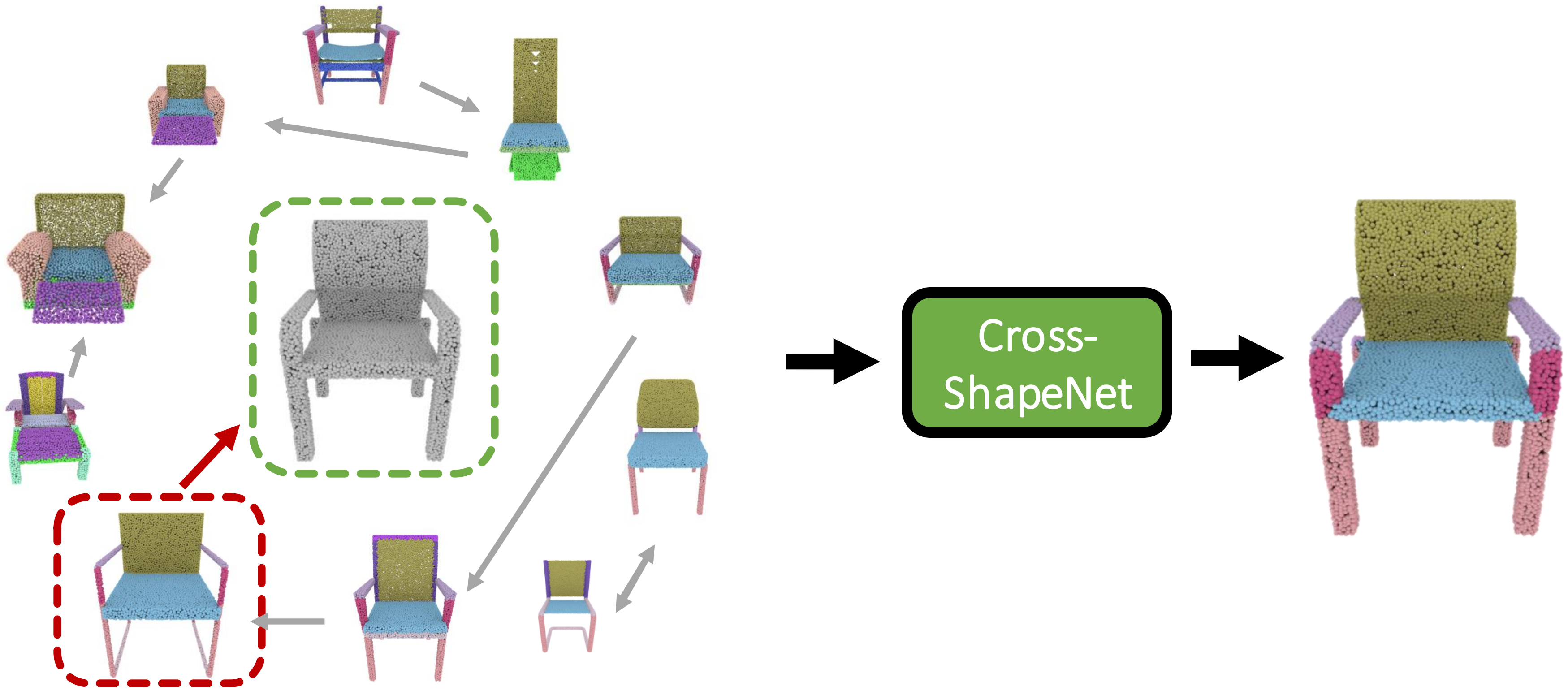
Cross-Shape Attention for Part Segmentation of 3D Point Clouds



Marios Loizou†[1] **Siddhant Garg †[2]** **Dmitry Petrov †[2]**

Melinos Averkiou[1] **Evangelos Kalogerakis[2]**

[1]University of Cyprus / CYENS CoE *[2]University of Massachusetts Amherst*

† Equal Contribution

Goal: learn more coordinated feature representations

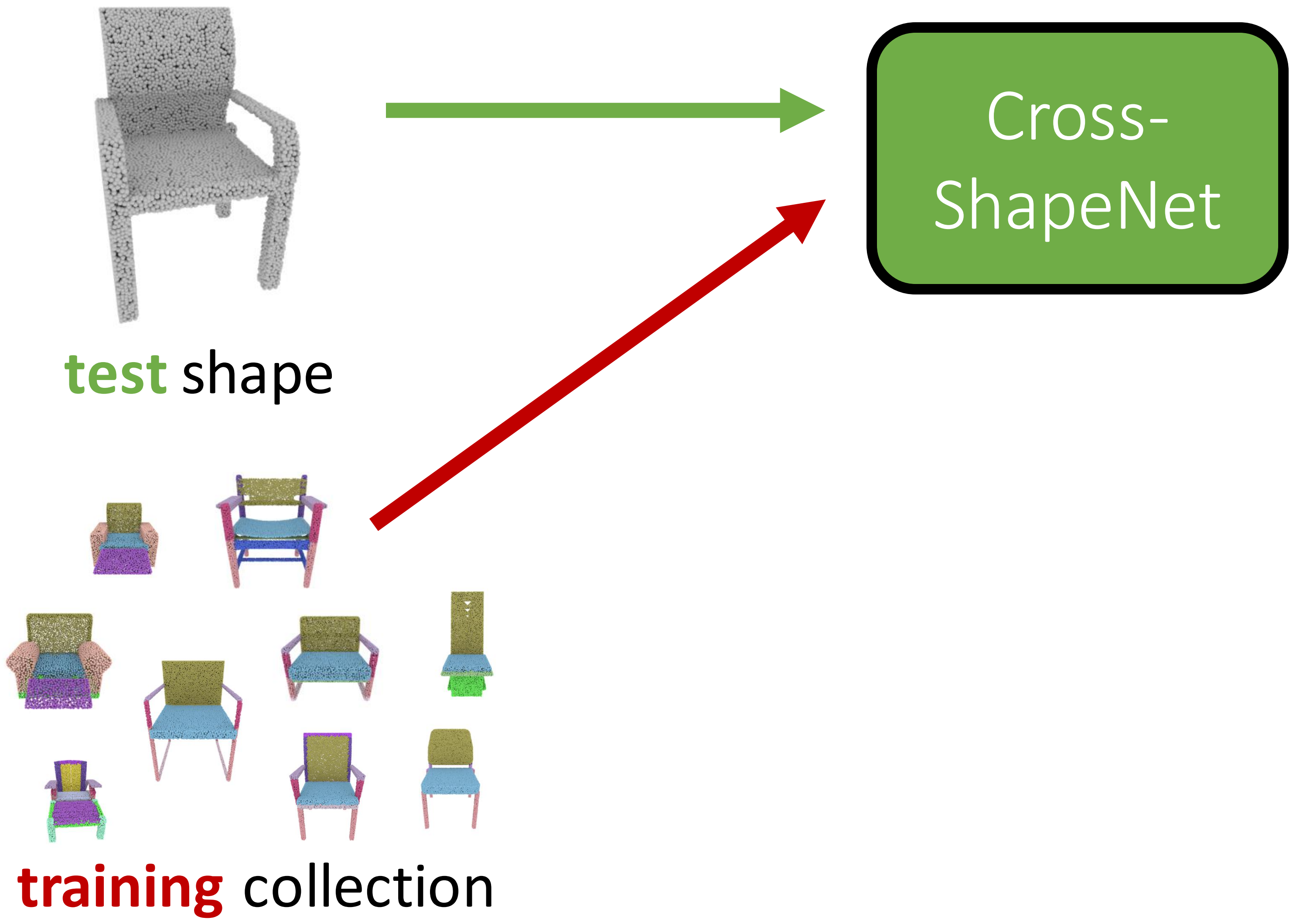
test shape

Goal: learn more coordinated feature representations

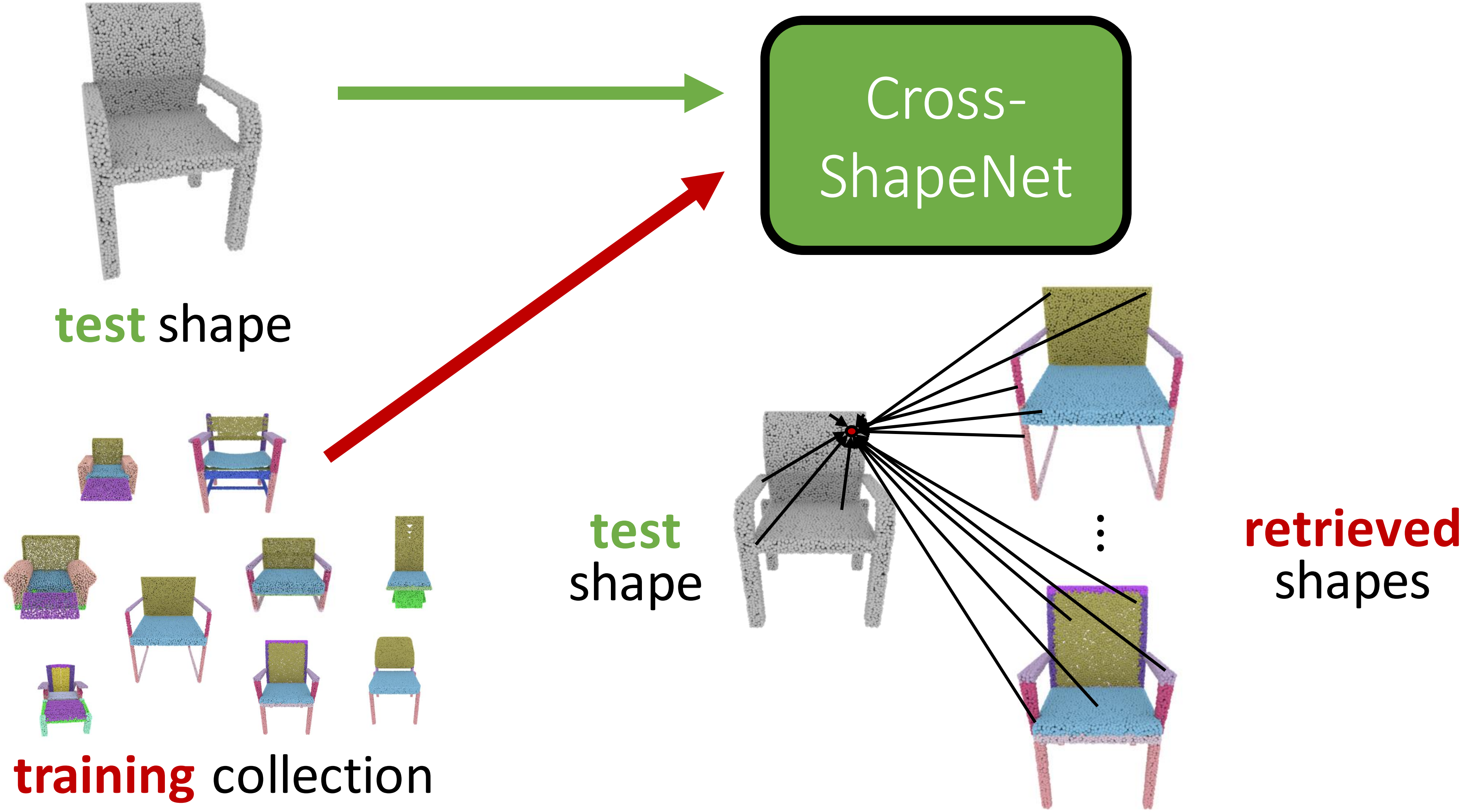
test shape

training collection

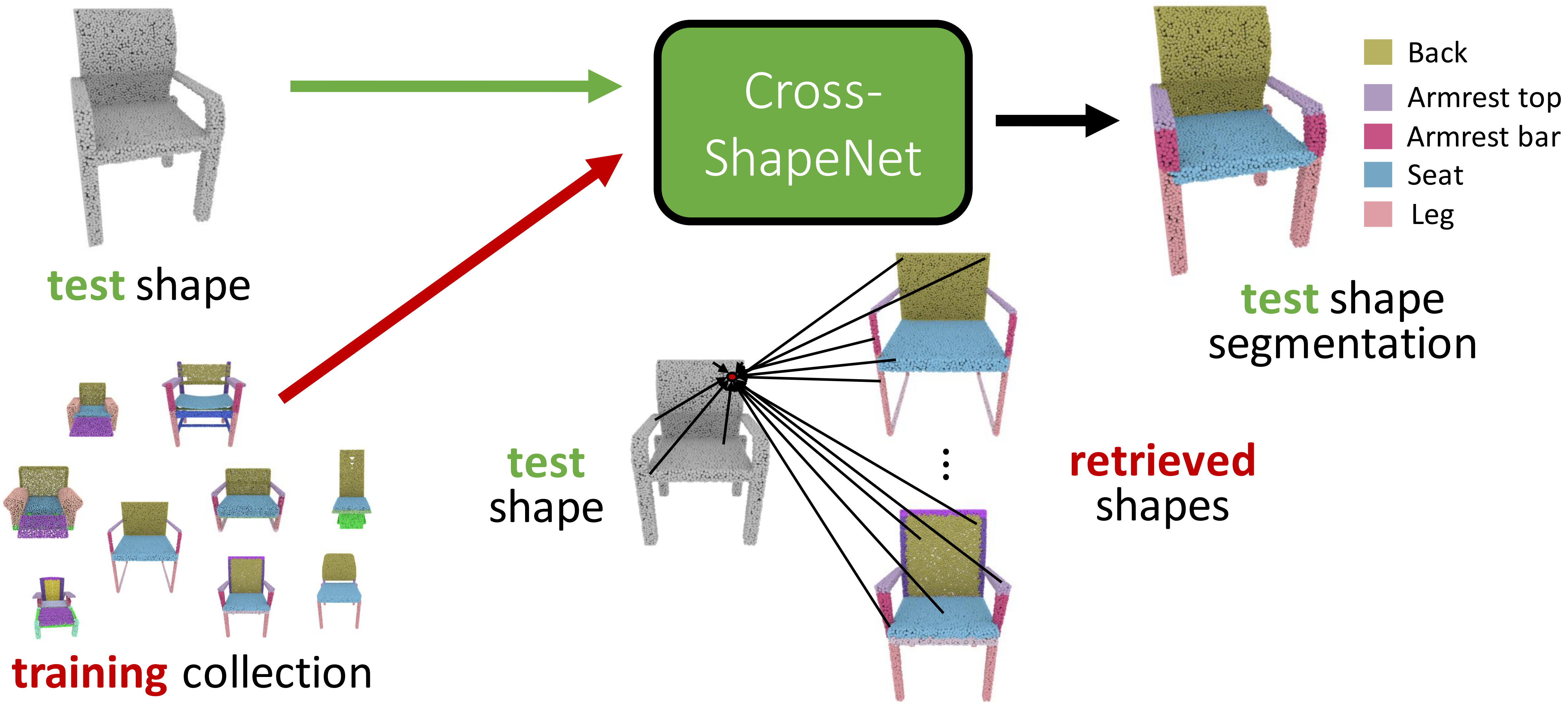
Goal: learn more coordinated feature representations



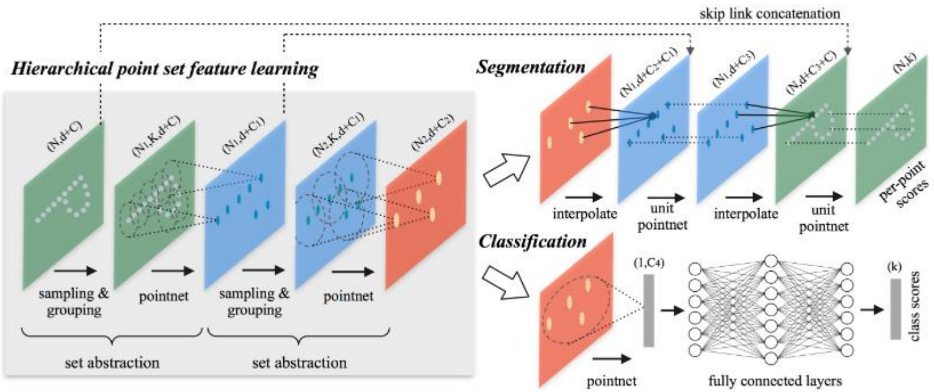
Goal: learn more coordinated feature representations



Goal: learn more coordinated feature representations

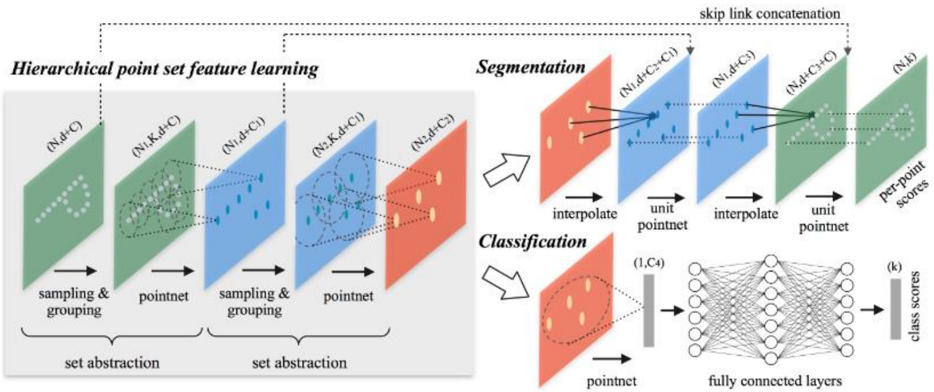


Prior work: Point-based networks

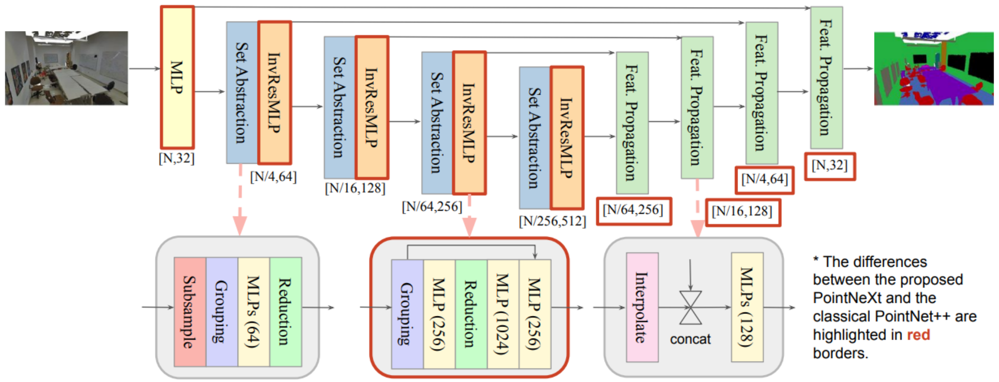


PointNet++ [Qi et al. 2017]

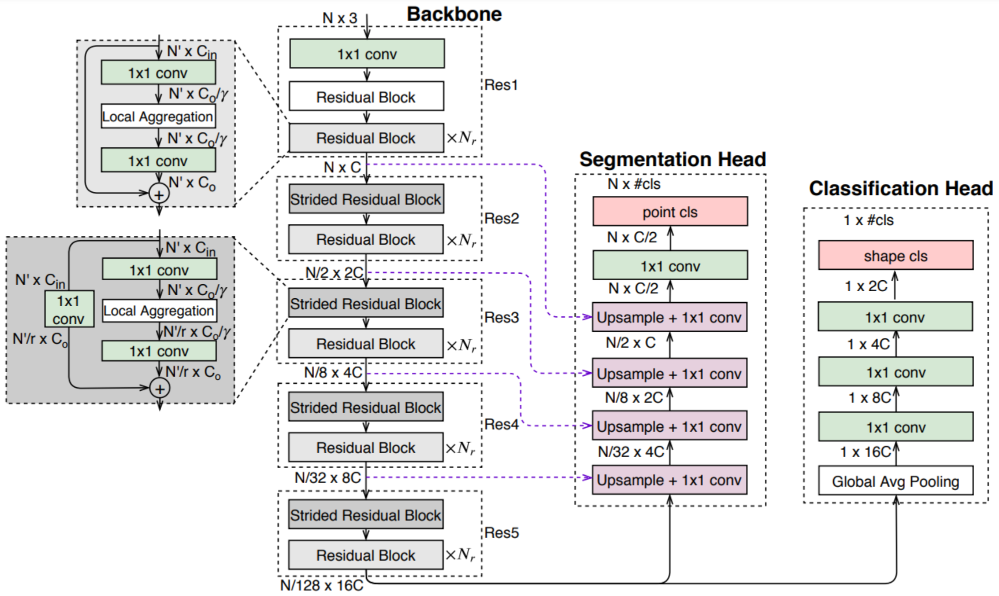
Prior work: Point-based networks



PointNet++ [Qi et al. 2017]

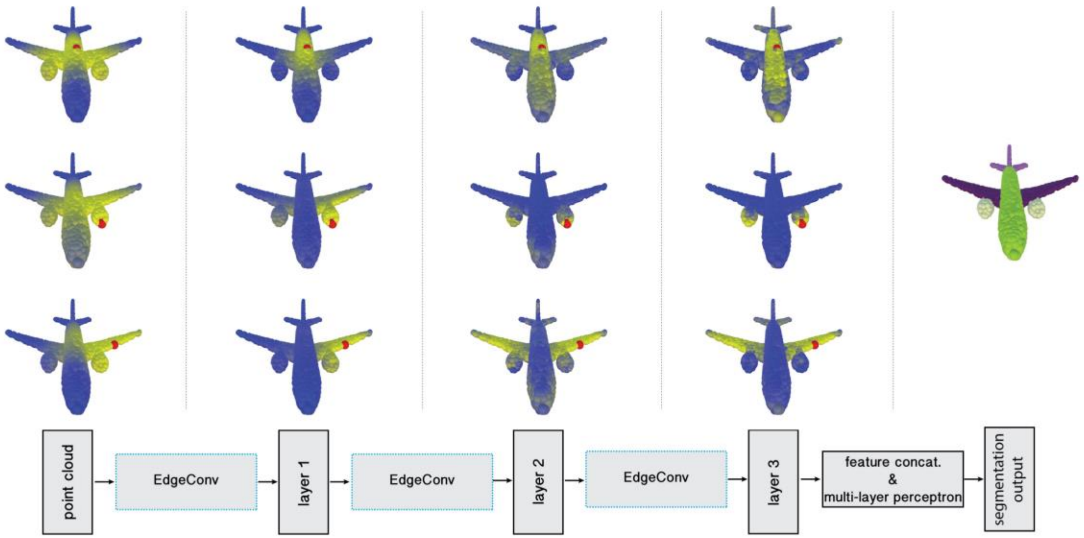


PointNeXt [Qian et al. 2022]

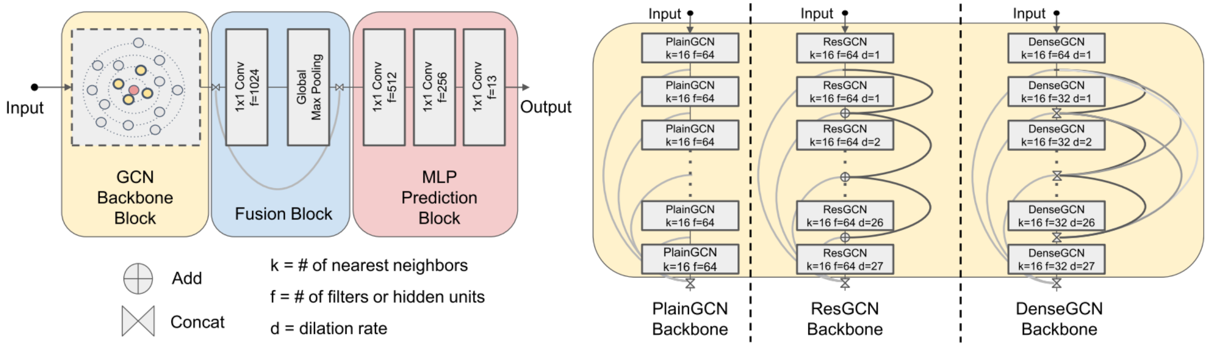


CloserLook3D [Liu et al. 2020]

Prior work: GCNs for non-Eucledian data

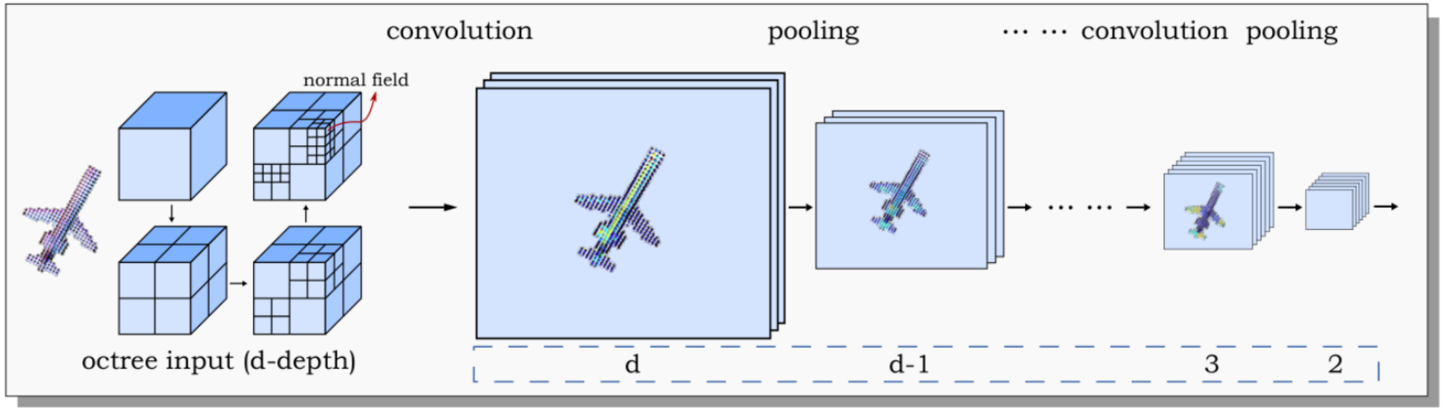


DGCNN [Wang et al. 2019]

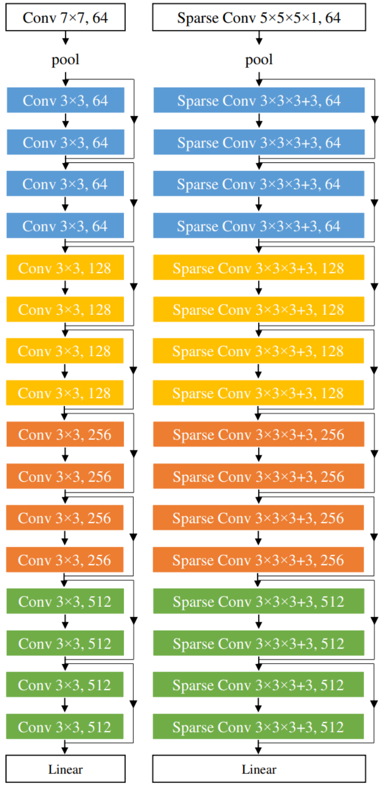


DeepGCNs [Li et al. 2023]

Prior work: Volumetric networks

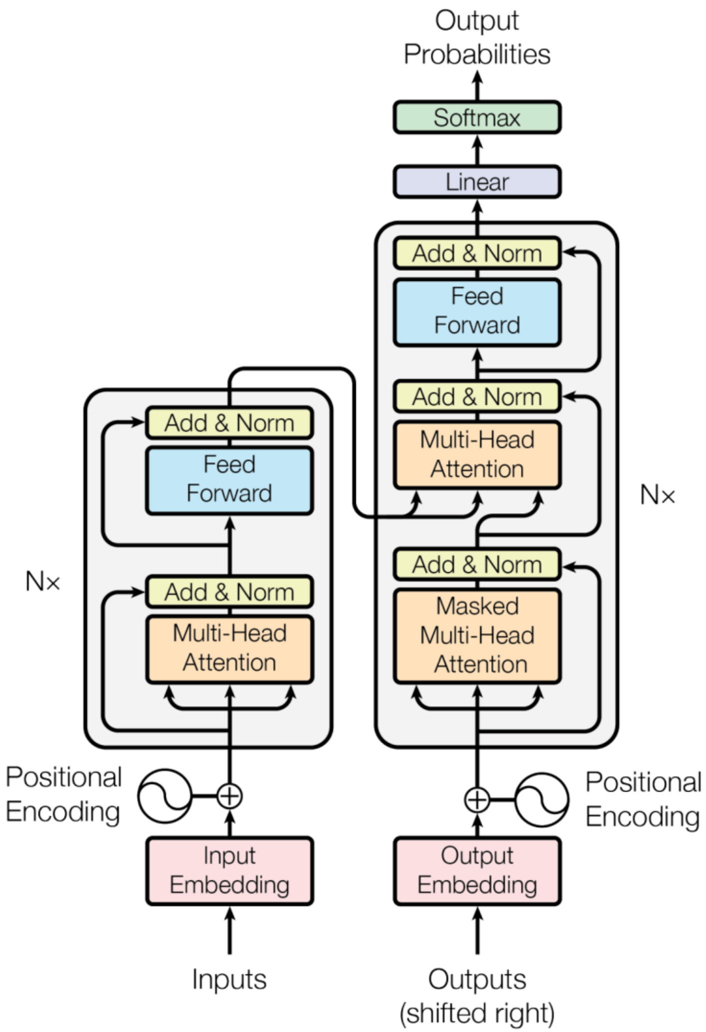


O-CNN [Wang et al. 2017]

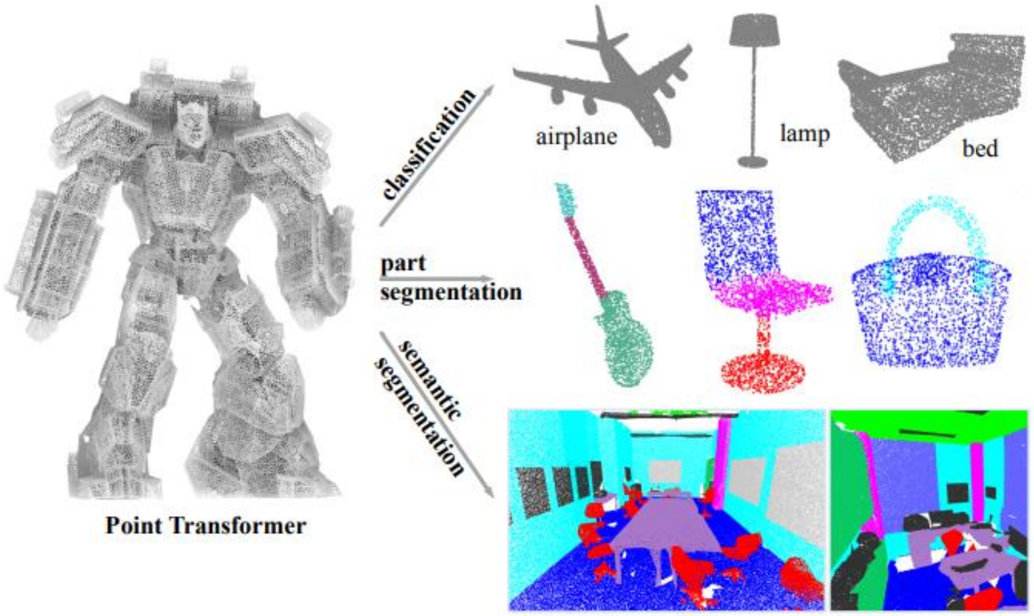


MinkowskiNet [Choy et al. 2019]

Prior work: Attention is All You Need



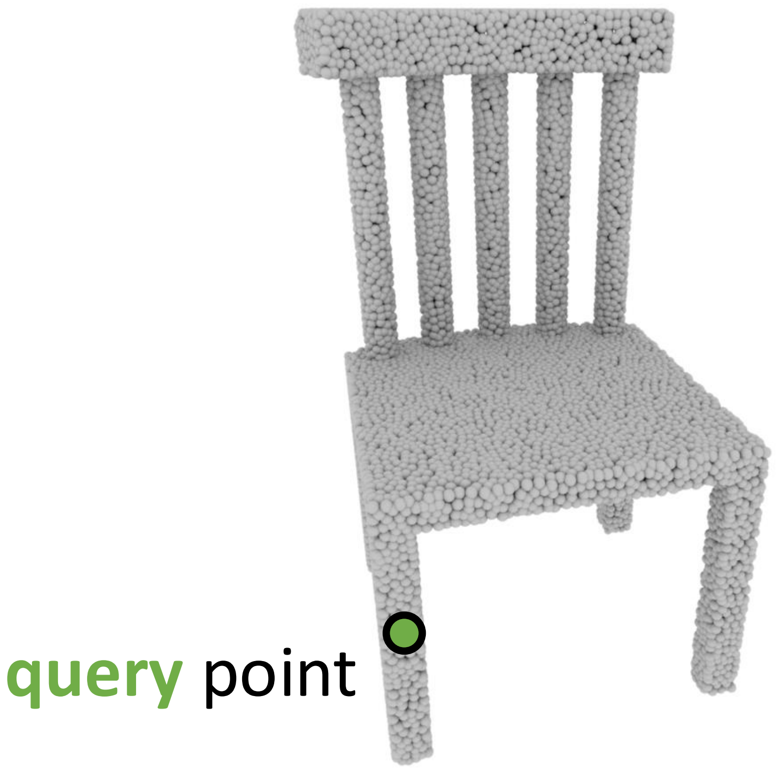
Transformer [Vaswani et al. 2017]



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

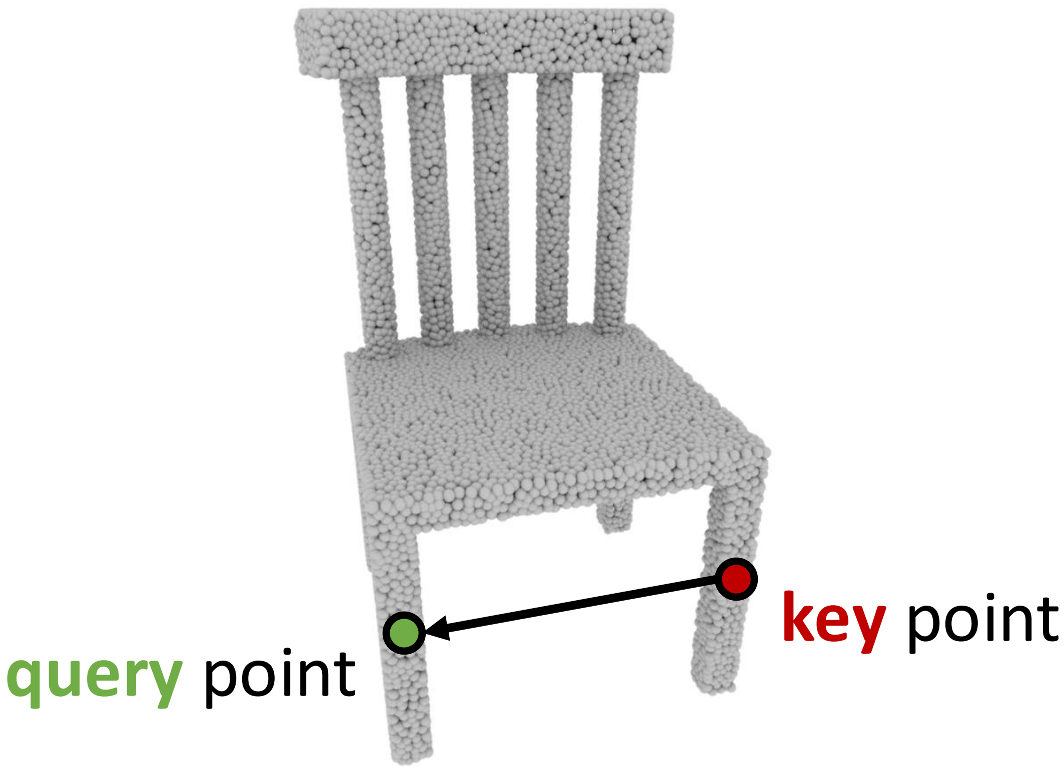
Why use **attention** for 3D representations?

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



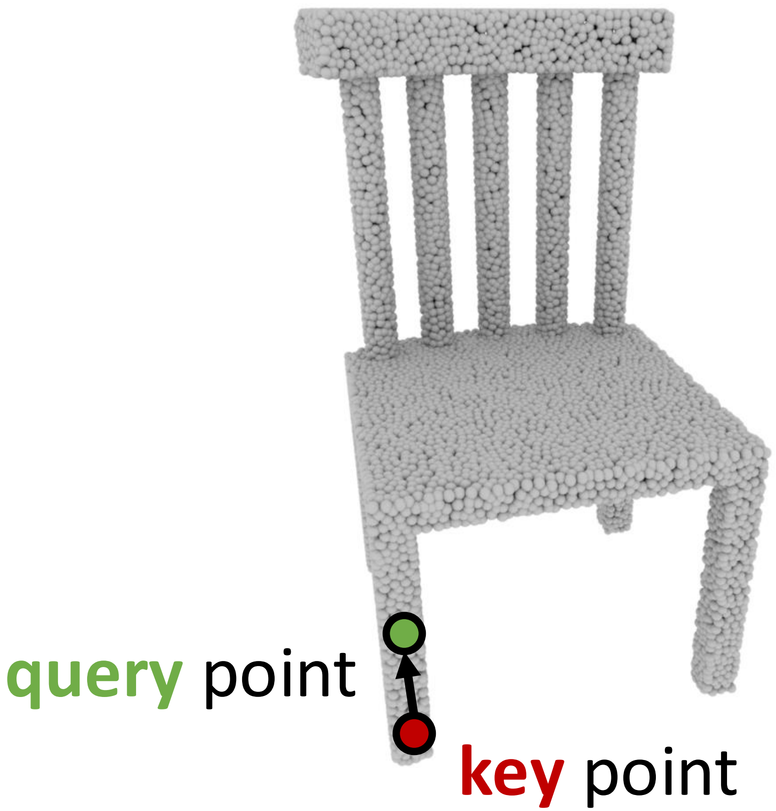
Why use **attention** for 3D representations?

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



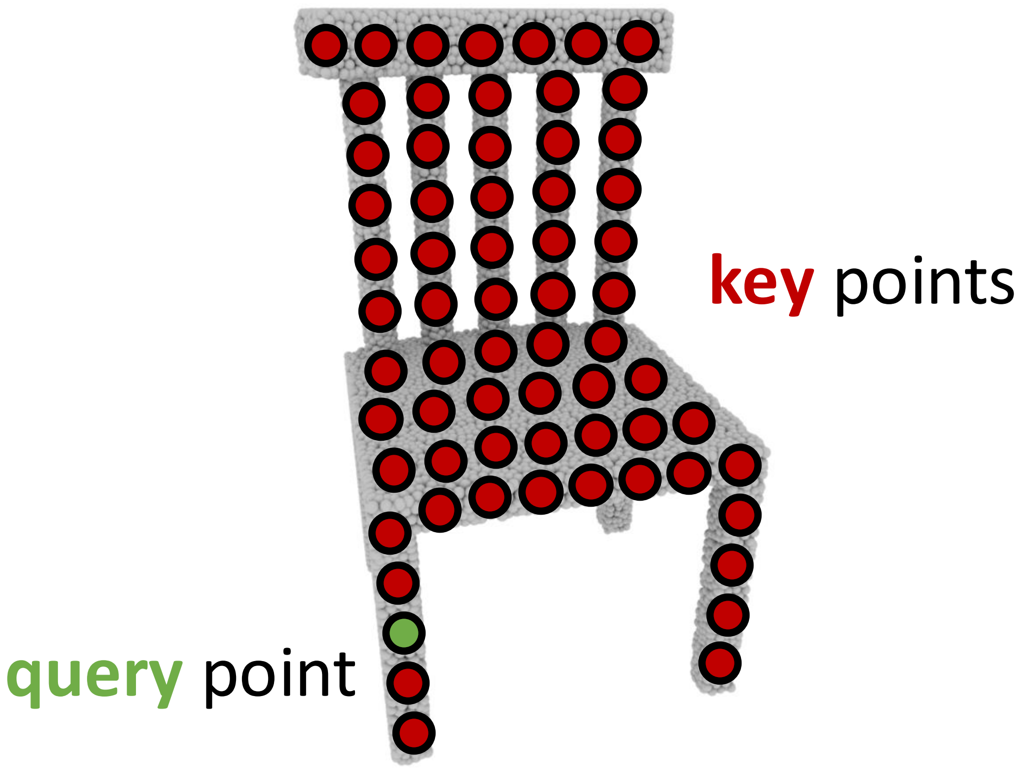
Why use **attention** for 3D representations?

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



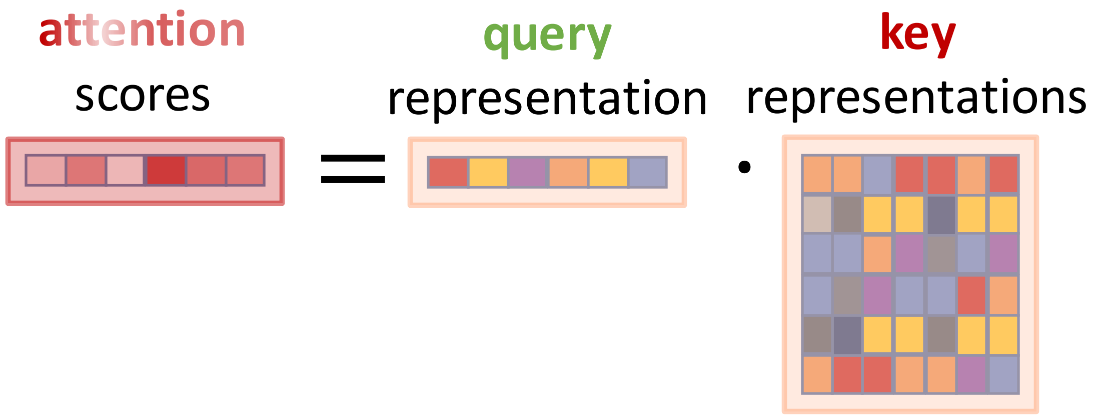
Why use **attention** for 3D representations?

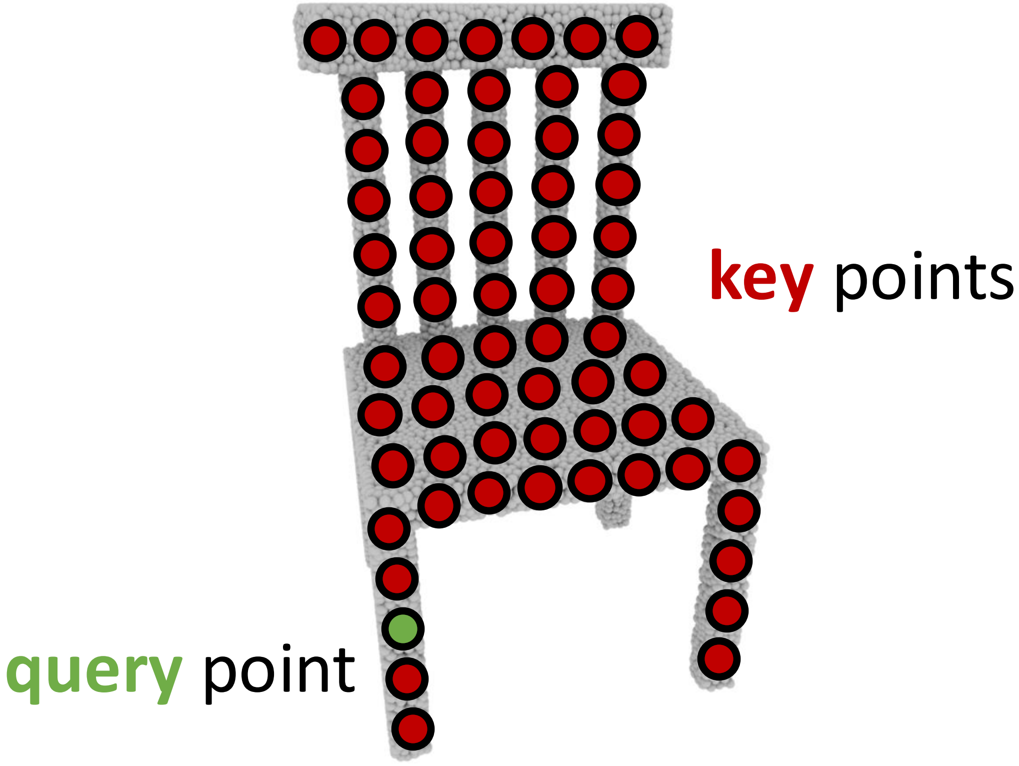
Encode points such that their features capture **relations** wrt the rest of the shape



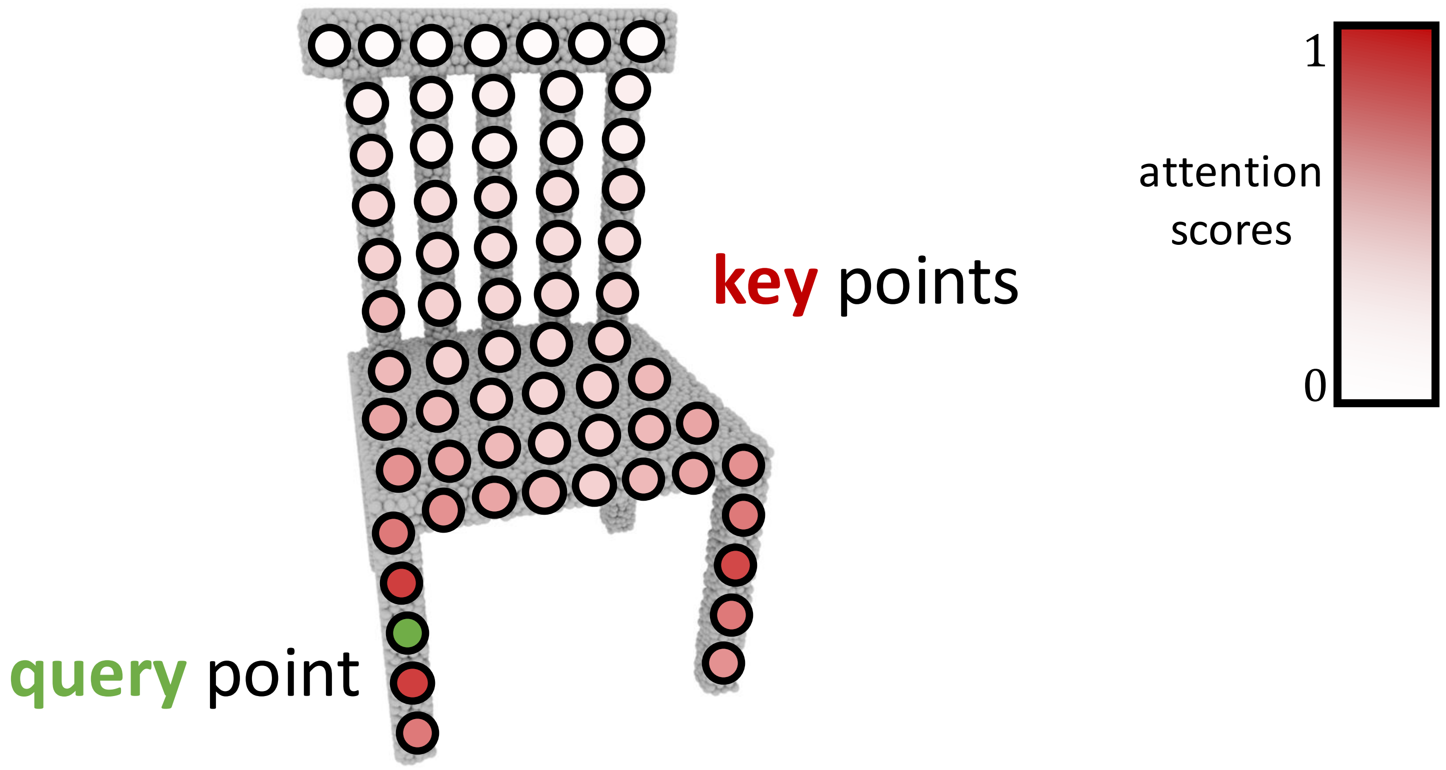
Why use **attention** for 3D representations?

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

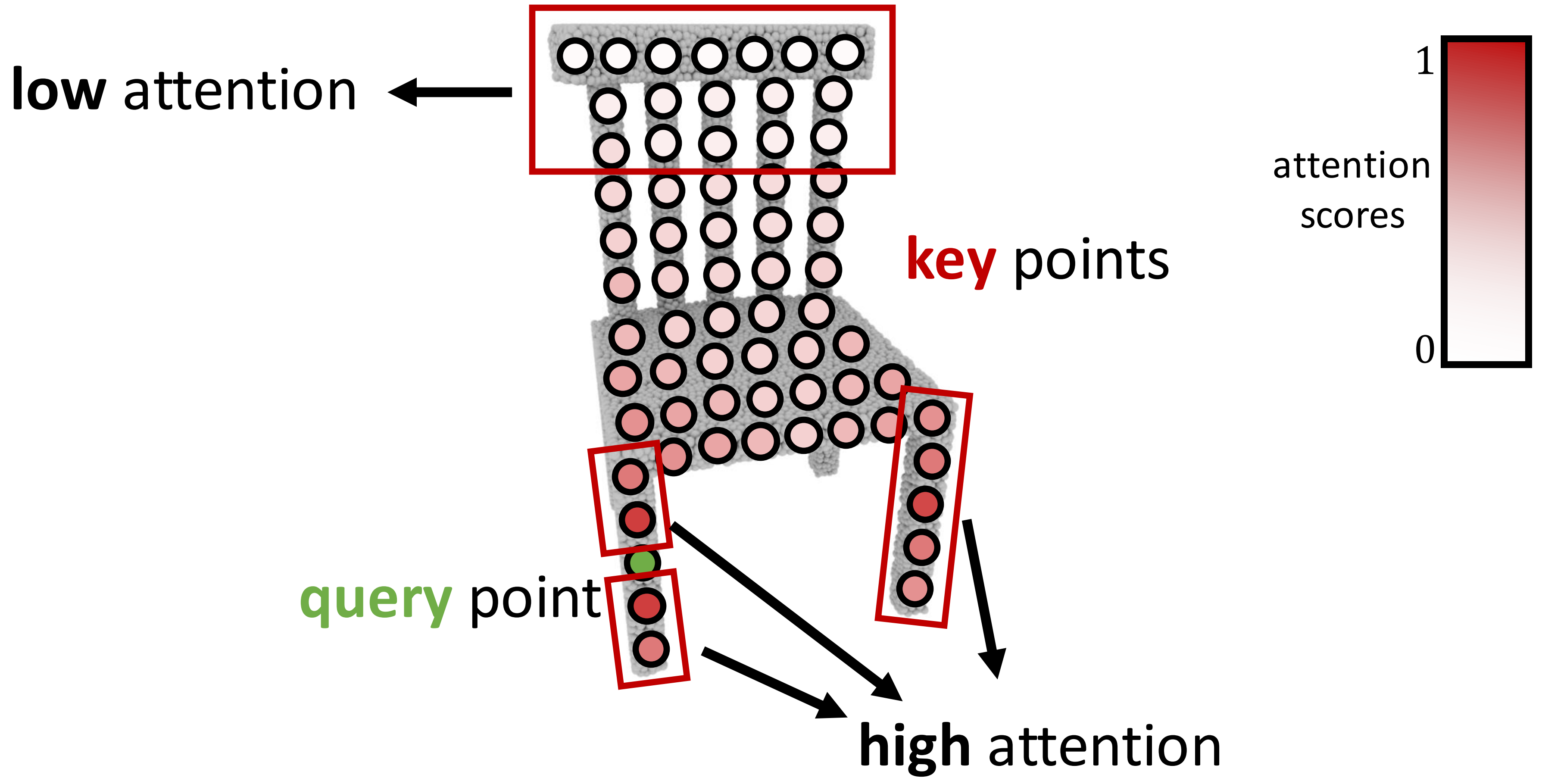




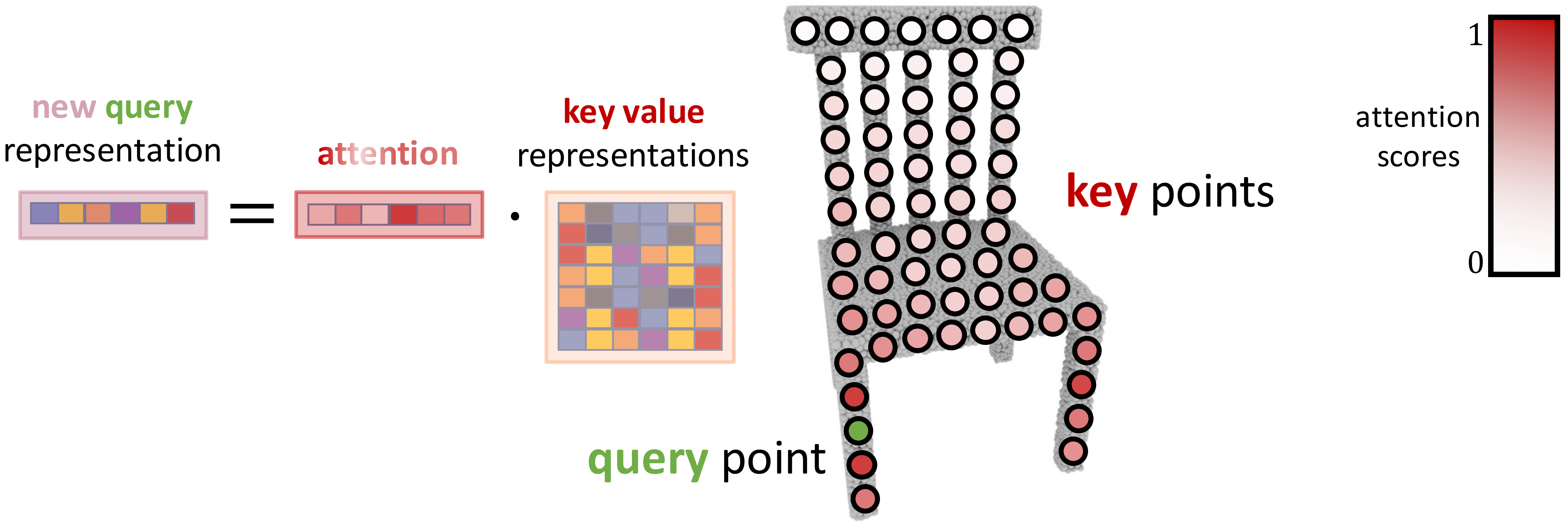
Why use **attention** for 3D representations?



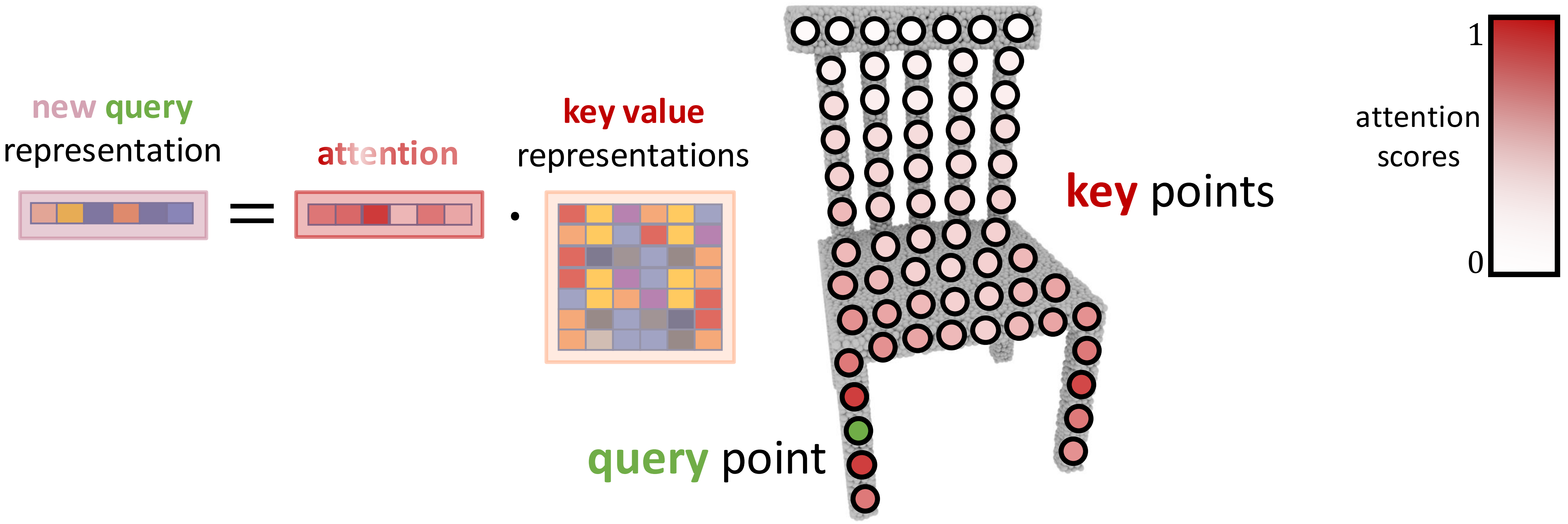
Why use **attention** for 3D representations?



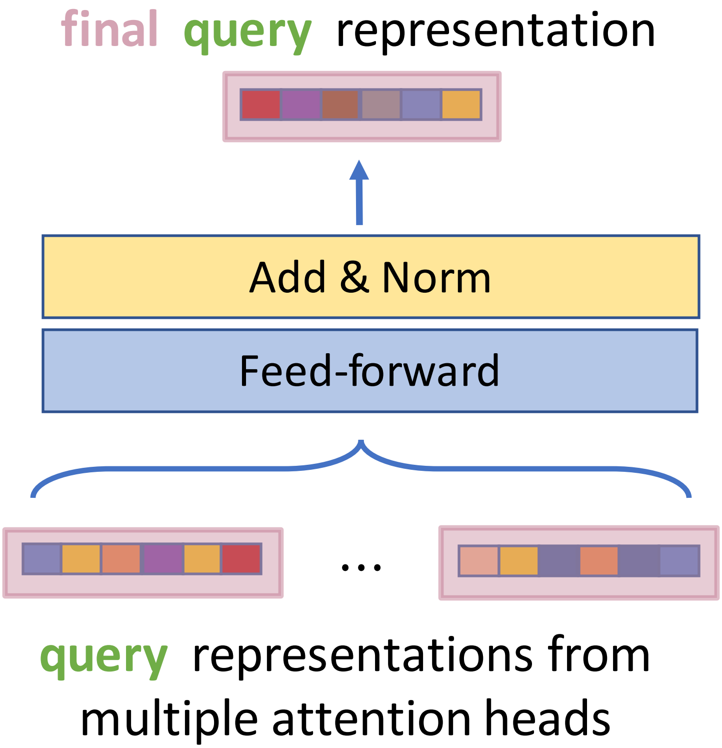
Why use **attention** for 3D representations?

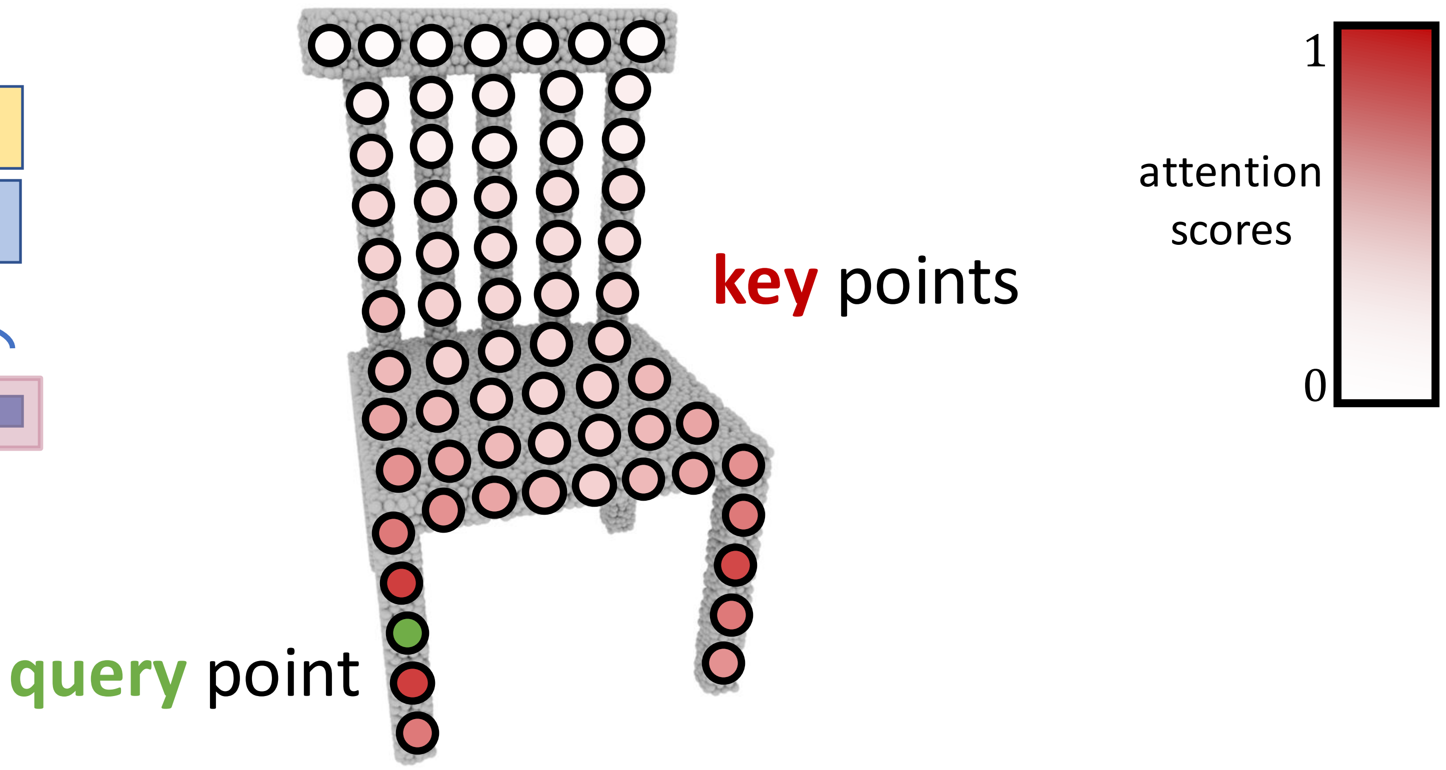


Why use **attention** for 3D representations?



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





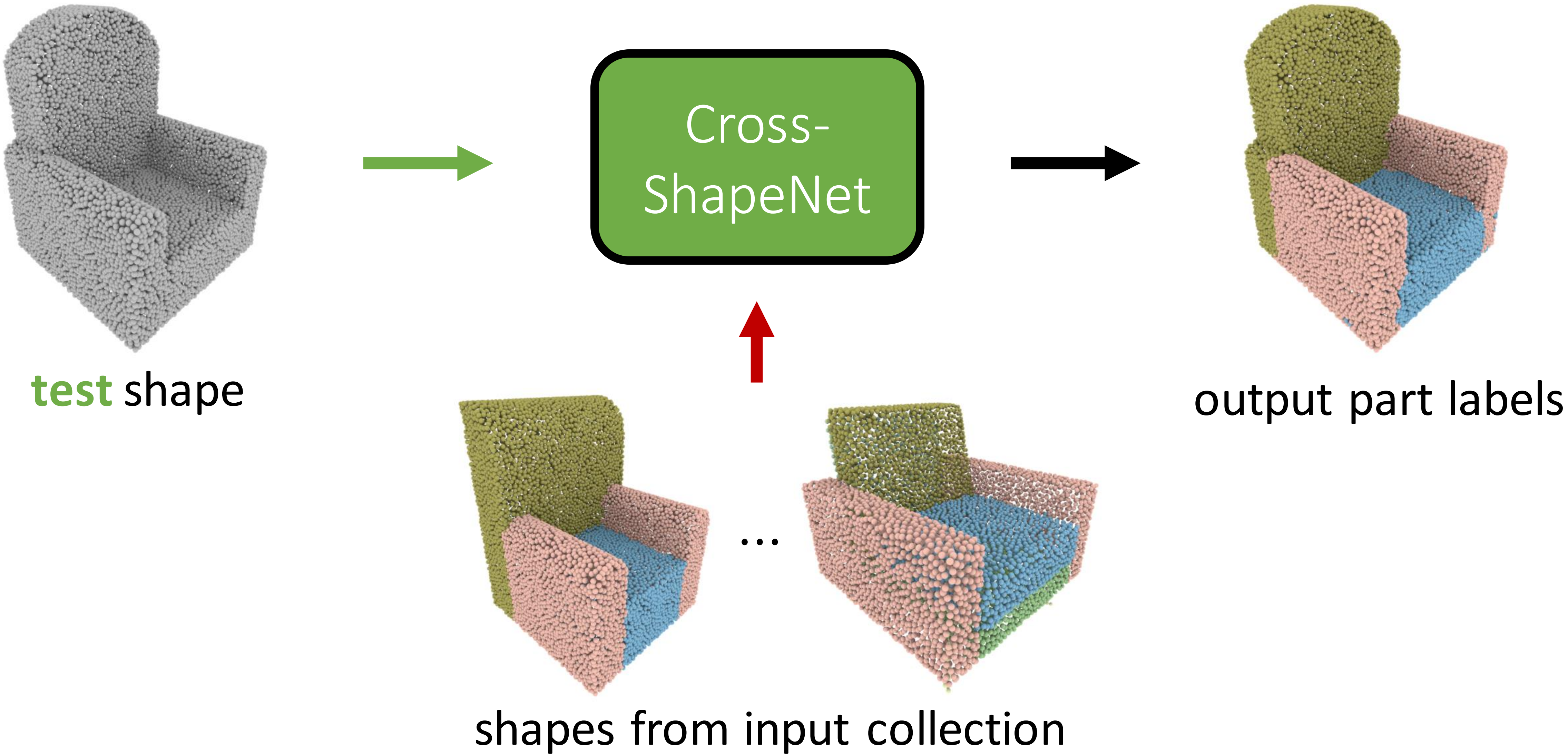
Motivation: Long-range interactions **across** shapes

test shape

output part labels

No interactions across shapes

Motivation: Long-range interactions across shapes

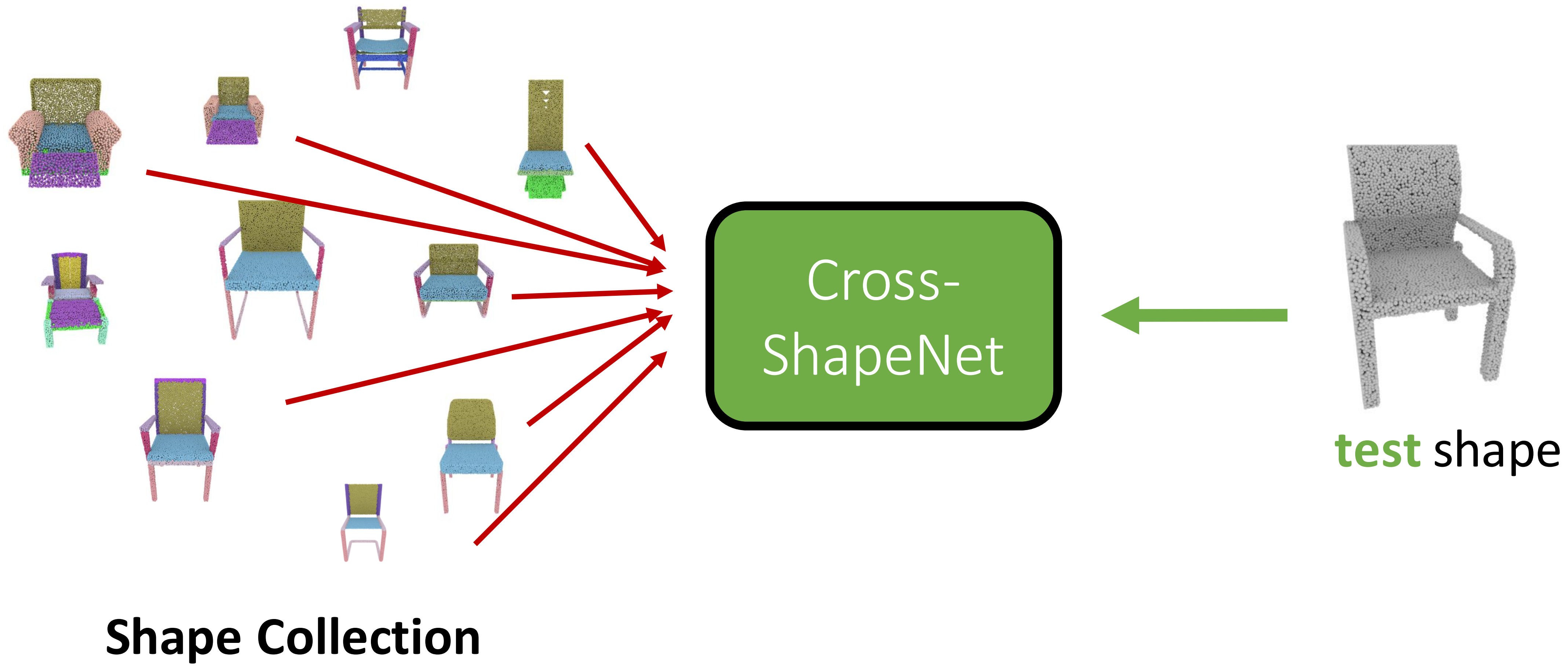


Key challenge: Retrieve compatible shapes

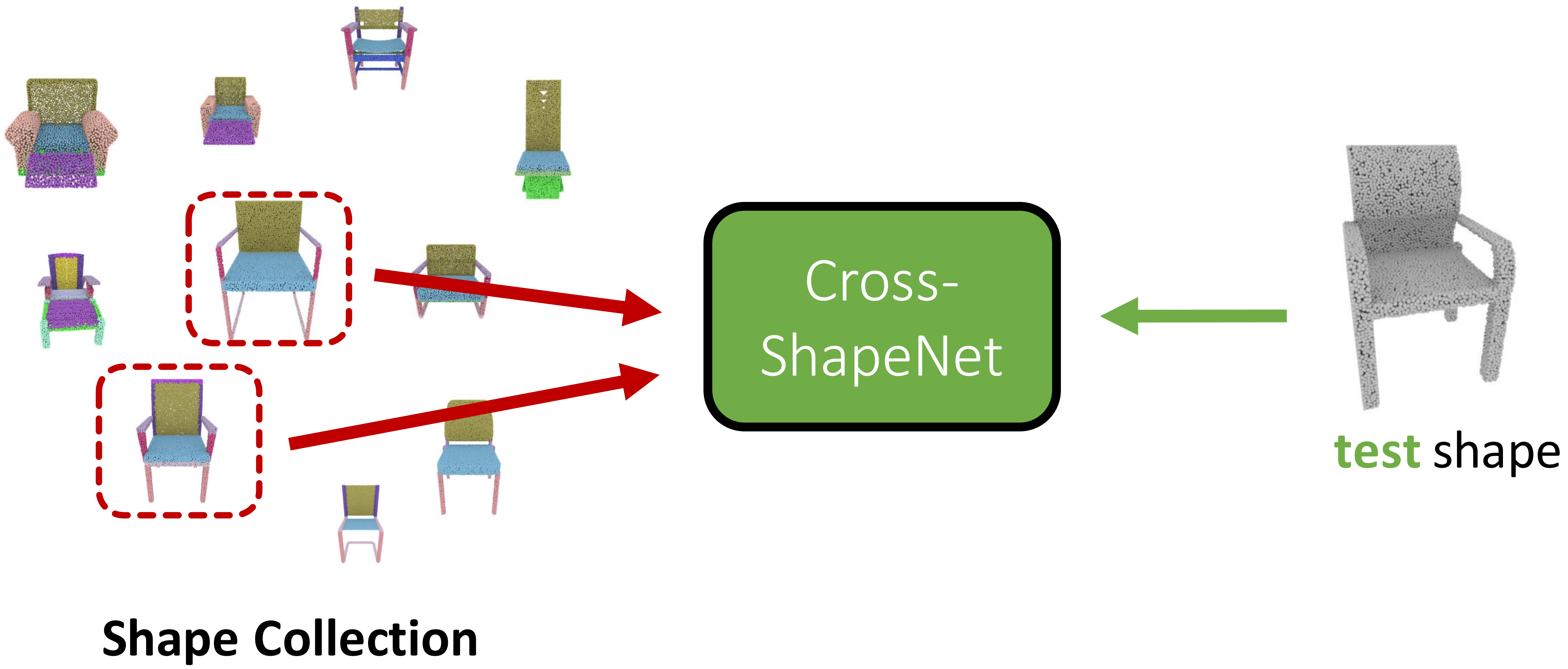
Shape Collection

test shape

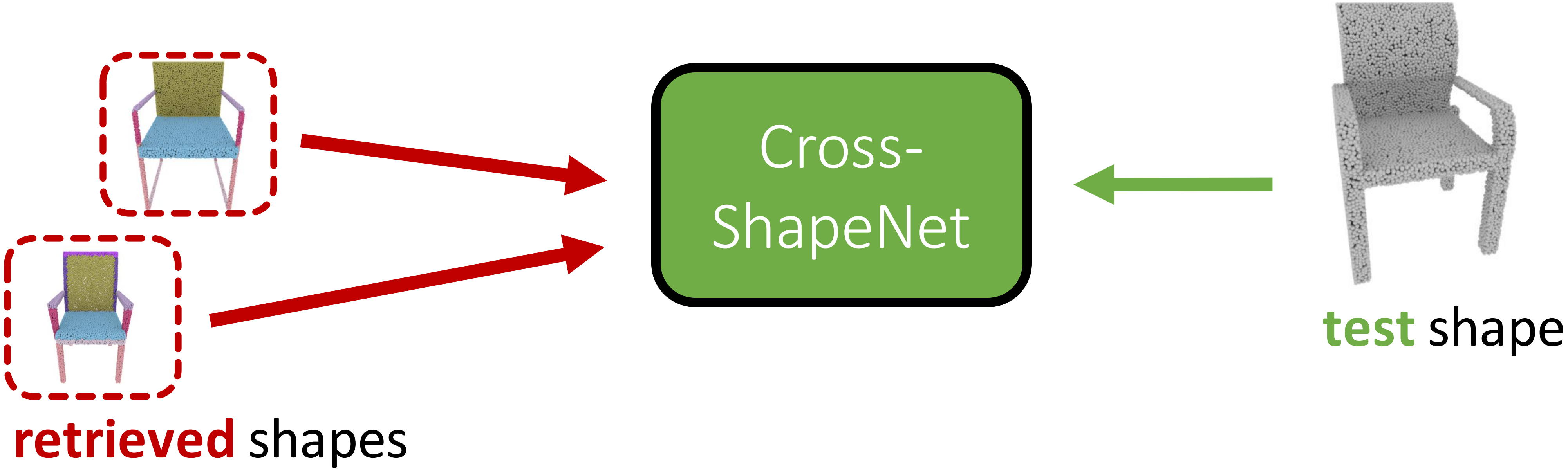
Key challenge: Retrieve compatible shapes



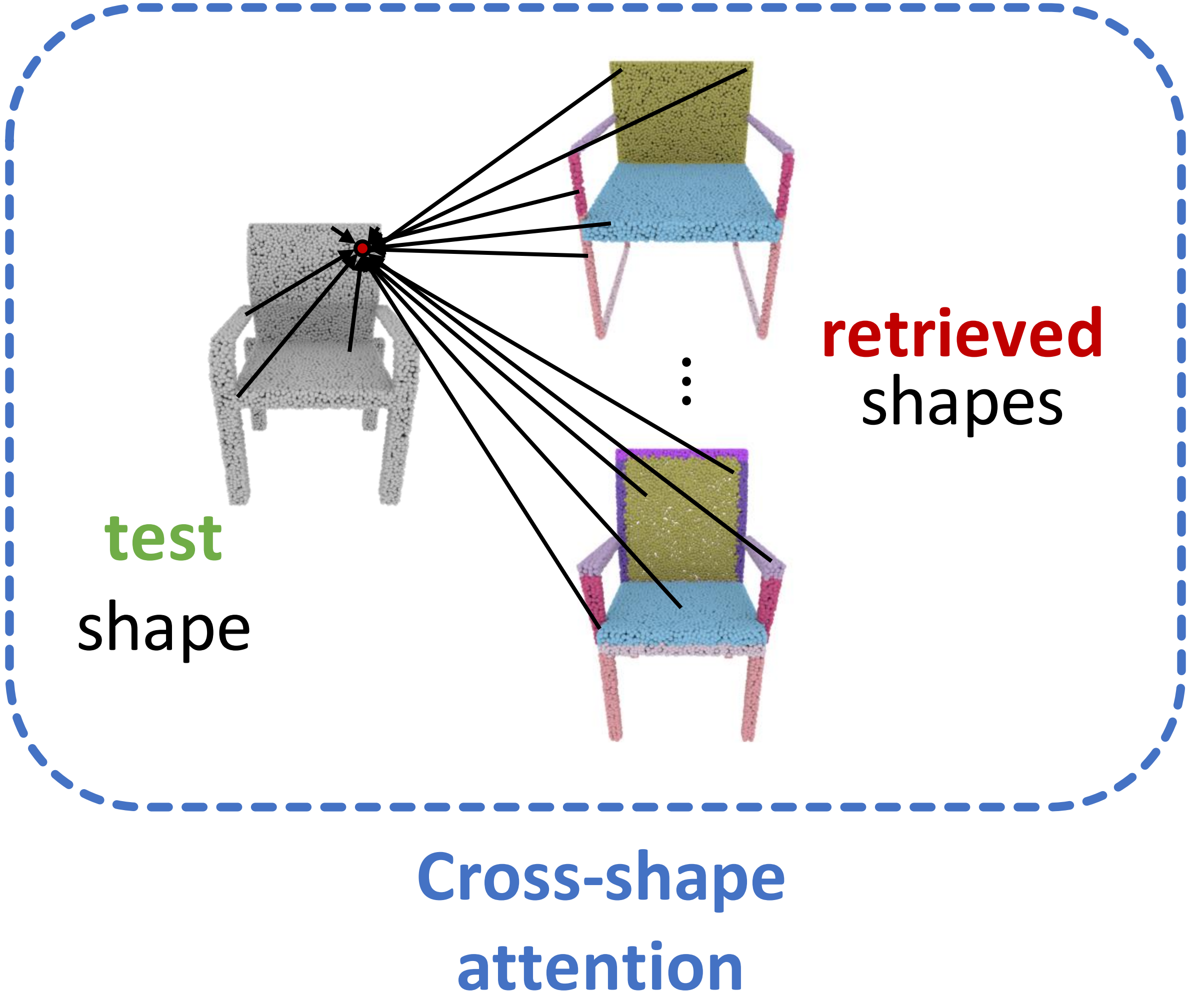
Key challenge: Retrieve compatible shapes



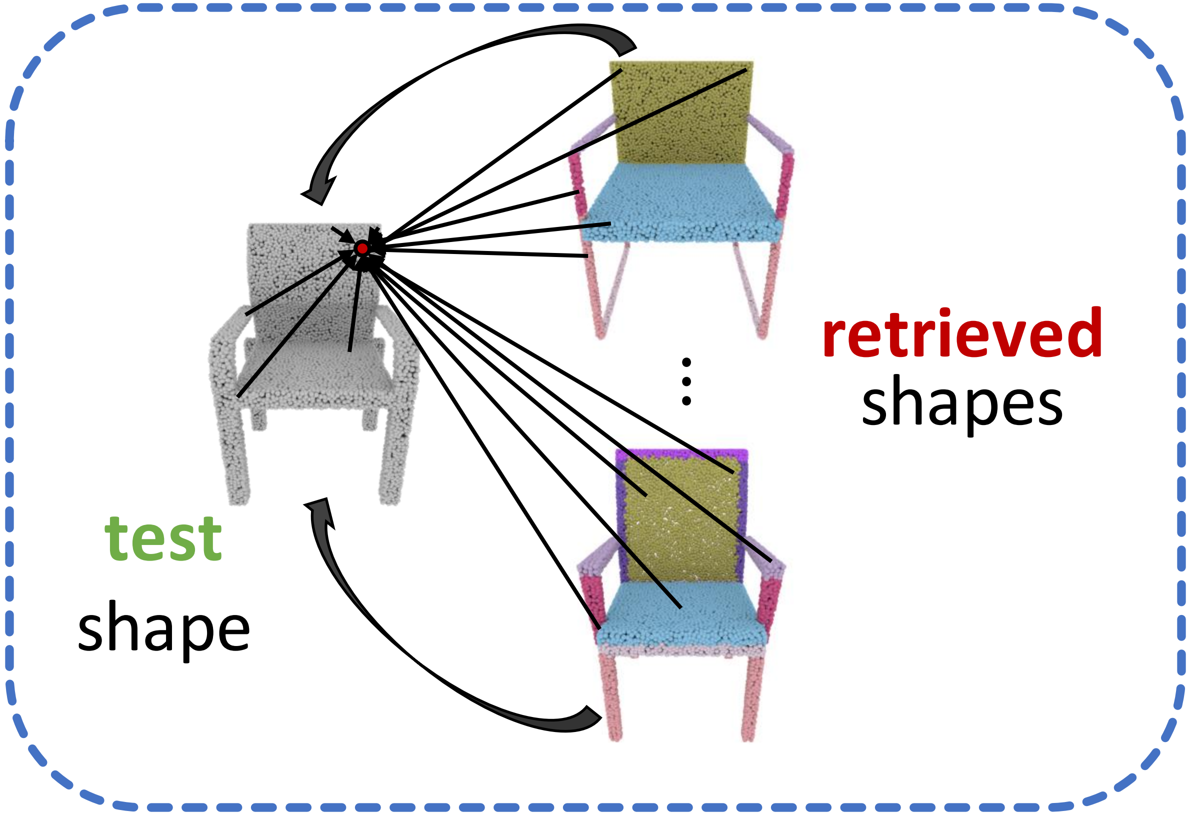
Key challenge: Combine multiple shapes



Key challenge: Combine multiple shapes

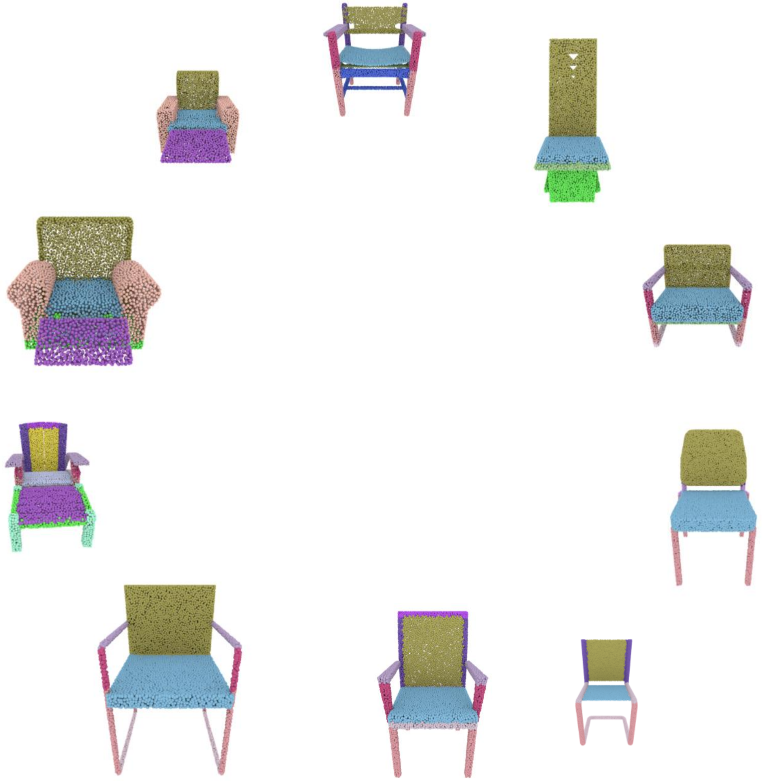


Key challenge: Combine multiple shapes



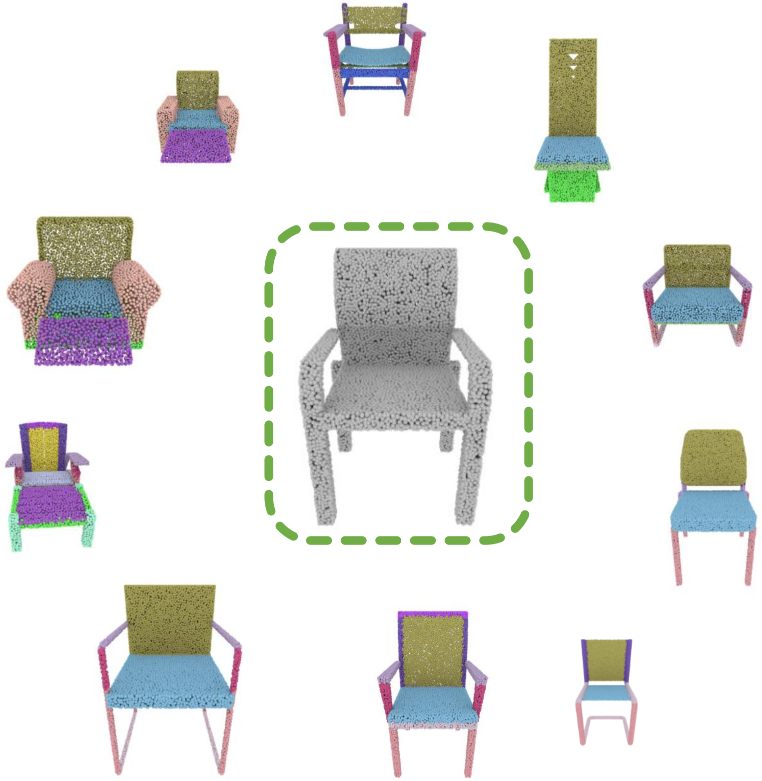
Cross-shape attention

Pipeline



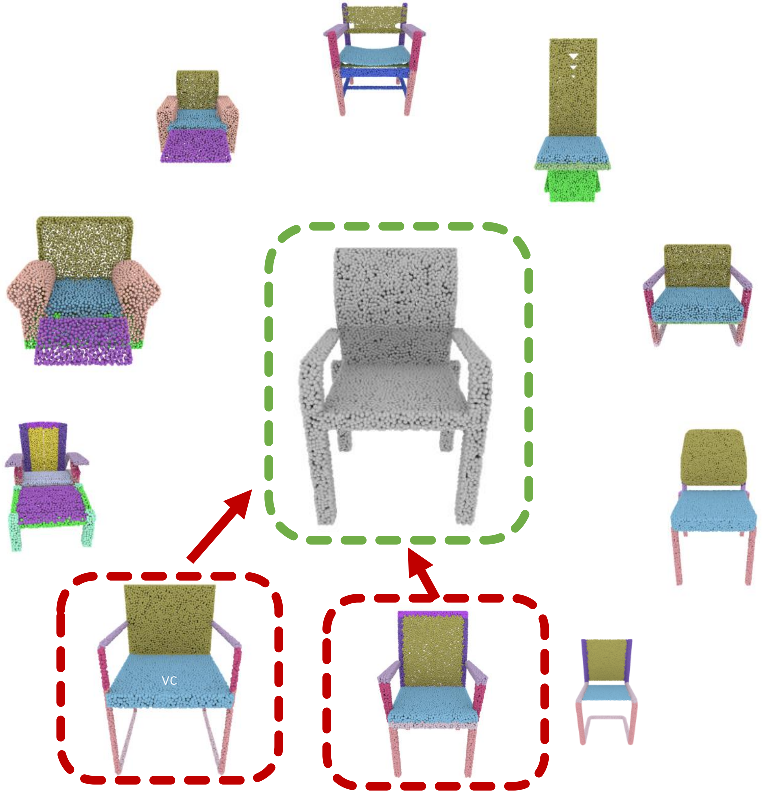
Shape Collection

Pipeline



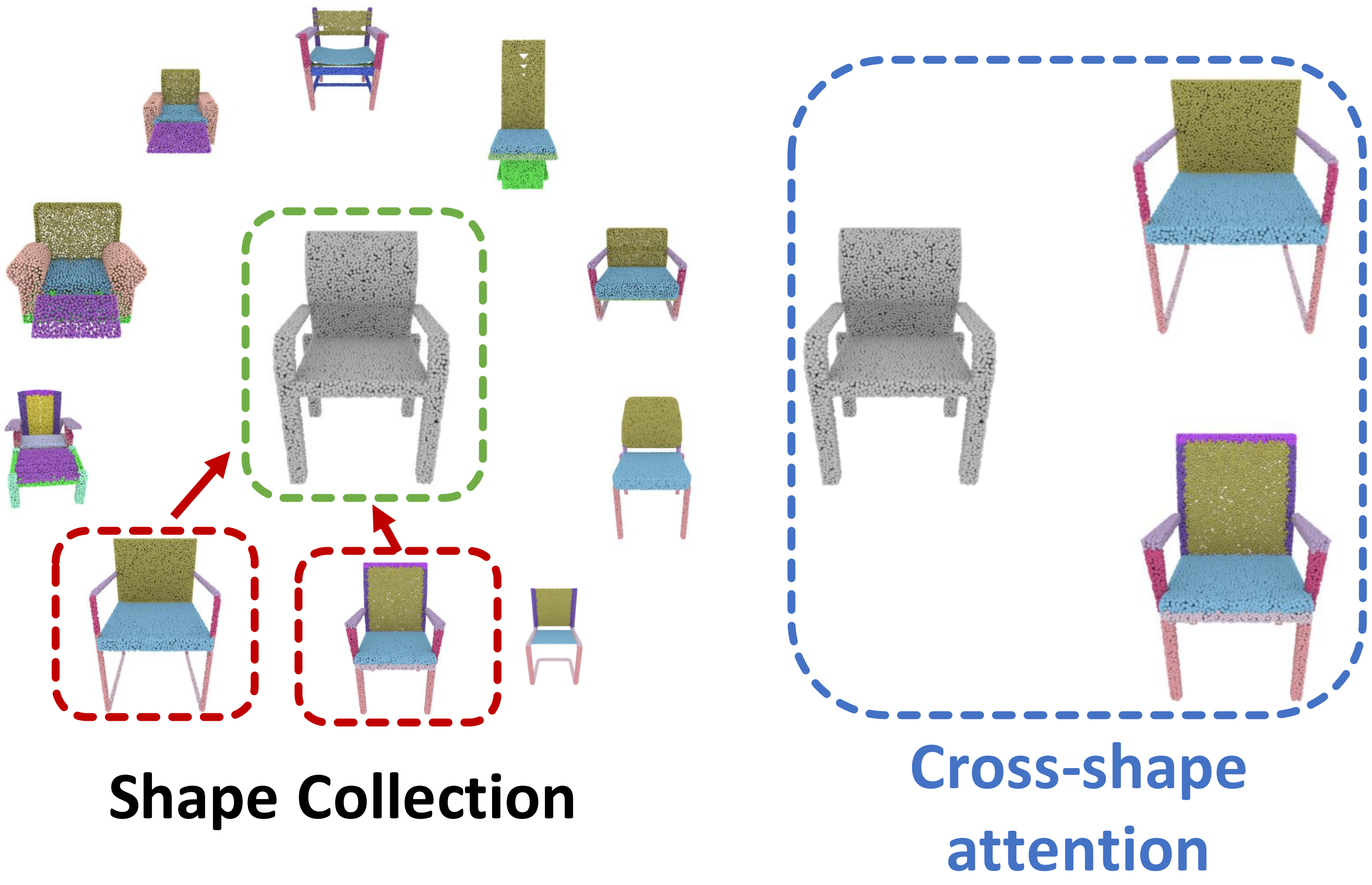
Shape Collection

Pipeline

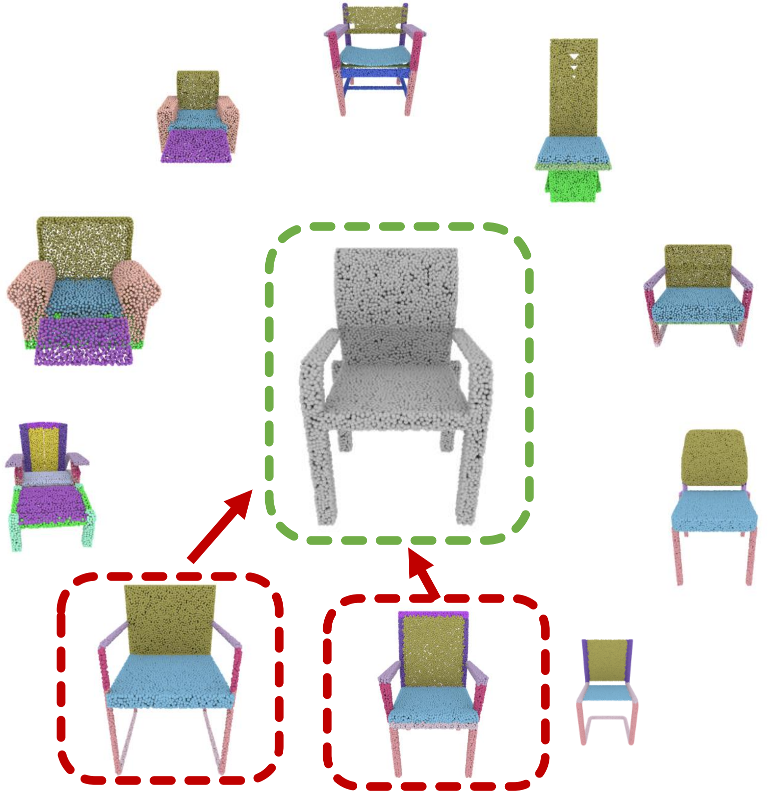


Shape Collection

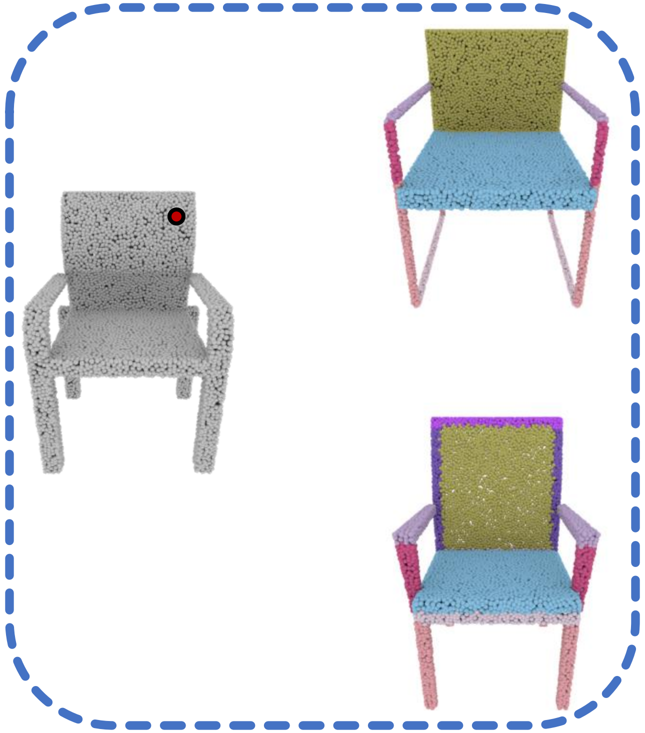
Pipeline



Pipeline

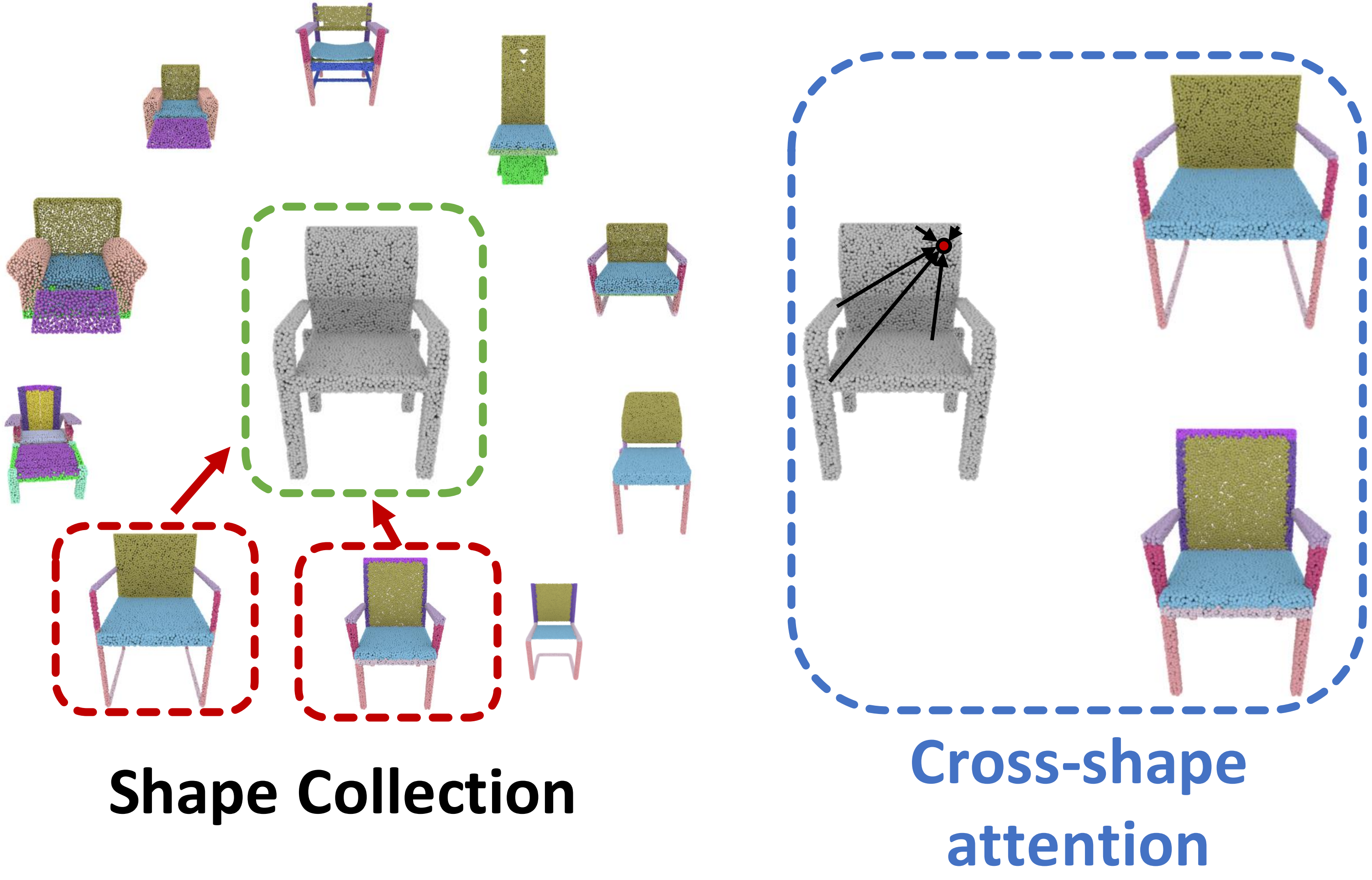


Shape Collection

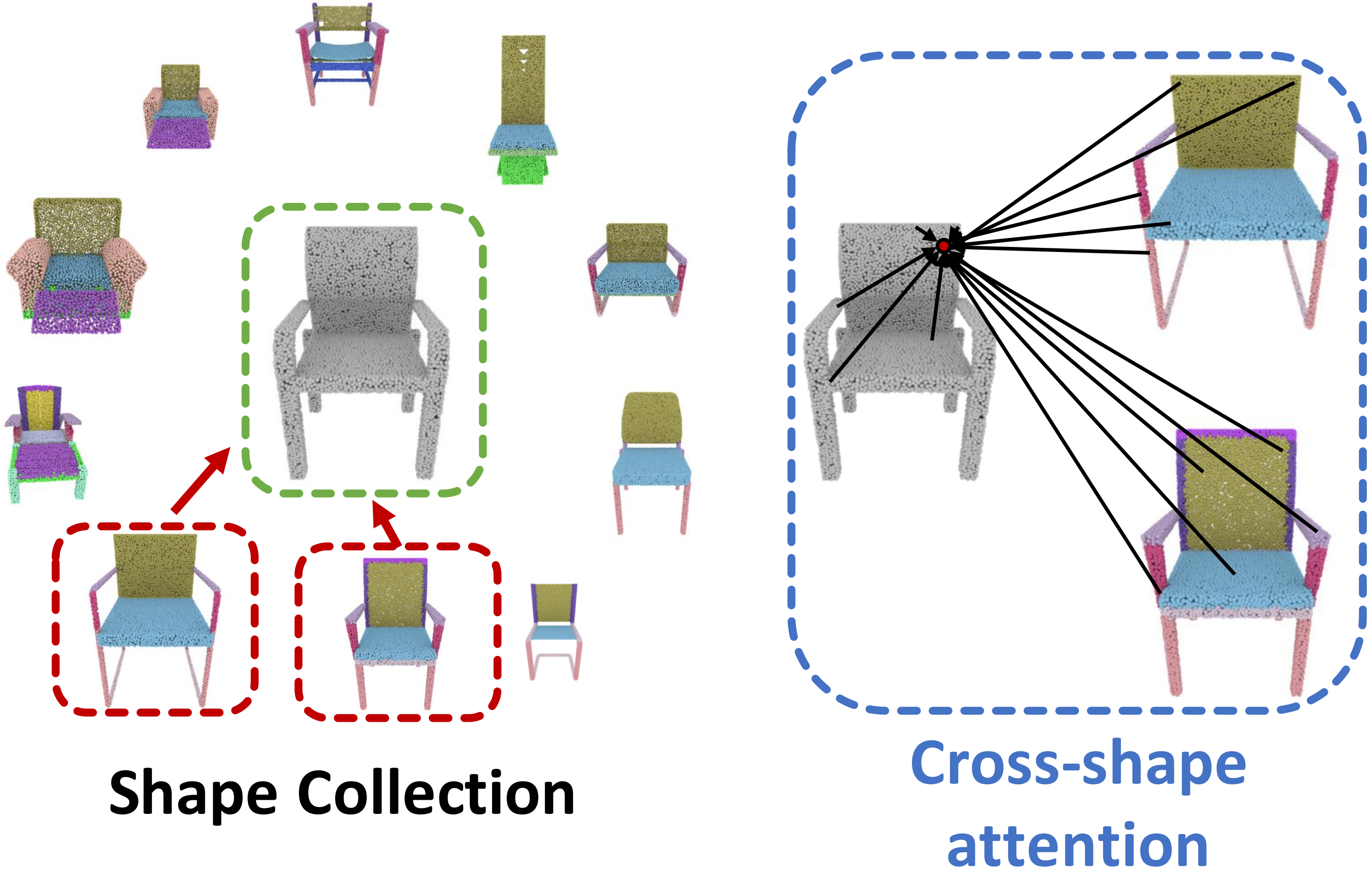


Cross-shape attention

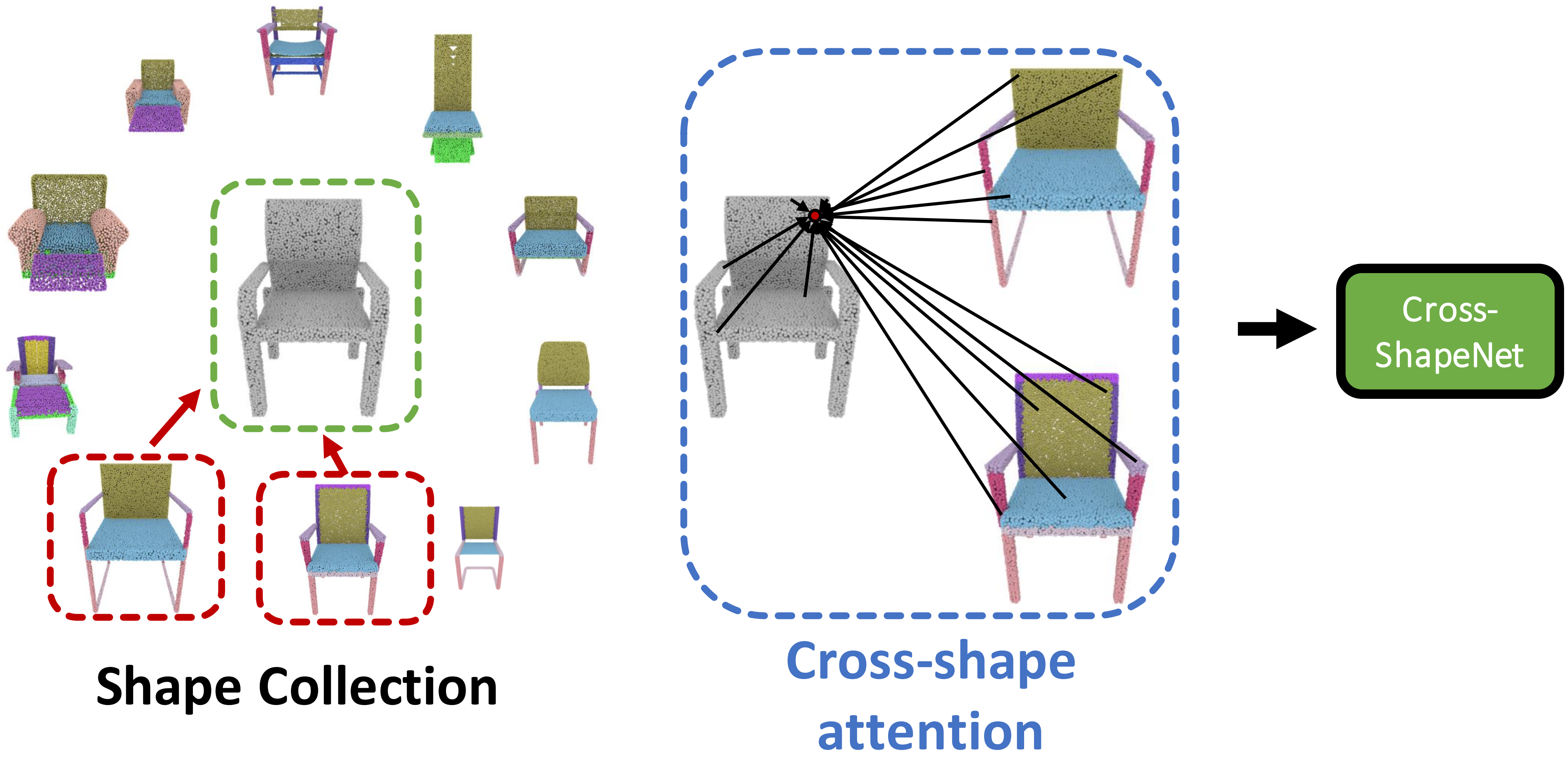
Pipeline



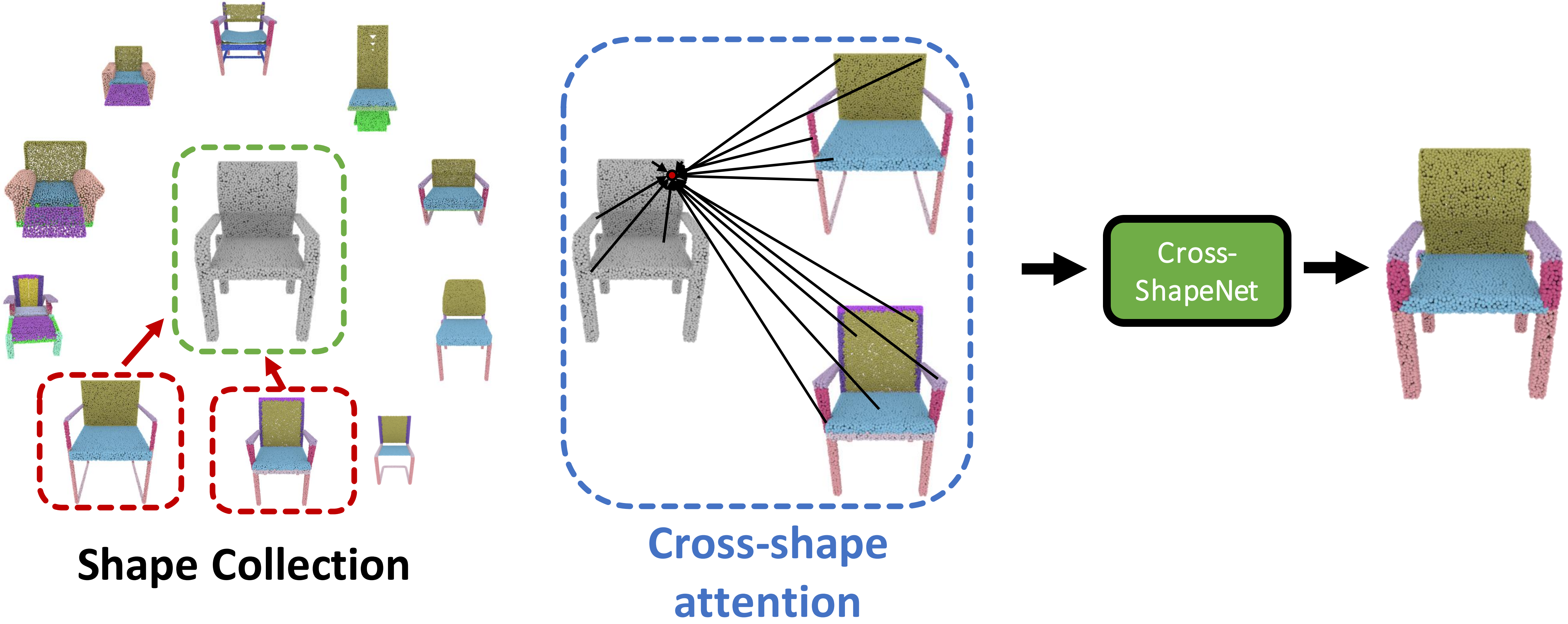
Pipeline



Pipeline

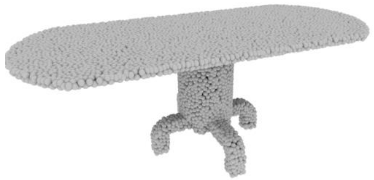


Pipeline



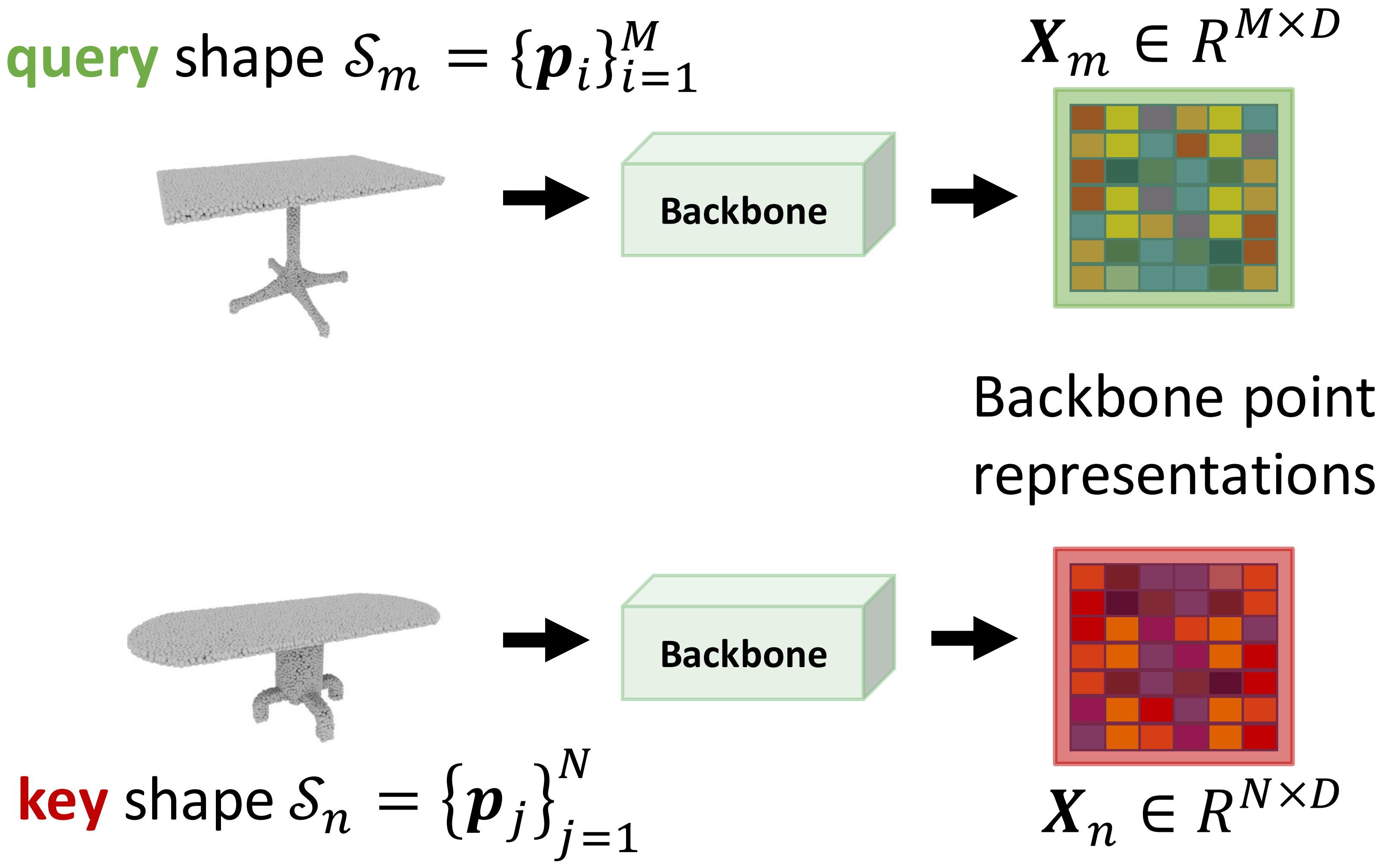
Cross-Shape Attention

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



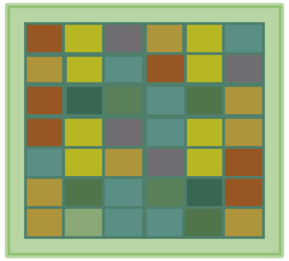
key shape $\mathcal{S}_n = \{\boldsymbol{p}_j\}_{j=1}^{N}$

Cross-Shape Attention

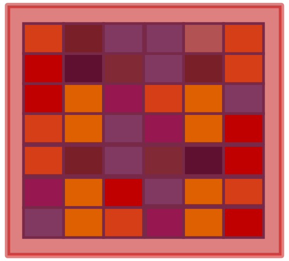


Cross-Shape Attention

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



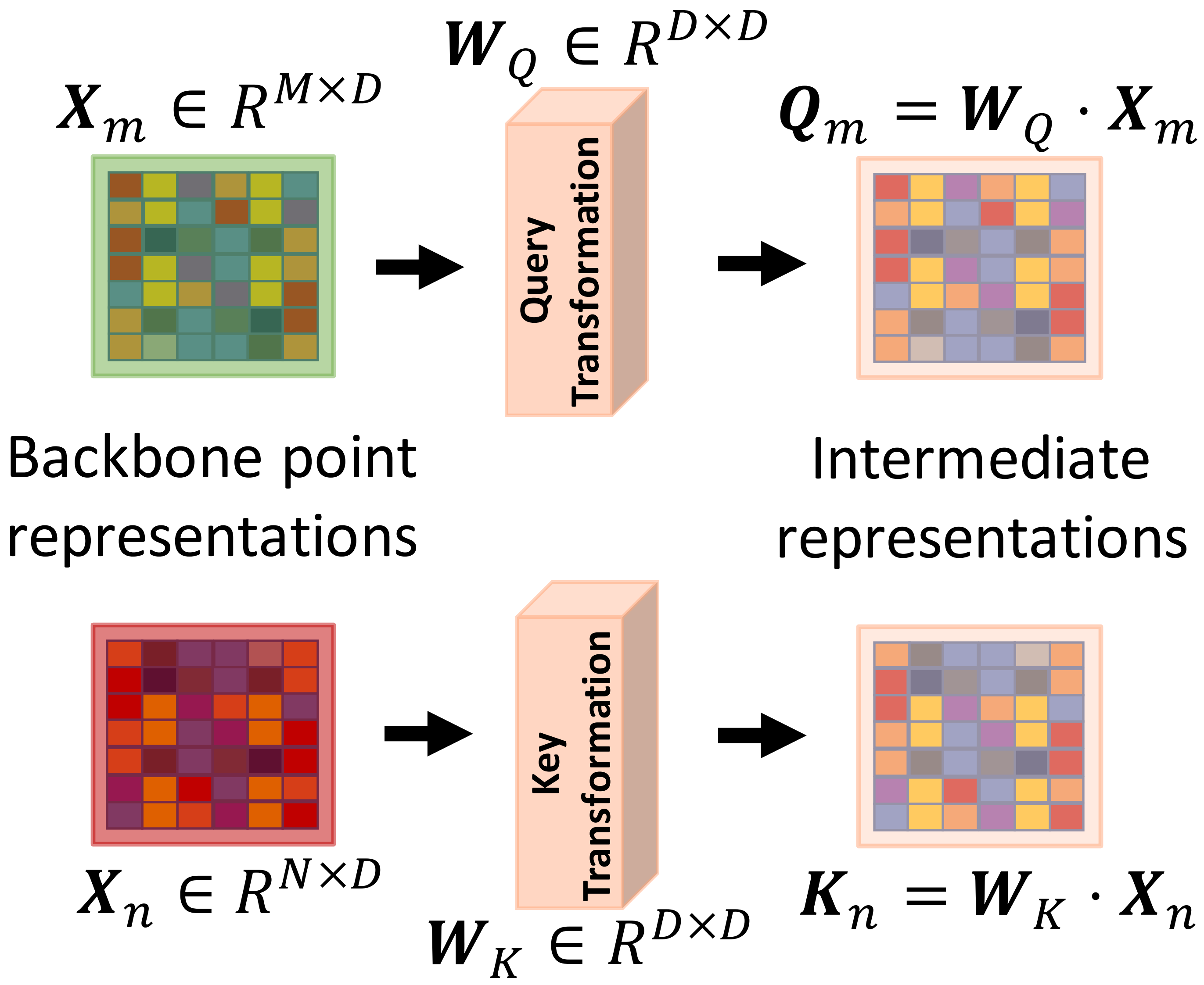
Backbone point representations



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

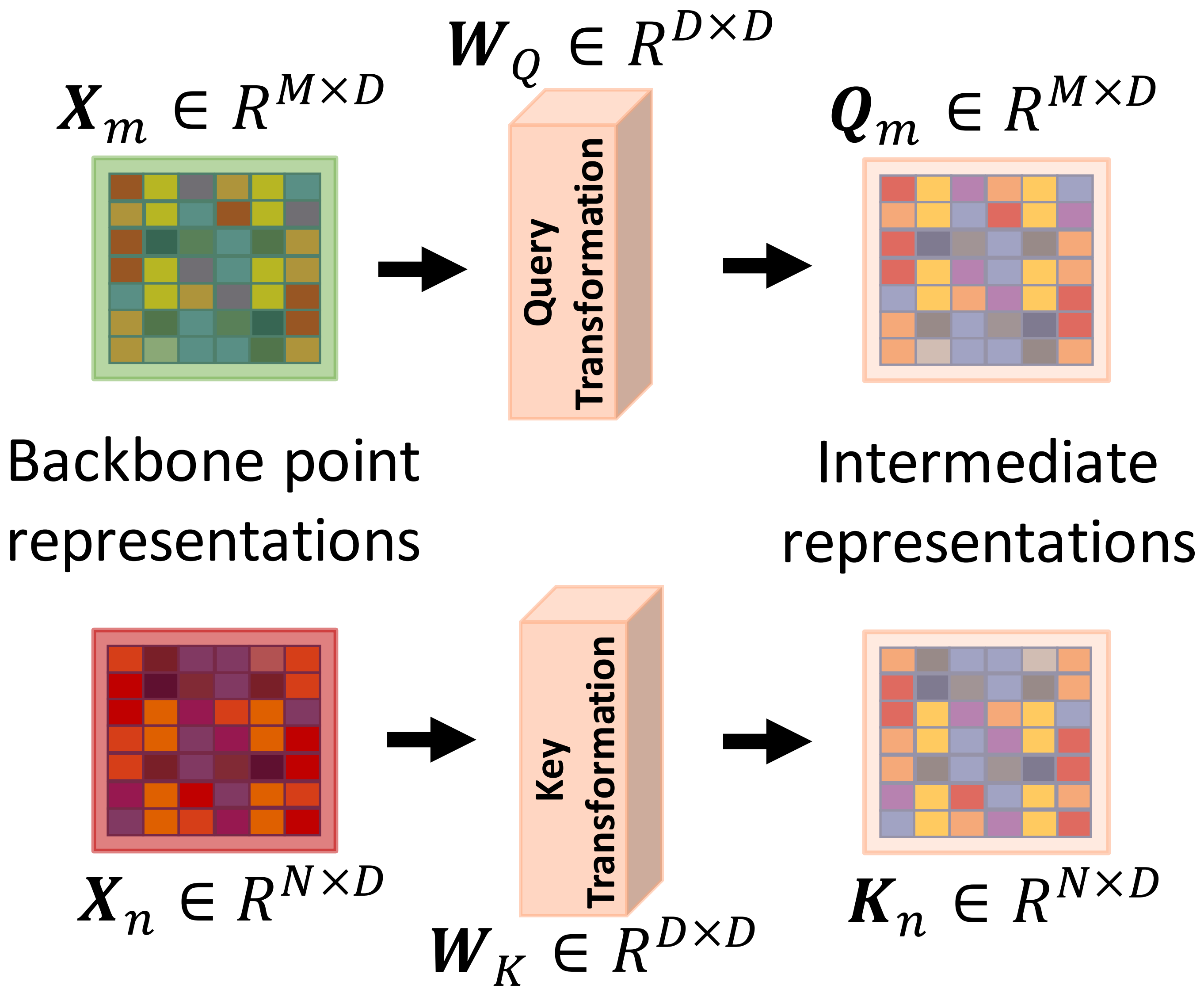
Transformer [Vaswani et al. 2017]

Cross-Shape Attention



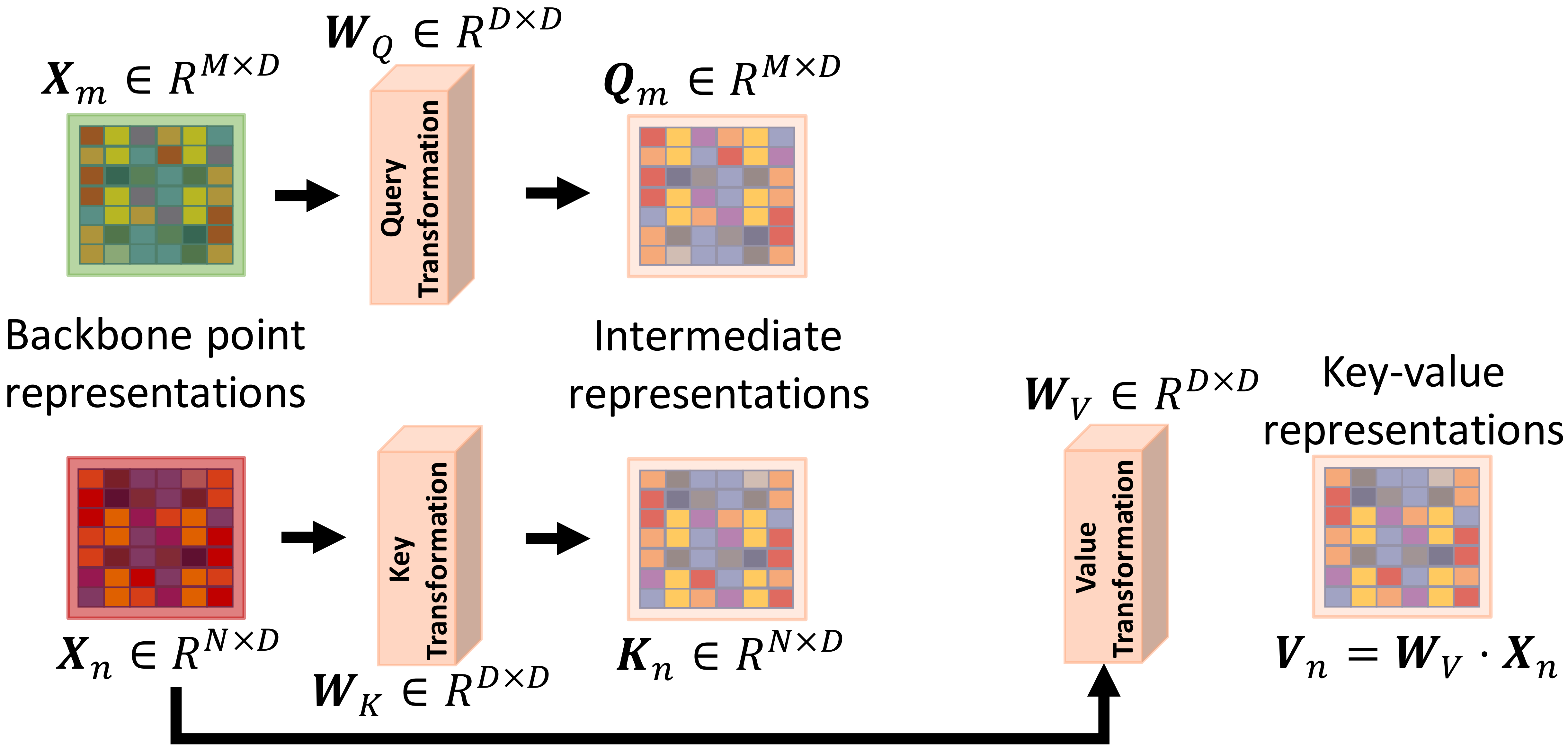
Transformer [Vaswani et al. 2017]

Cross-Shape Attention



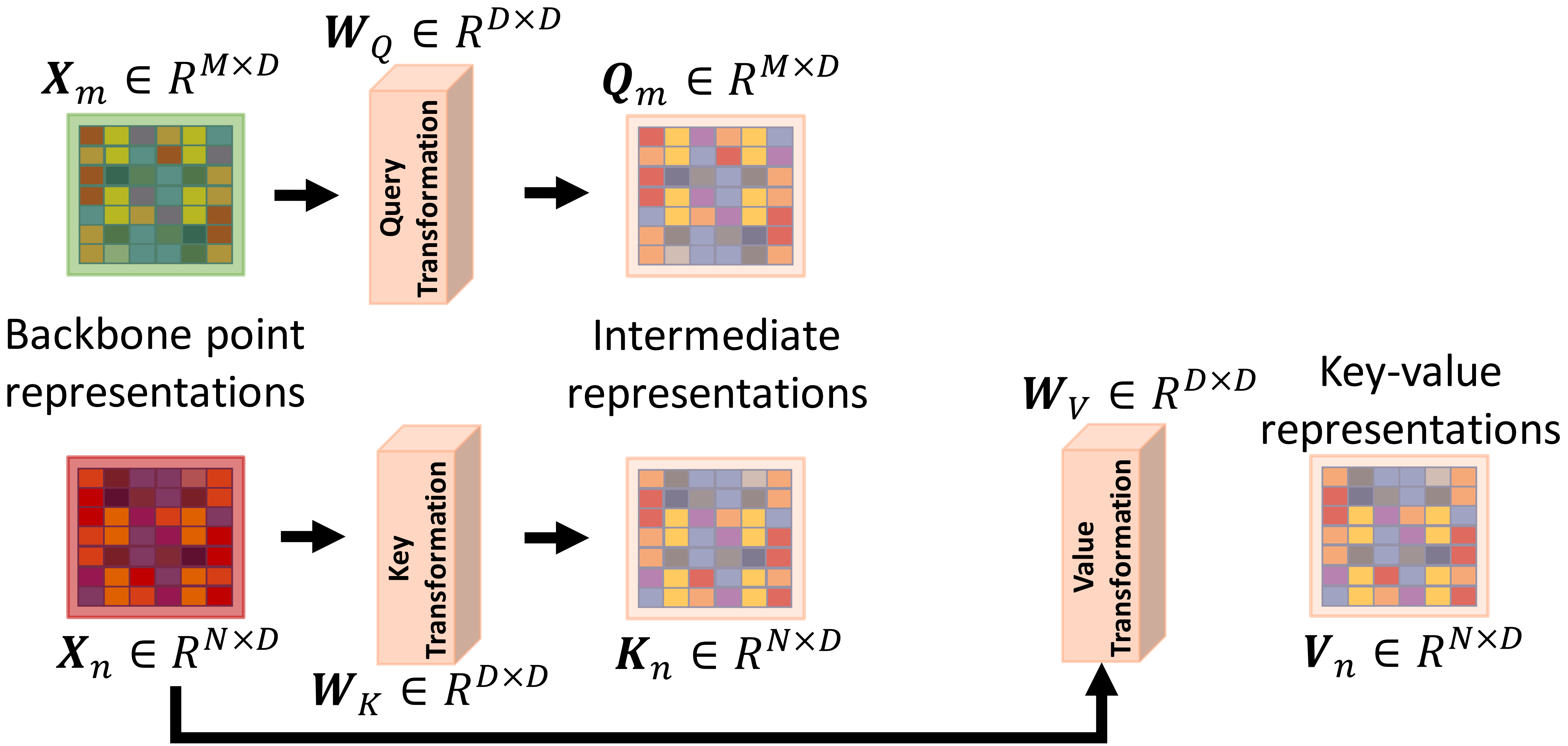
Transformer [Vaswani et al. 2017]

Cross-Shape Attention



Transformer [Vaswani et al. 2017]

Cross-Shape Attention

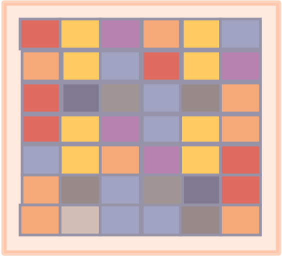


Transformer [Vaswani et al. 2017]

Cross-Shape Attention

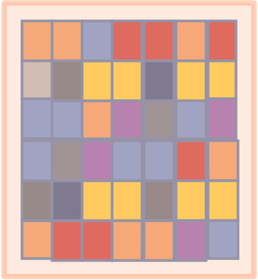
Query
representations

$\boldsymbol{Q}_m$



Key
representations

$\mathbf{K}_n^T$



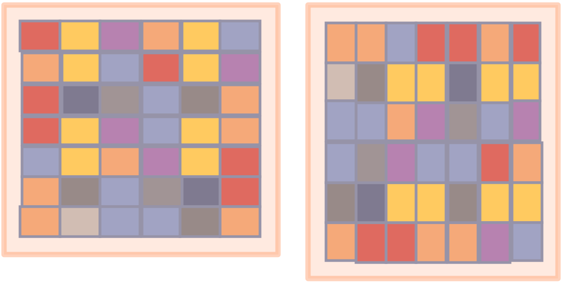
Transformer [Vaswani et al. 2017]

Cross-Shape Attention

$\boldsymbol{Q}_m \in R^{M\times D}$

$\boldsymbol{K}_n \in R^{N\times D}$

Query representations | Key representations



$$\frac{\boldsymbol{Q}_m \mathbf{K}_n^T}{\sqrt{D}}$$

Transformer [Vaswani et al. 2017]

Cross-Shape Attention

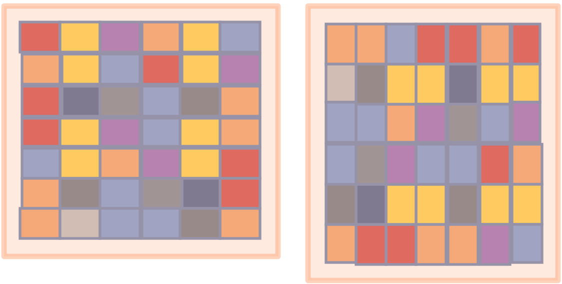
Query representations

Key representations

$\boldsymbol{Q}_m$ $\mathbf{K}_n^T$

$\boldsymbol{Q}_m \in R^{M \times D}$

$\boldsymbol{K}_n \in R^{N \times D}$



$\sqrt{D}$

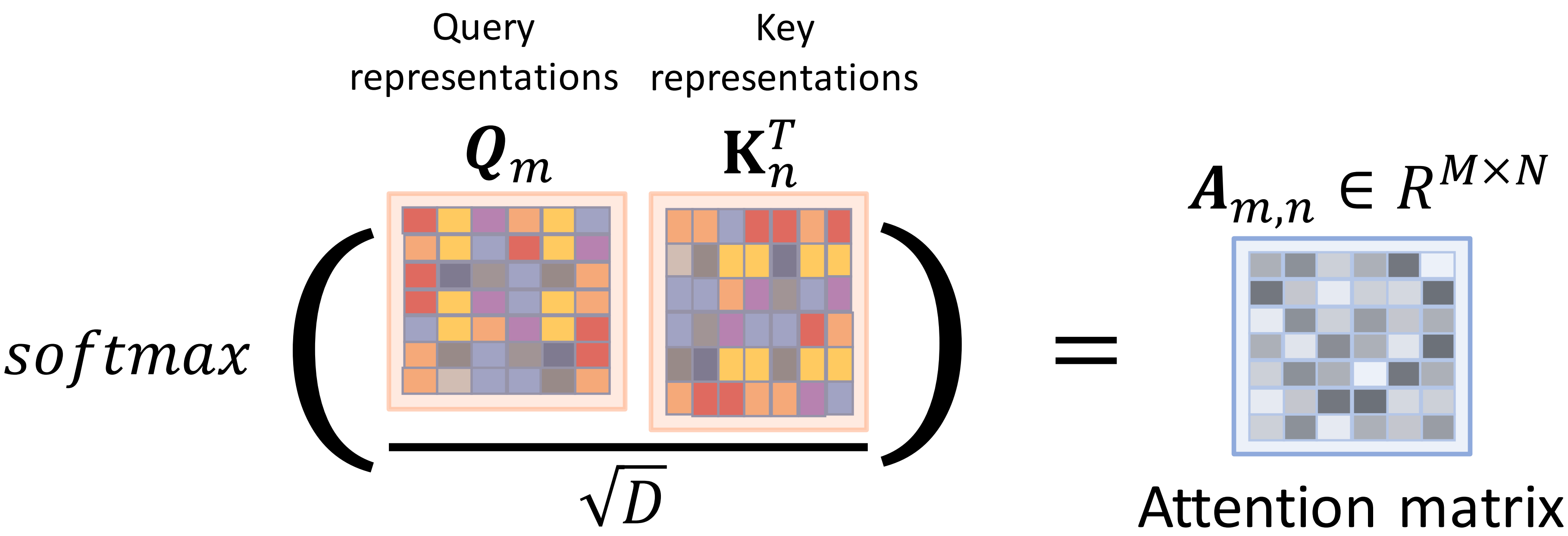
$$Var\left(\frac{\boldsymbol{Q}_{i,:}\boldsymbol{K}_{:,j}^T}{\sqrt{D}}\right) = 1,$$

$$\forall i = 1, \cdots, M$$

$$\forall j = 1, \cdots, N$$

Transformer [Vaswani et al. 2017]

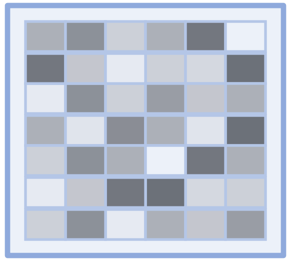
Cross-Shape Attention



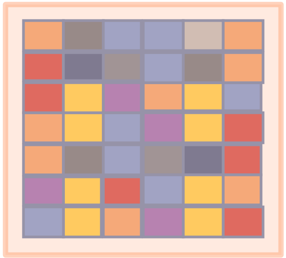
Transformer [Vaswani et al. 2017]

Cross-Shape Attention

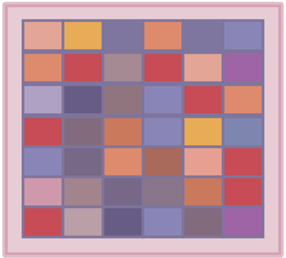
$$\boldsymbol{A}_{m,n} \cdot \boldsymbol{V}_n = \boldsymbol{X'}_m^{(CSA)} \in R^{M \times D}$$



Cross-shape attention matrix



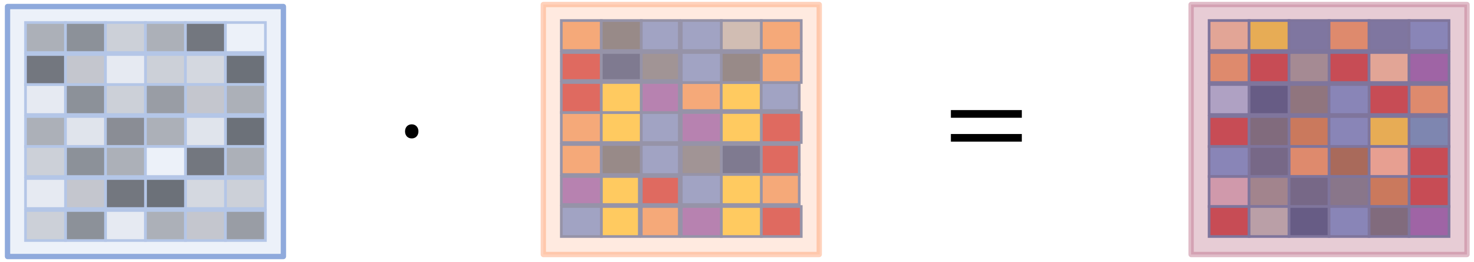
Key shape value representations



Cross-shape attention representations

Cross-Shape Attention

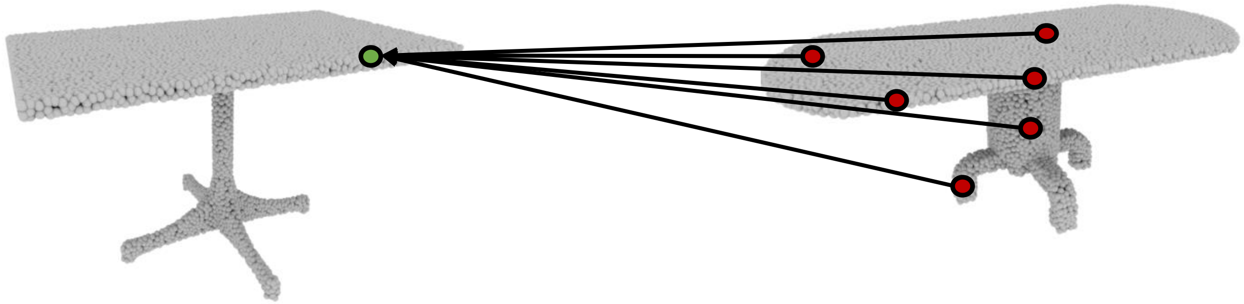
$$\boldsymbol{A}_{m,n} \cdot \boldsymbol{V}_n = \boldsymbol{X'}_m^{(CSA)} \in R^{M \times D}$$



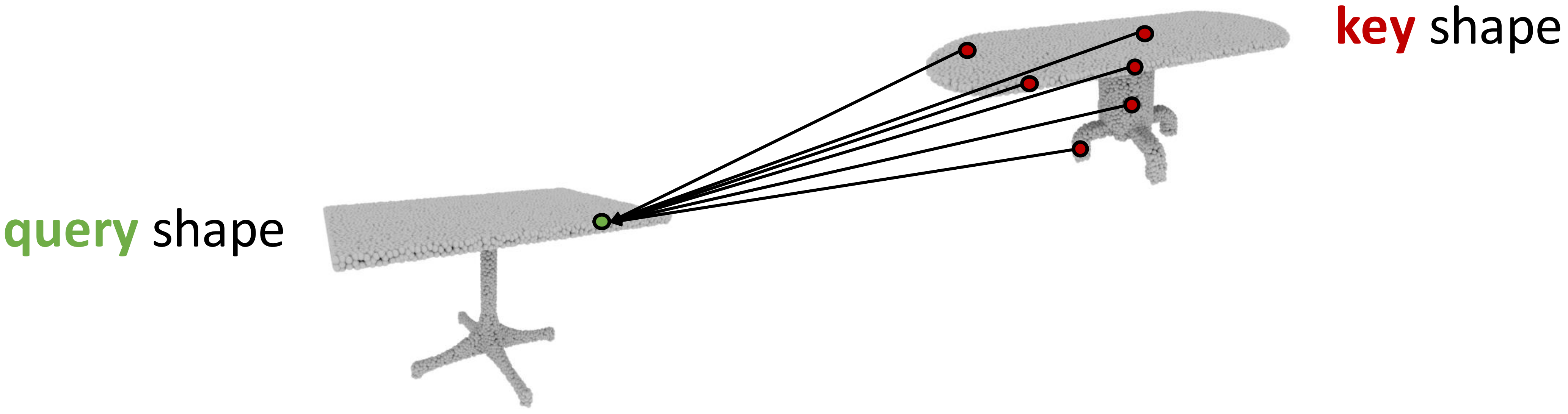
Cross-shape attention matrix

Key shape value representations

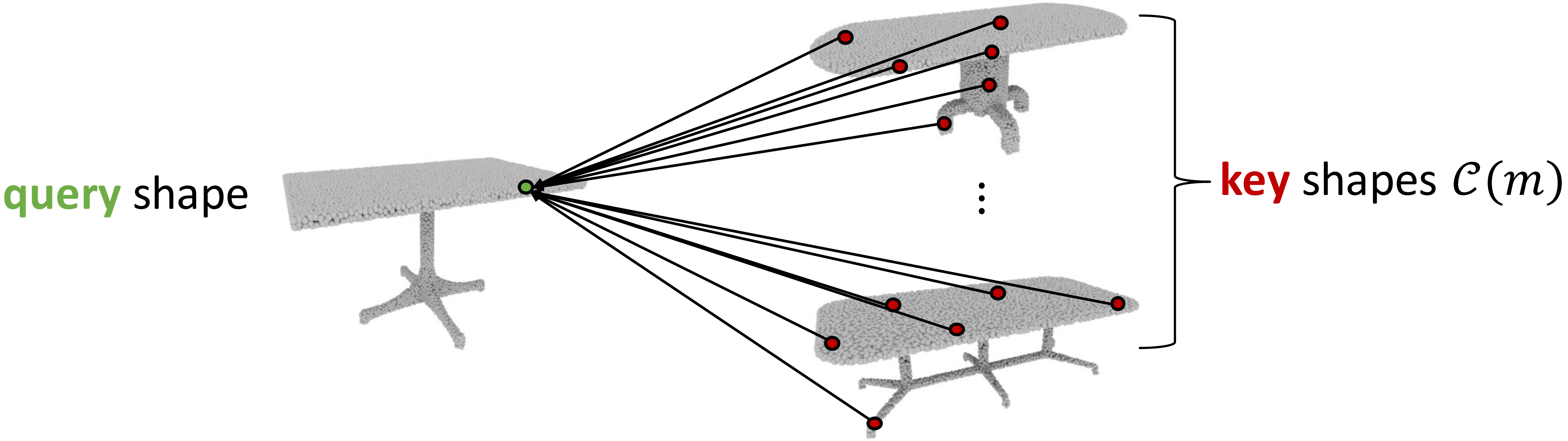
Cross-shape attention representations



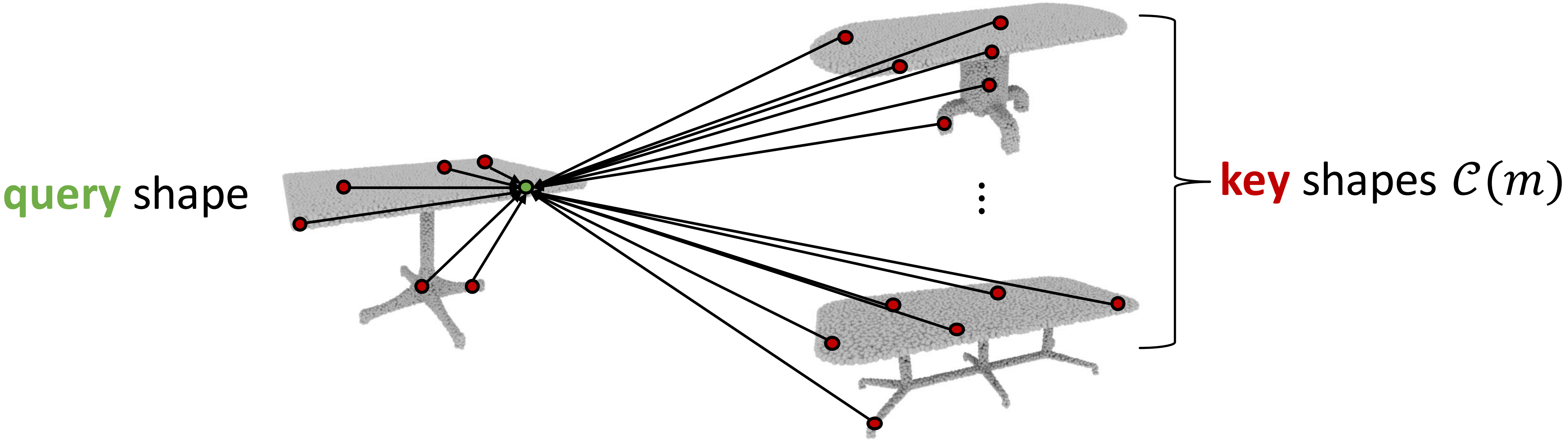
Cross-Shape Attention for multiple shapes



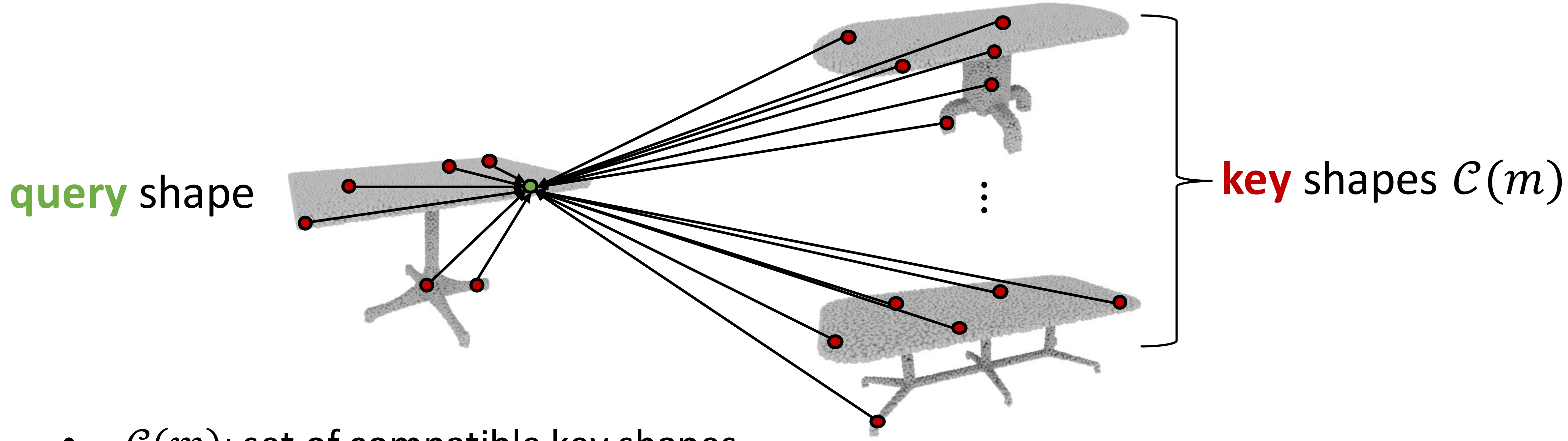
Cross-Shape Attention for multiple shapes



Cross-Shape Attention for multiple shapes



Cross-Shape Attention for multiple shapes



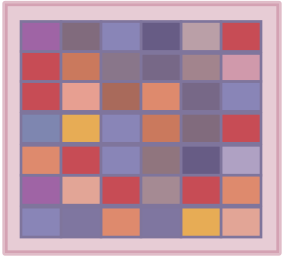
- $\mathcal{C}(m)$: set of compatible key shapes
- $c(m,n)$: compatibility function between query shape S_m and key shape S_n

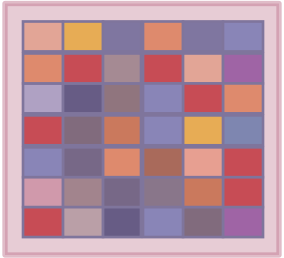
Cross-shape attention output

$$\boldsymbol{X}'_m = \sum_{n \in \{\mathcal{C}(m), m\}} c(m,n) \boldsymbol{A}_{m,n} \boldsymbol{V}_n$$

Compatibility function

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

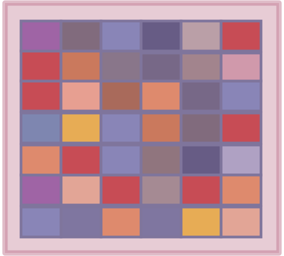




$\boldsymbol{X'}_n^{(SSA)} \in R^{N \times D}$

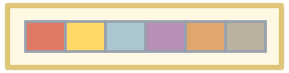
Compatibility function

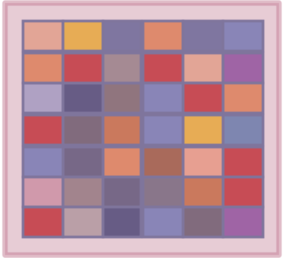
$\boldsymbol{X'}_m^{(SSA)} \in R^{M \times D}$



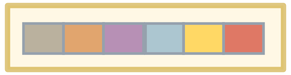
$\underset{i}{\text{avg}}\, \boldsymbol{X'}_{m,i}^{(SSA)}$

$\boldsymbol{y}_m^{(SSA)} \in R^{D}$





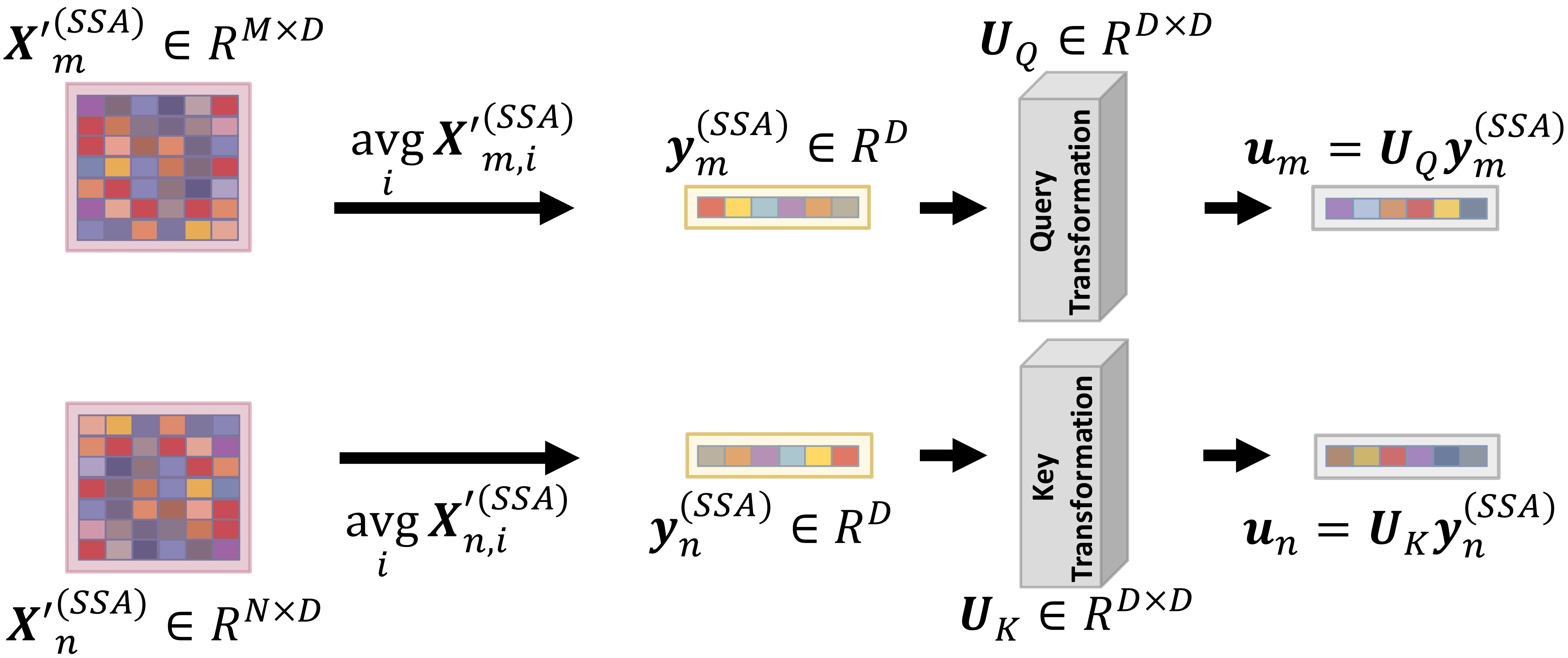
$\underset{i}{\text{avg}}\, \boldsymbol{X'}_{n,i}^{(SSA)}$



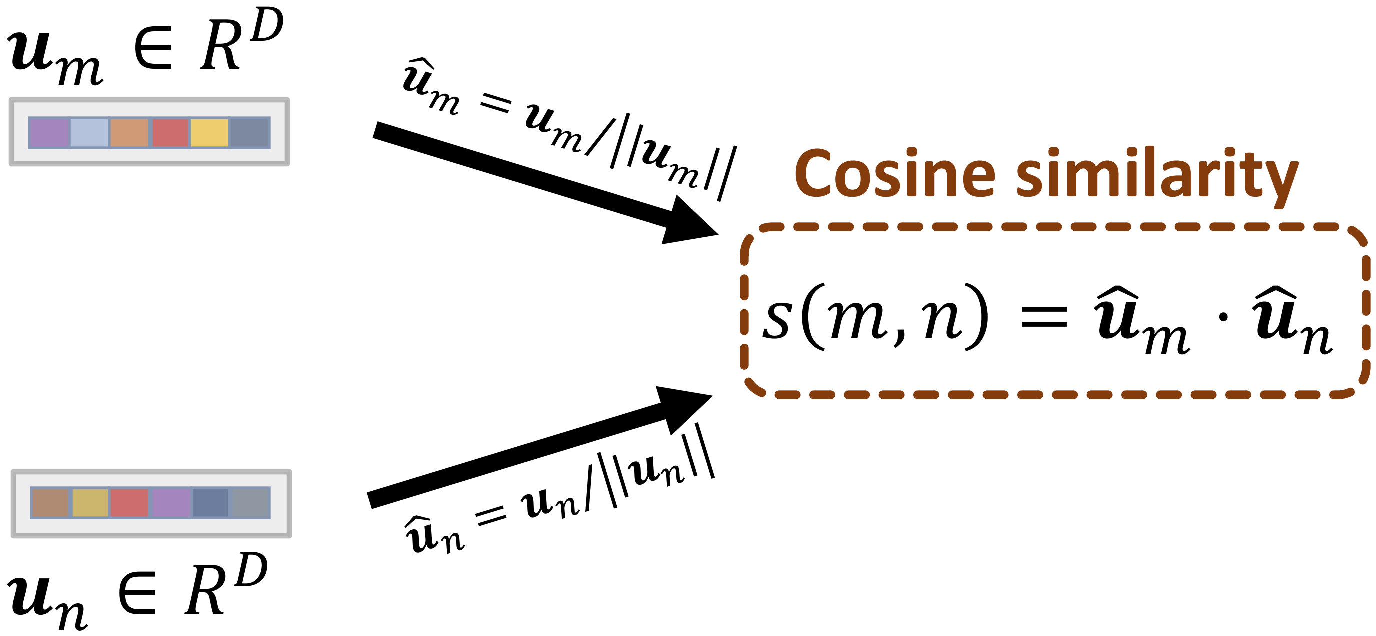
$\boldsymbol{y}_n^{(SSA)} \in R^{D}$

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

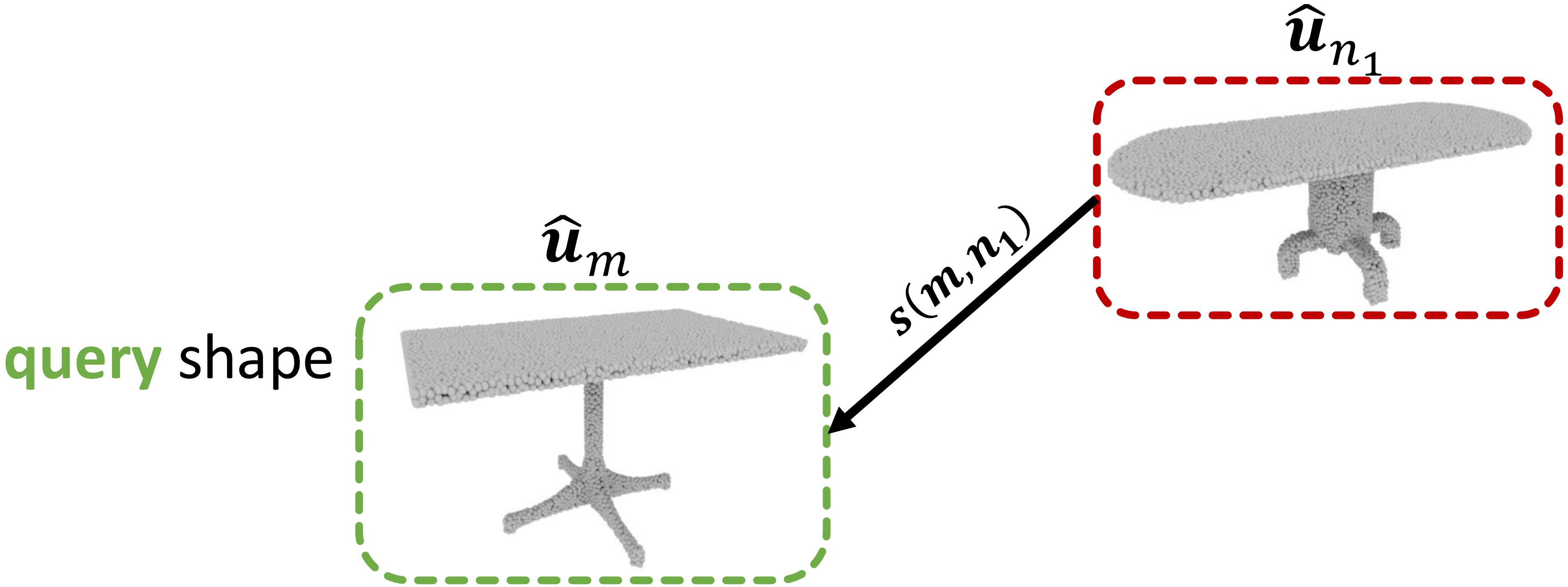
Compatibility function



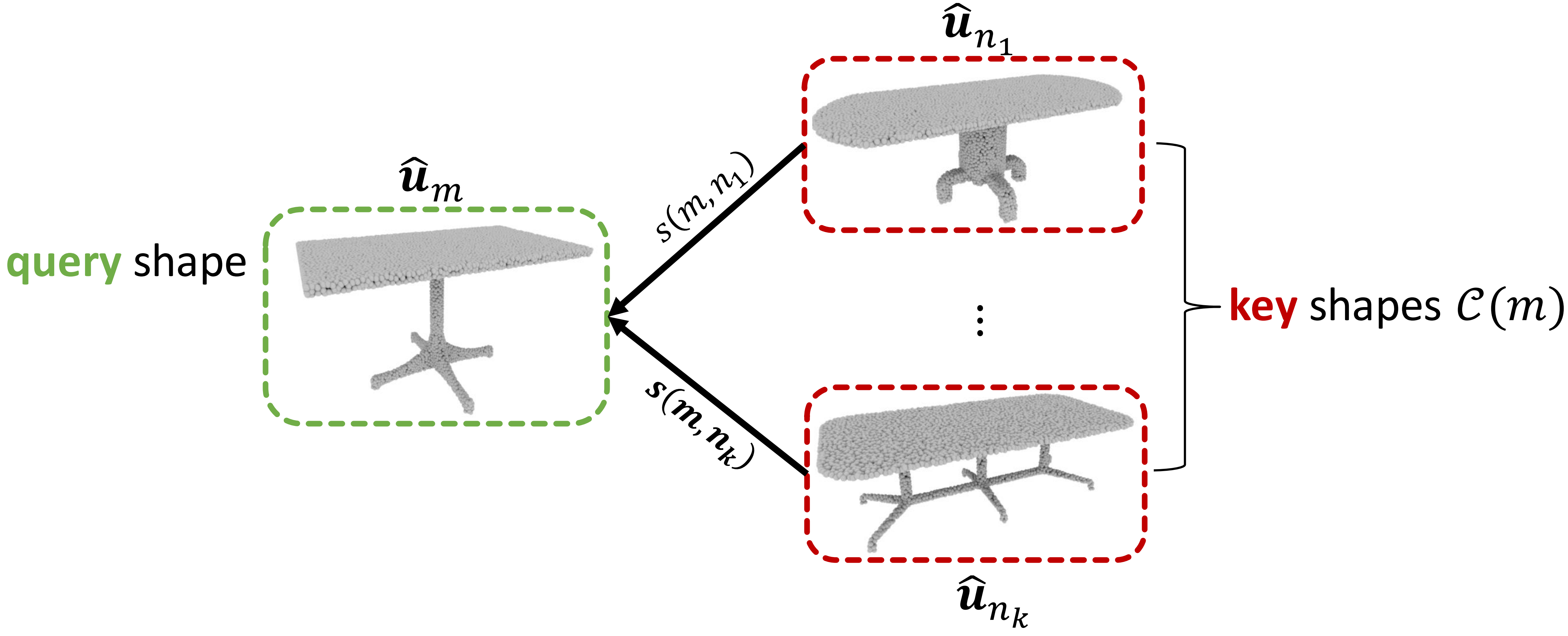
Compatibility function



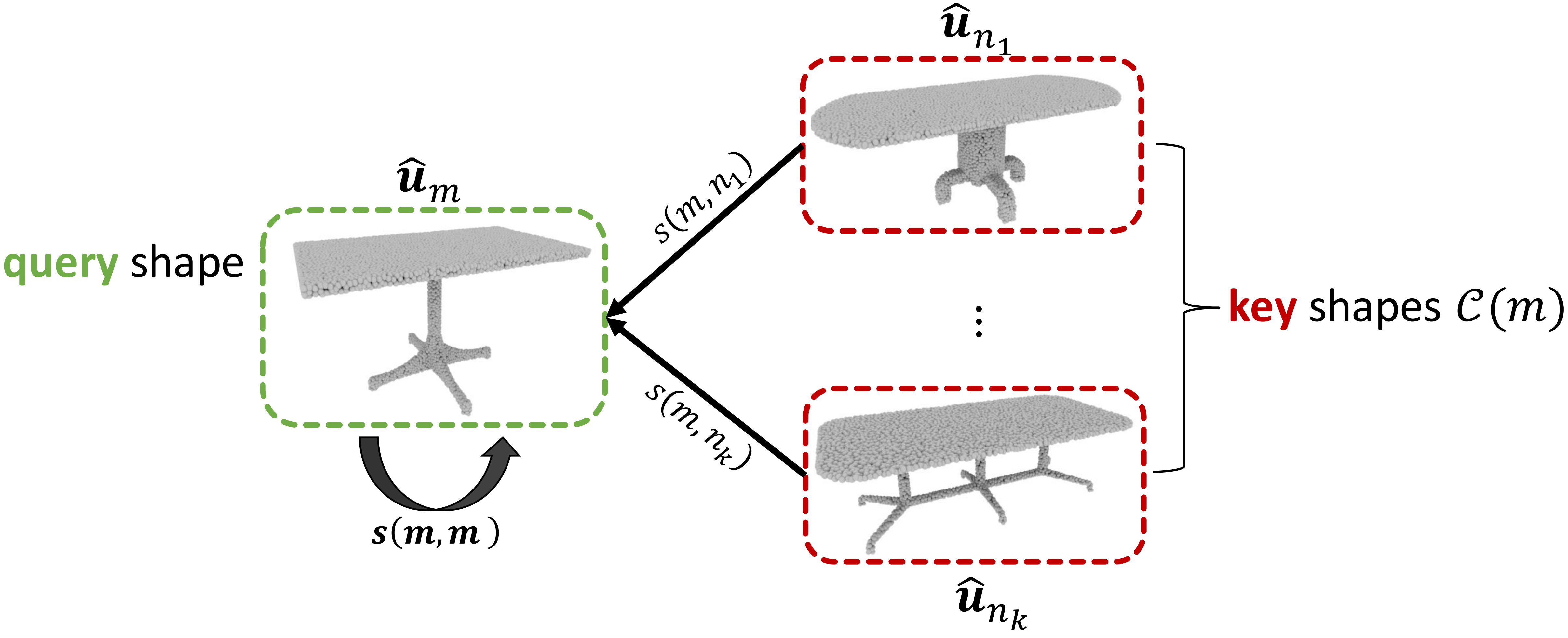
Compatibility function



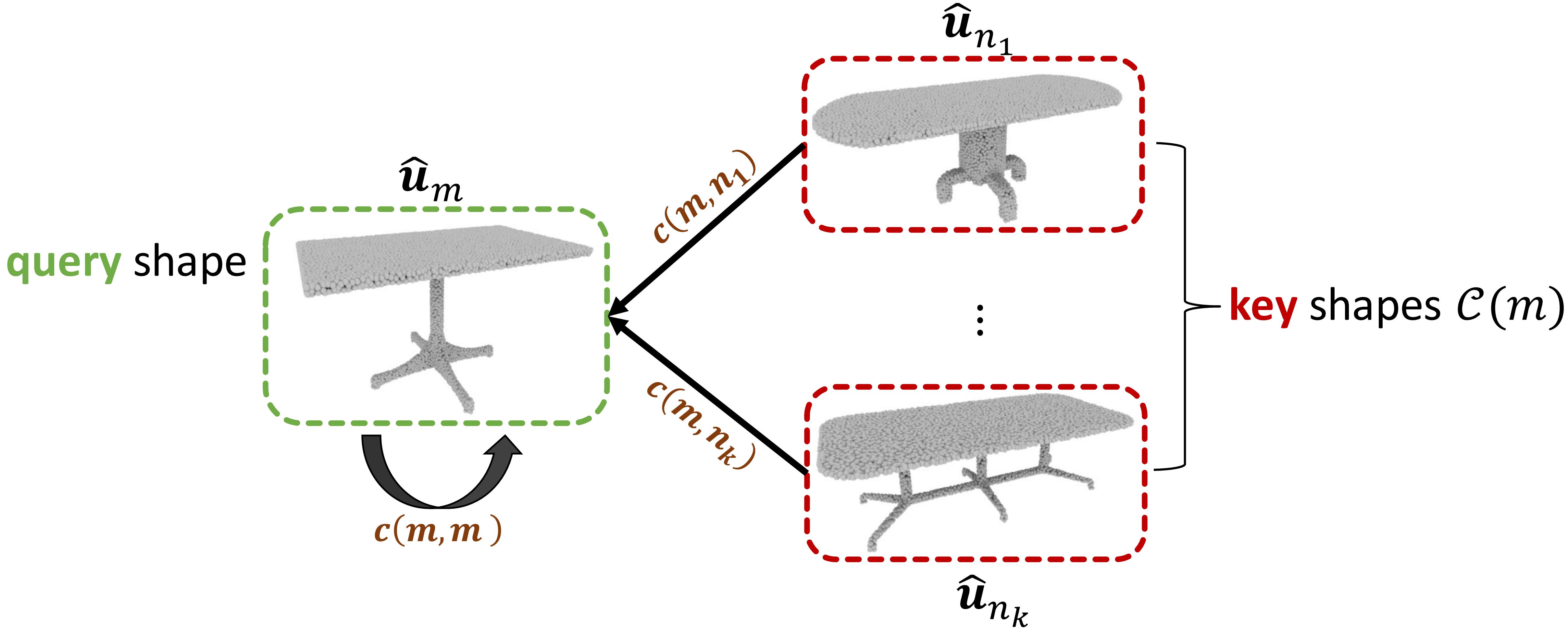
Compatibility function



Compatibility function

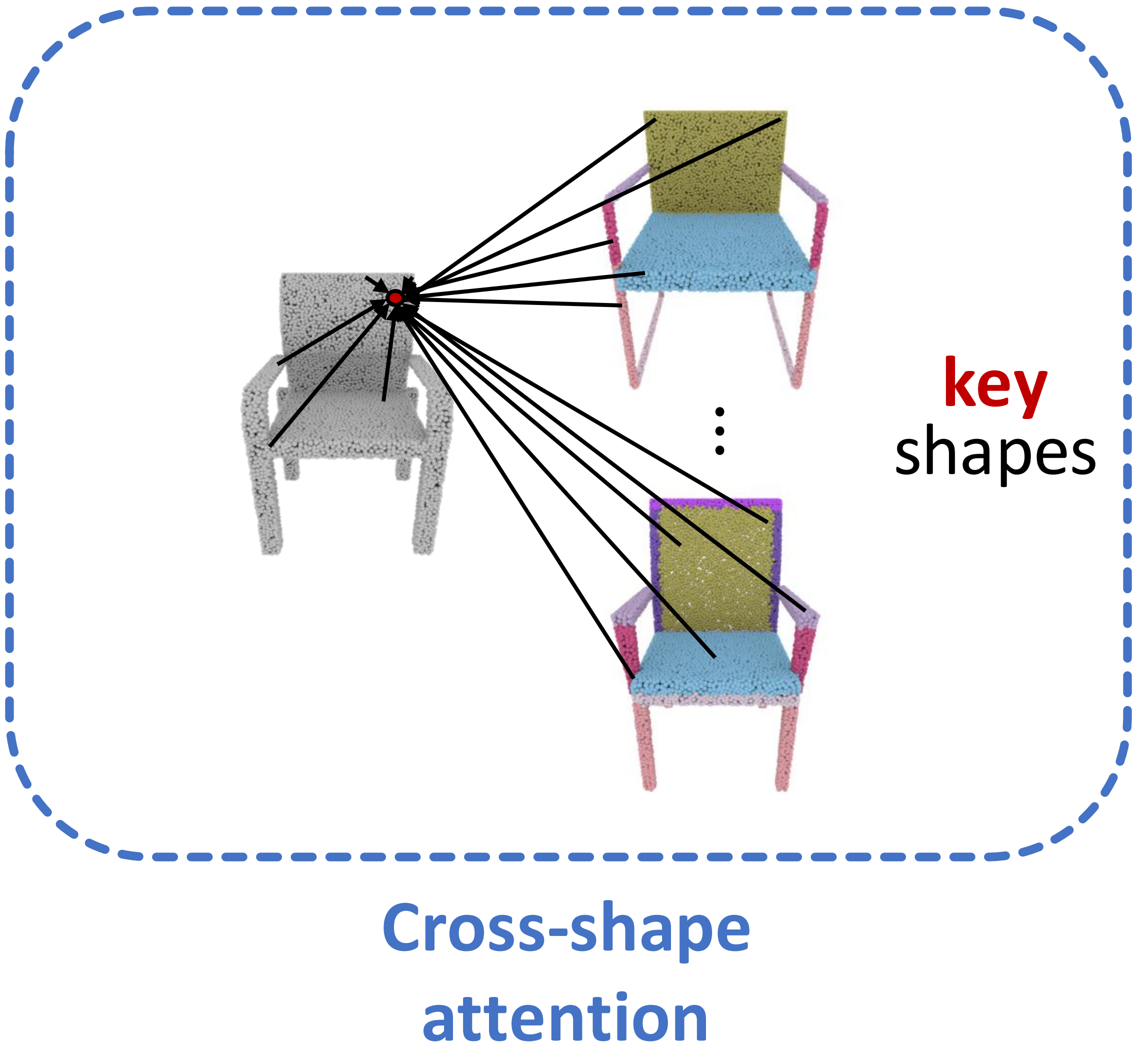


Compatibility function

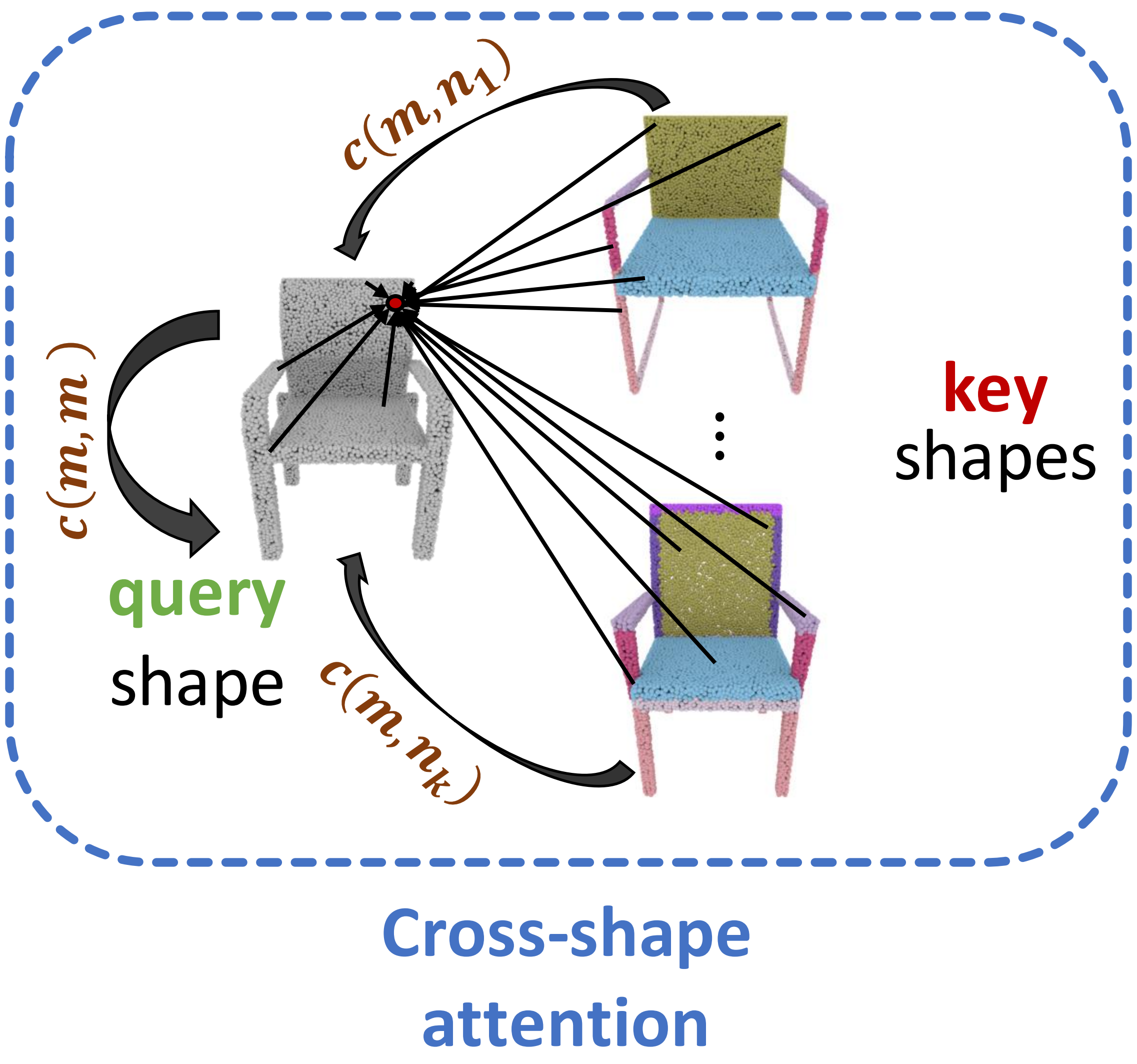


compatibility $$c(m,n) = \frac{e^{s(m,n)}}{\sum_{n \in \{\mathcal{C}(m), m\}} e^{s(m,n)}}$$

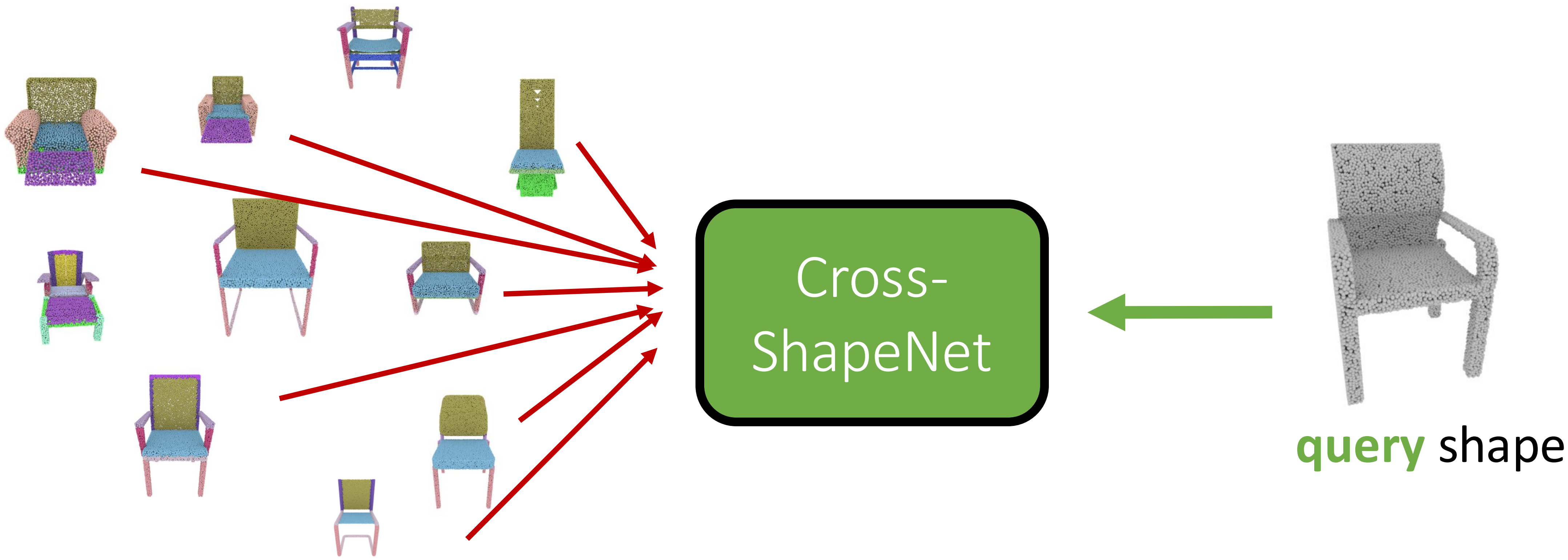
Cross-Shape Attention for multiple shapes



Cross-Shape Attention for multiple shapes

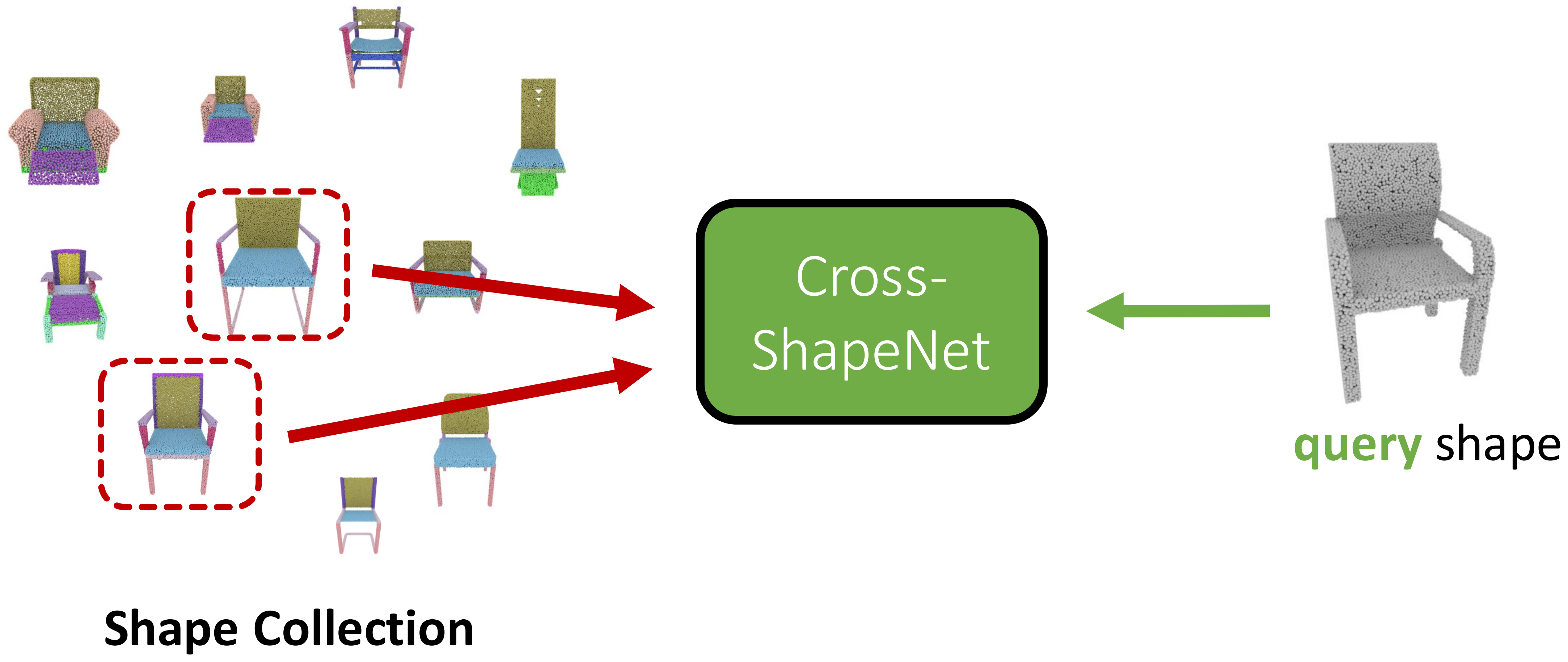


Retrieve compatible shapes

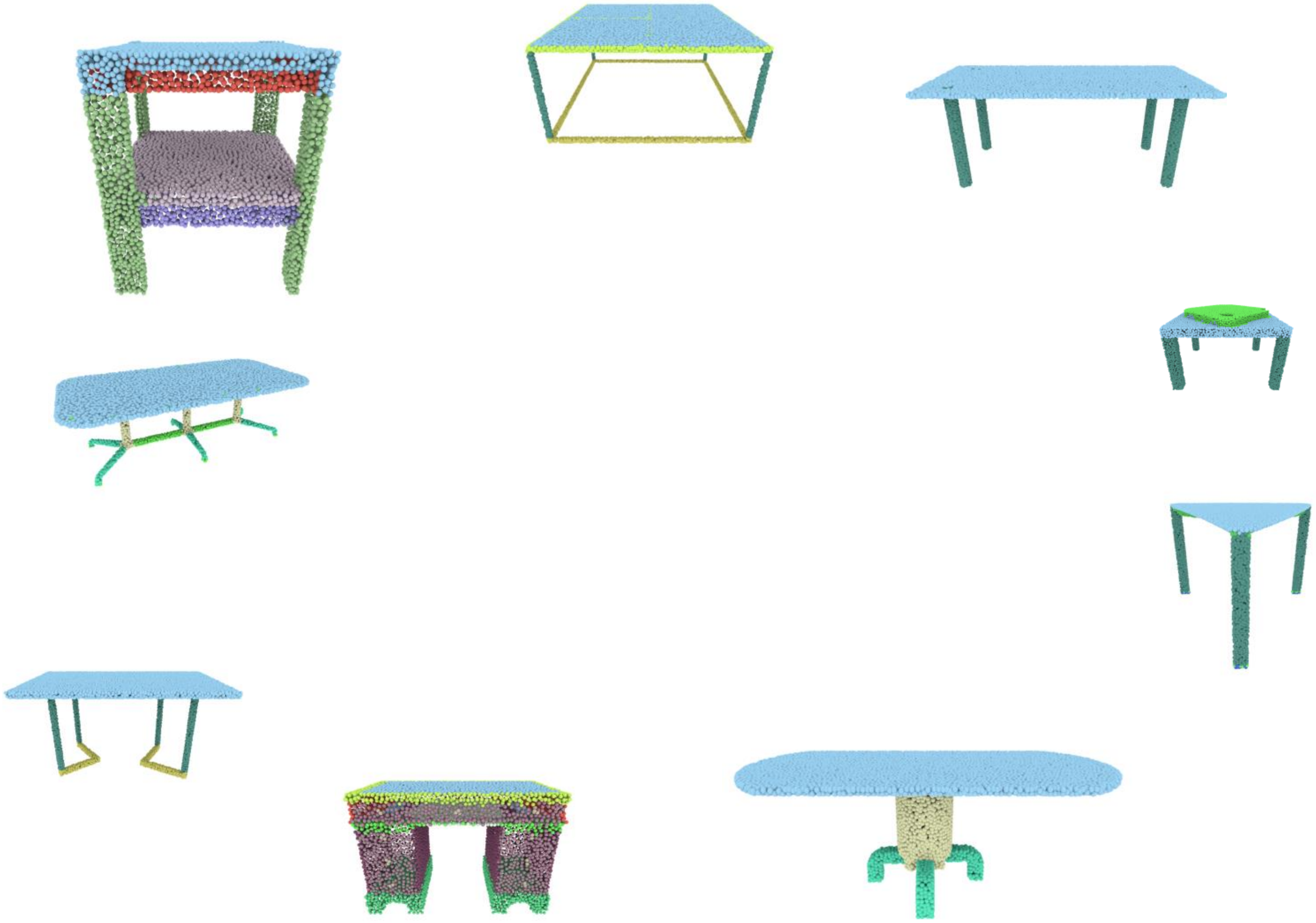


Shape Collection

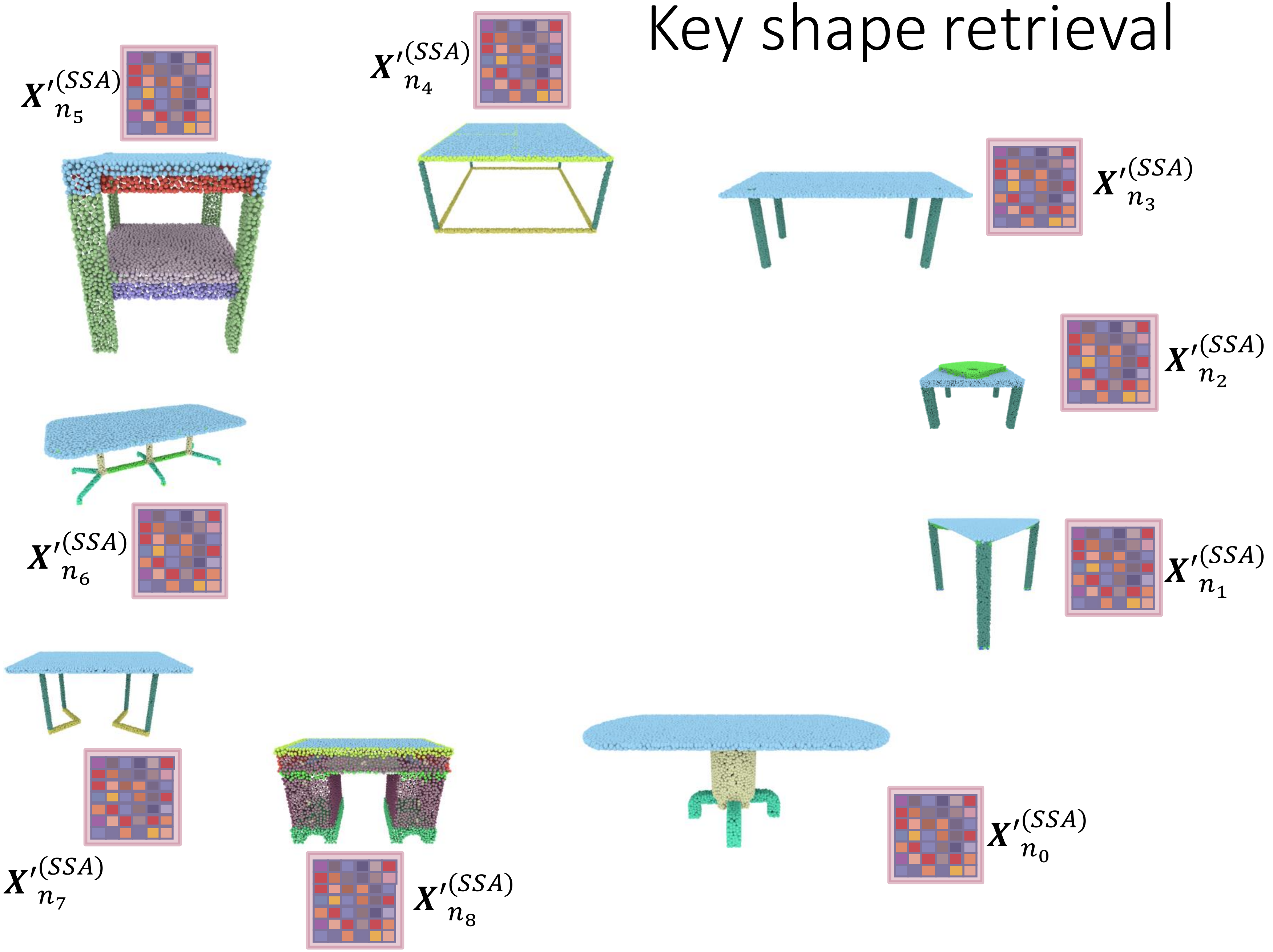
Retrieve compatible shapes



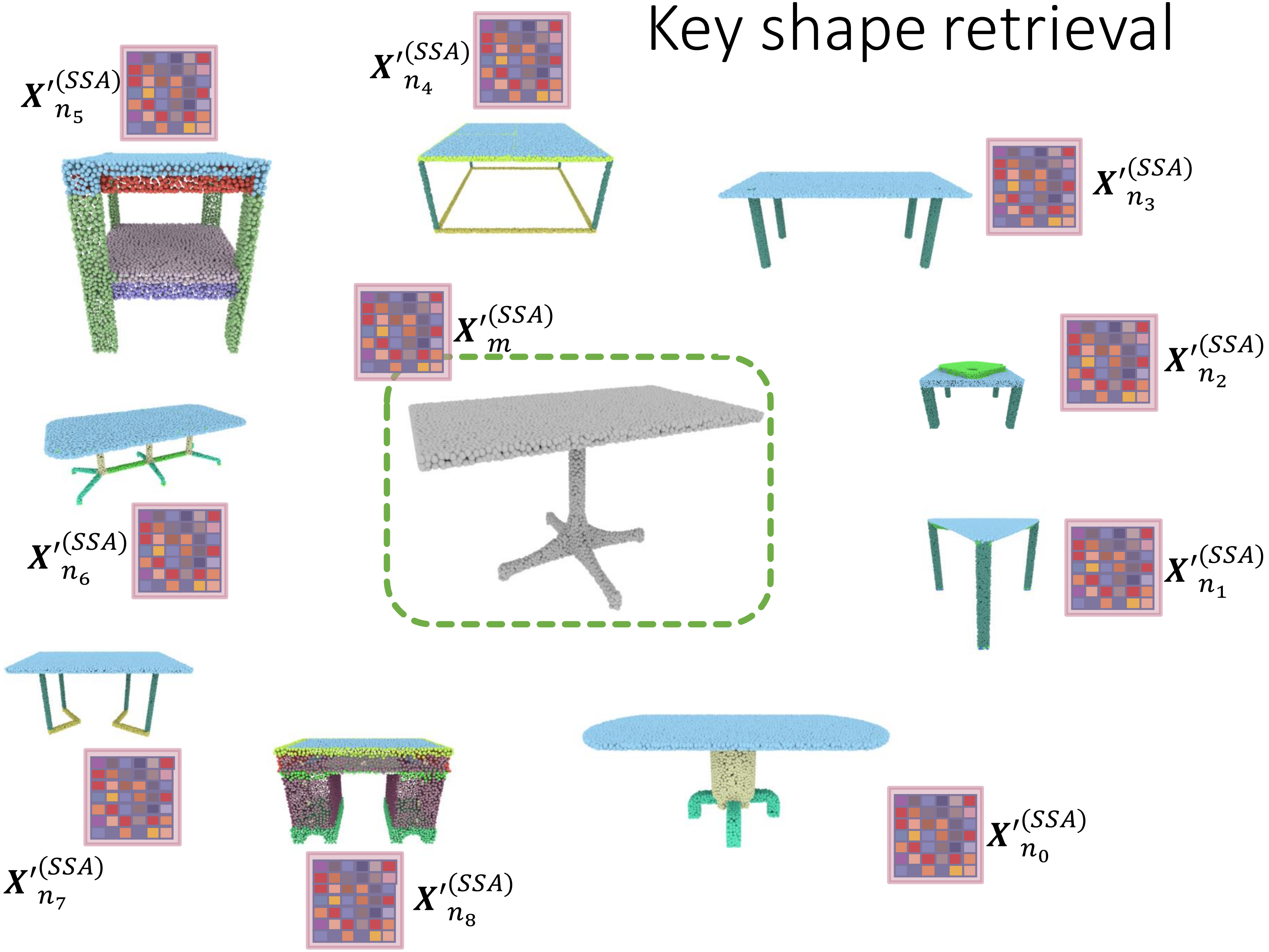
Key shape retrieval



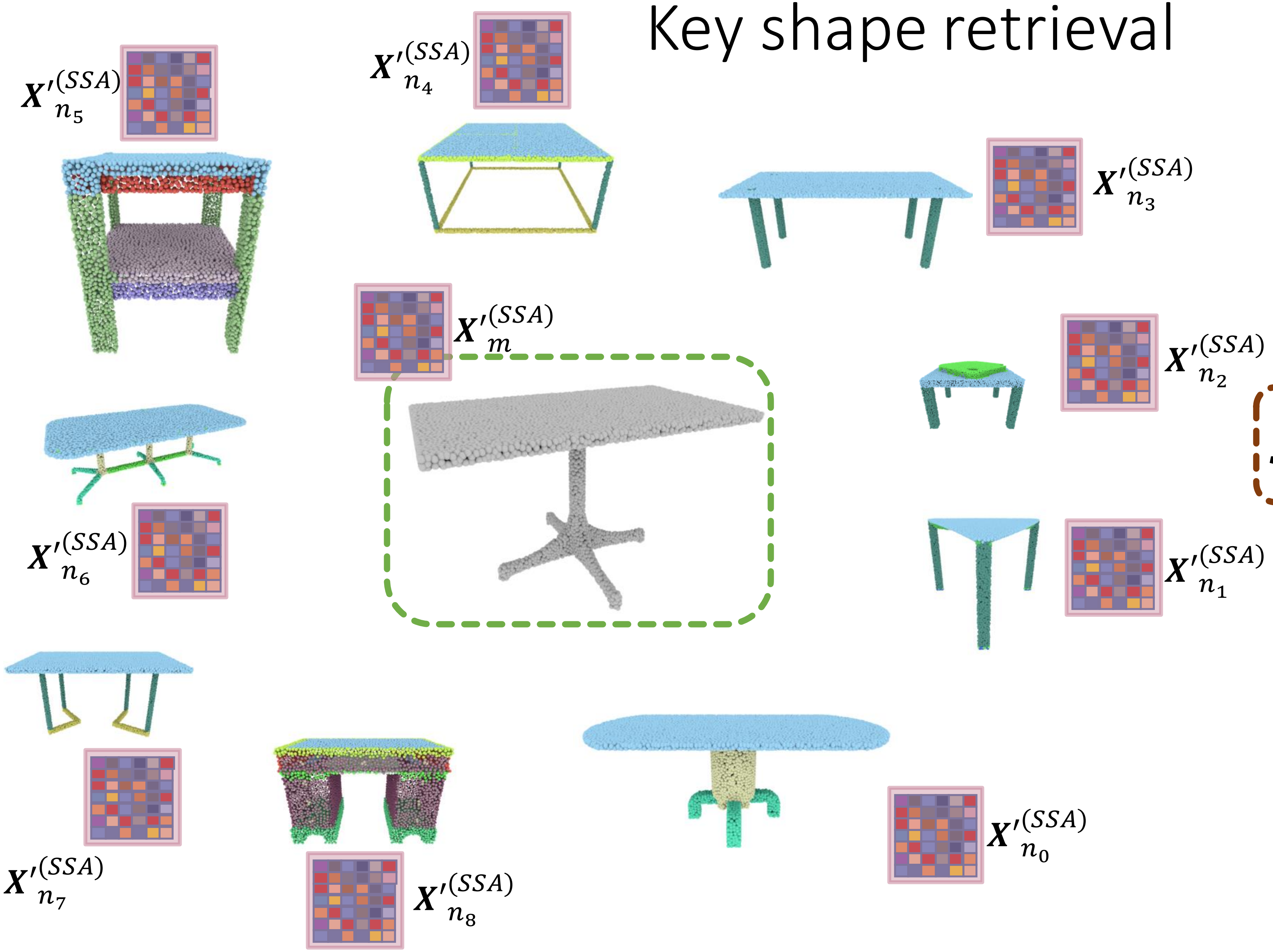
Key shape retrieval



Key shape retrieval



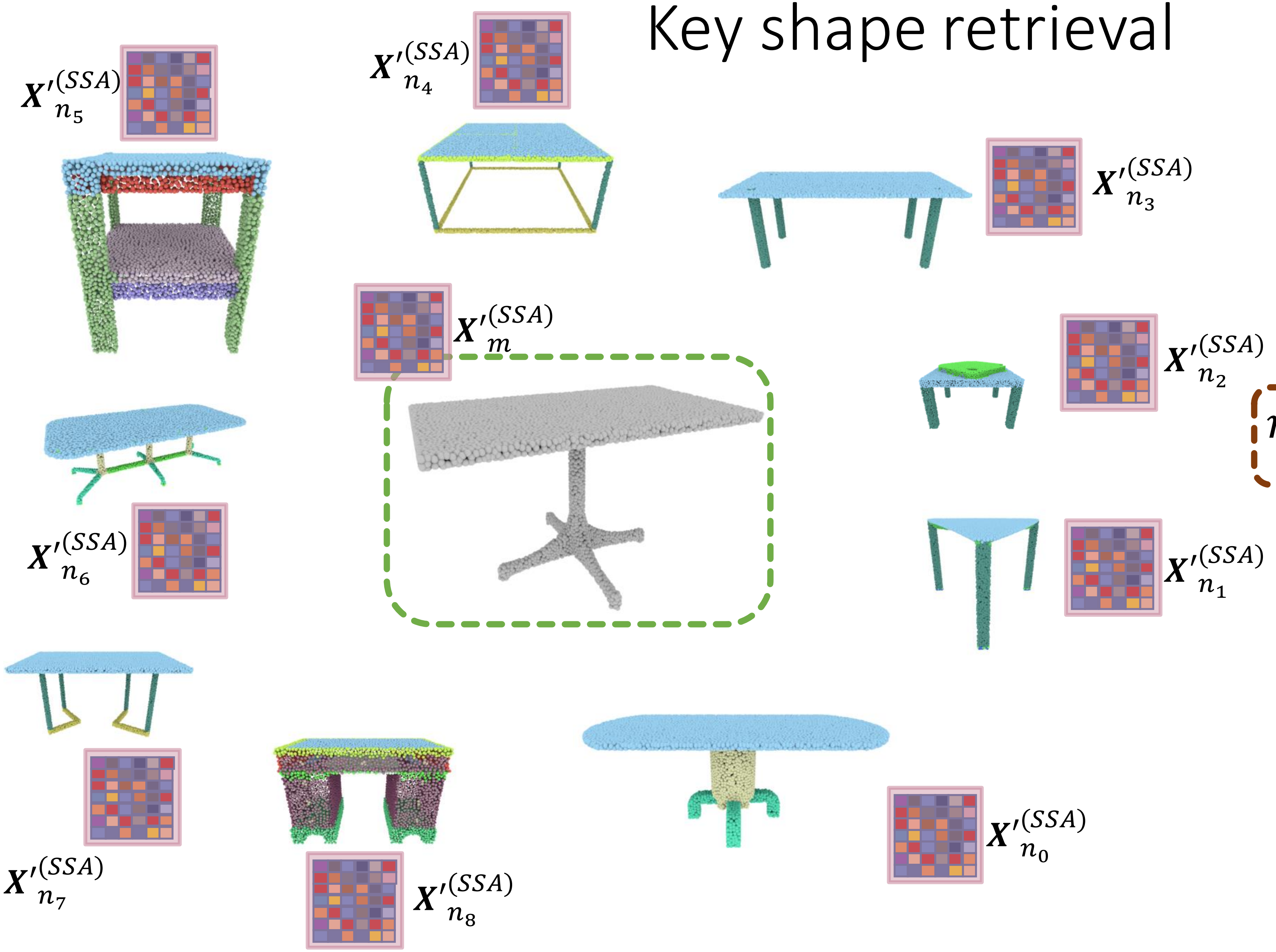
Key shape retrieval



Cosine similarity

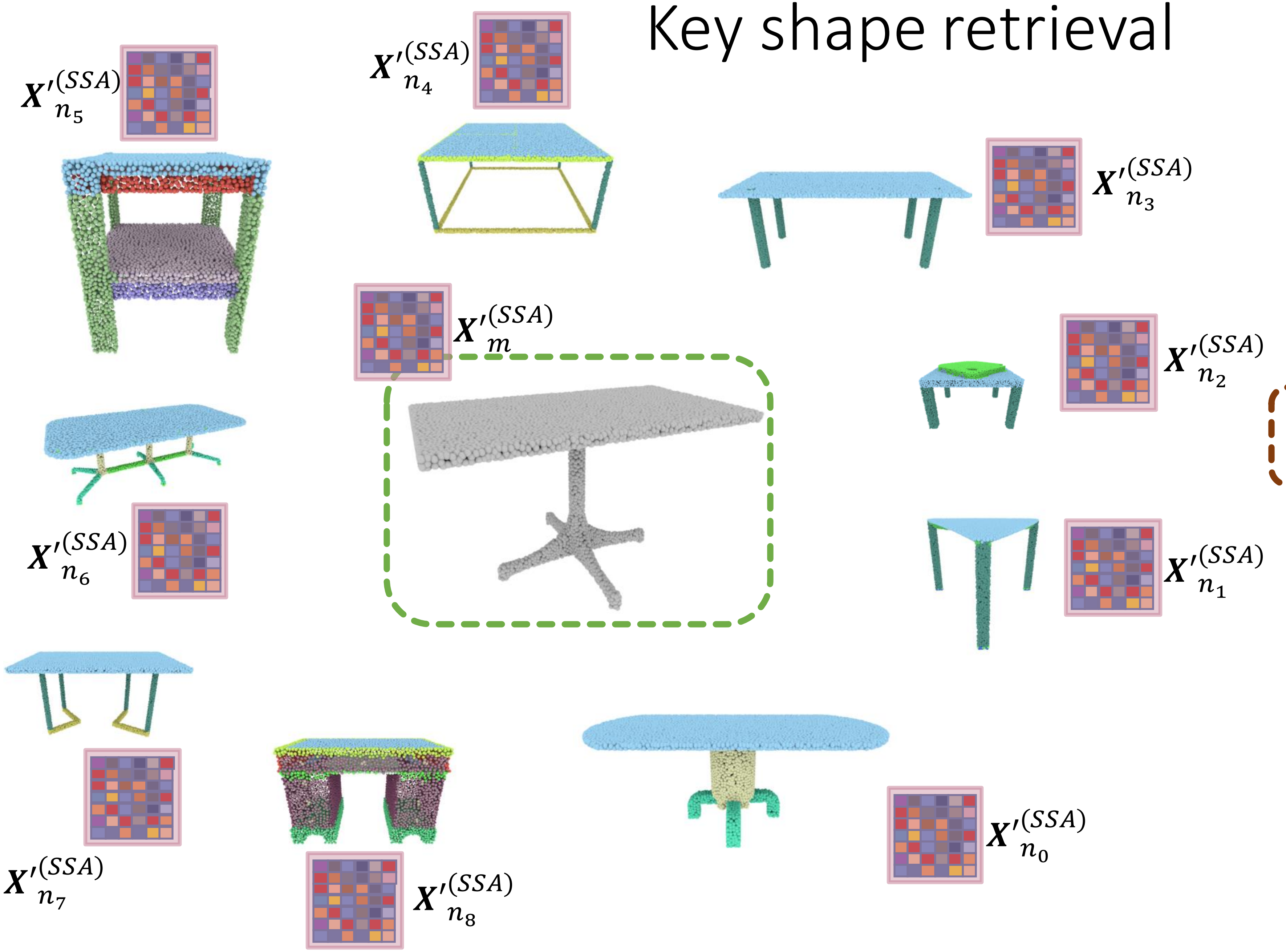
$$\boxed{S_{m,n_k} = X'^{(SSA)}_{m} \cdot \left(X'^{(SSA)}_{n_k}\right)^{\intercal}}$$

Key shape retrieval



$$\boxed{r_i(m, n_k) = \max_j \boldsymbol{S}_{m,n_k}[i, j]}$$

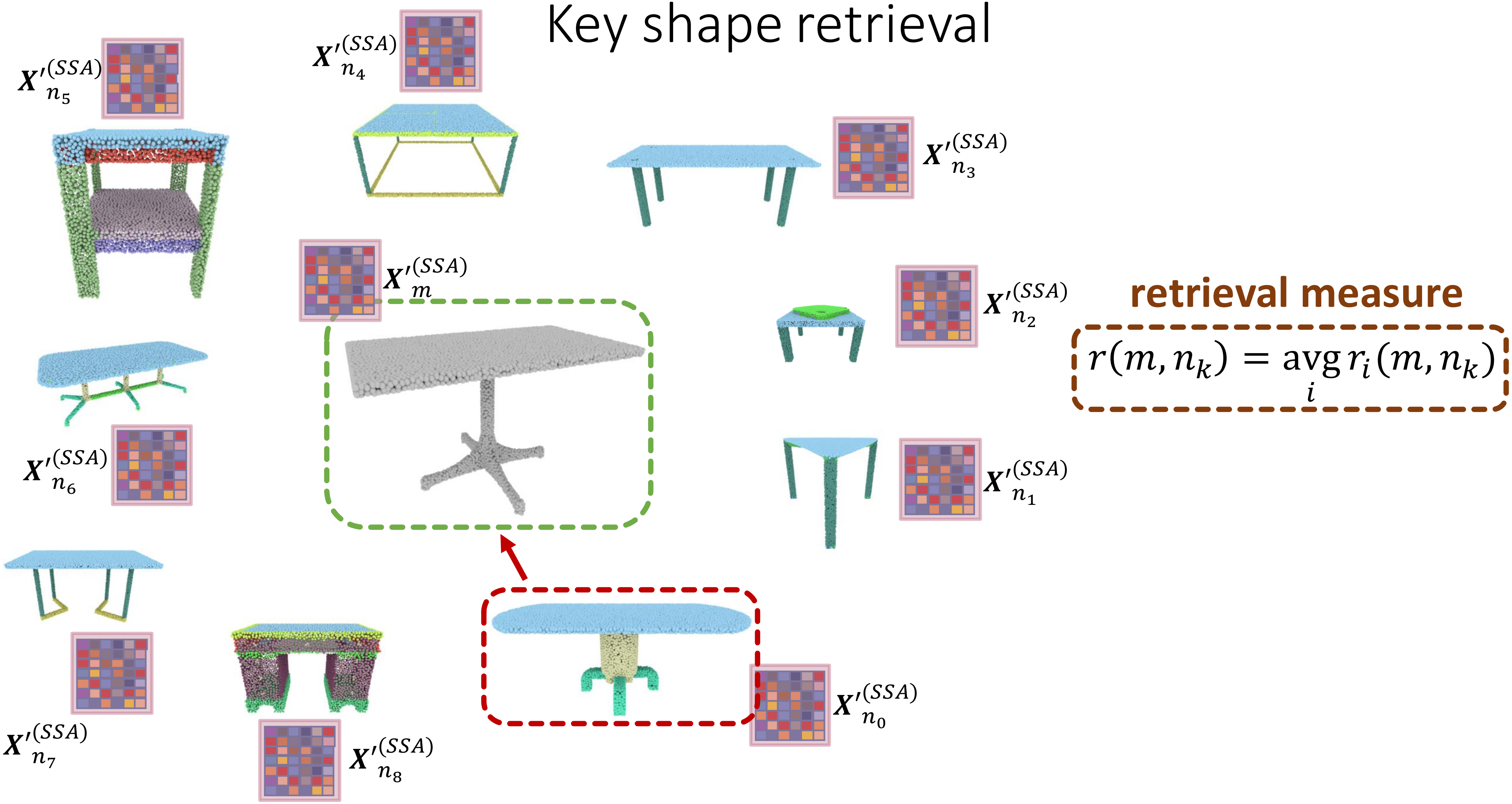
Key shape retrieval



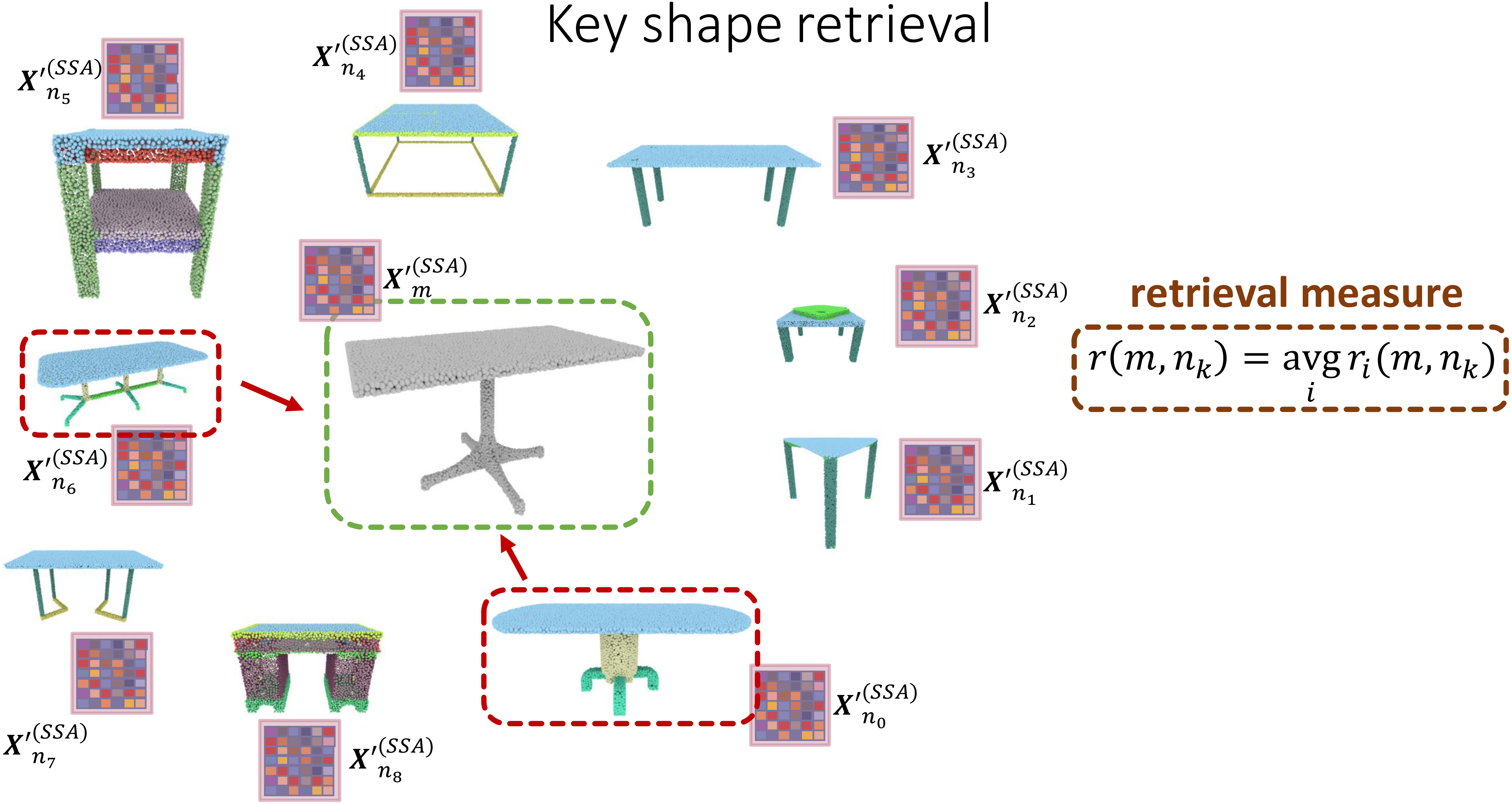
retrieval measure

$$\boxed{r(m, n_k) = \operatorname*{avg}_i r_i(m, n_k)}$$

Key shape retrieval

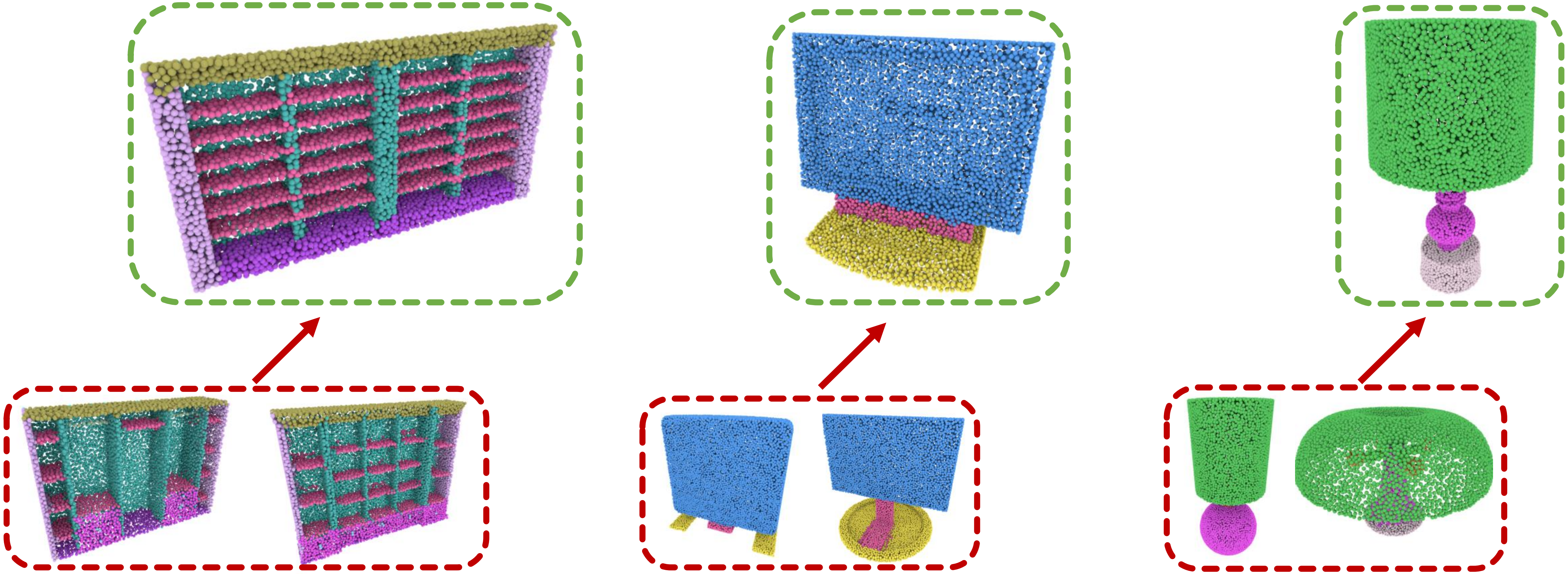


Key shape retrieval



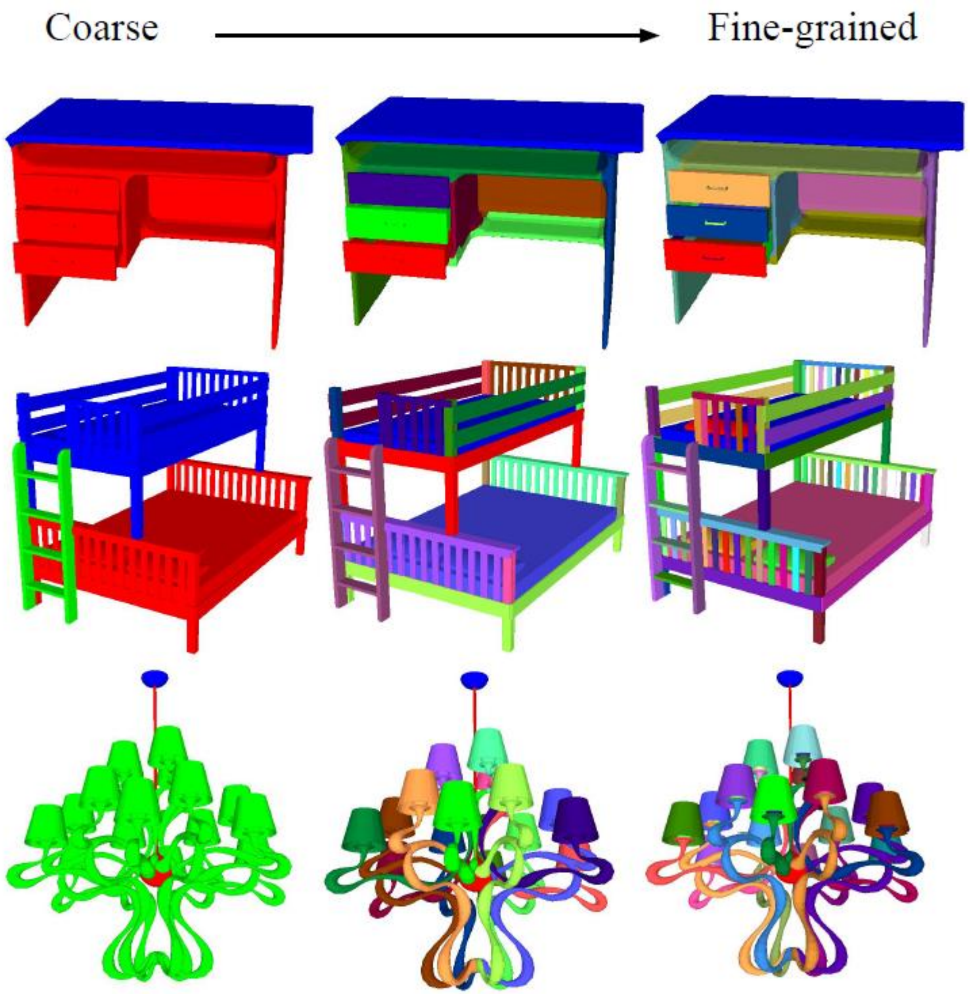
Key shape retrieval: Examples

query shapes



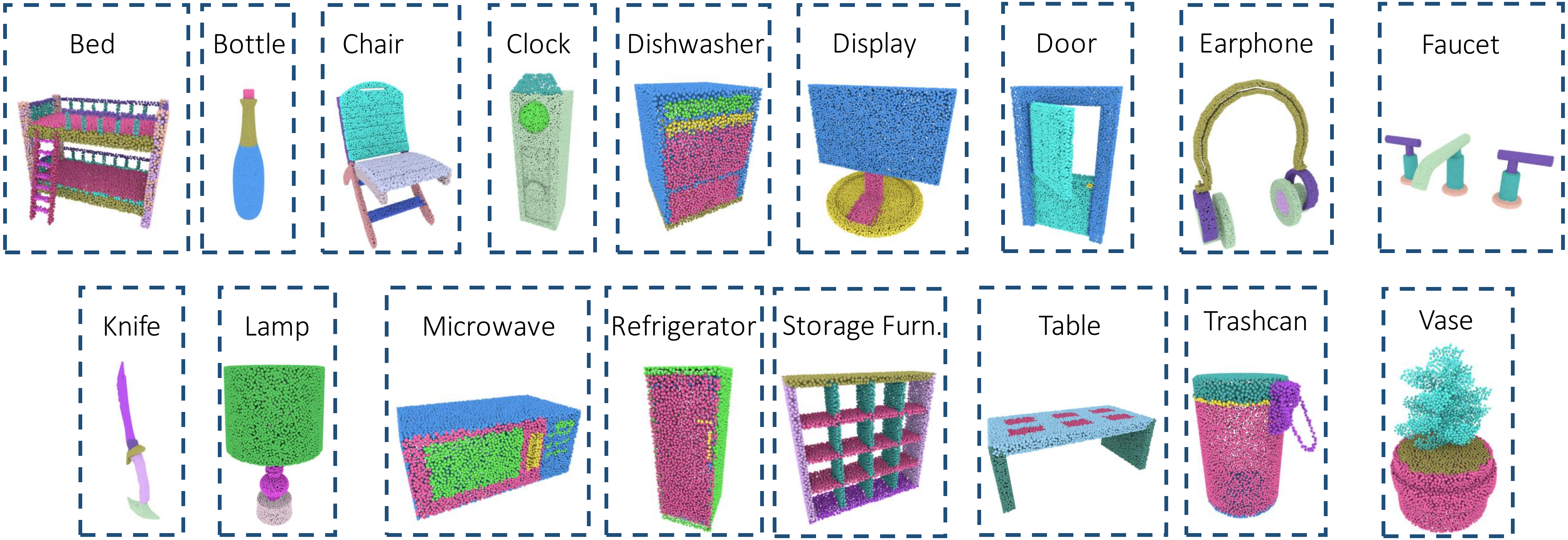
key shapes

PartNet dataset



[Mo et al. 2019]

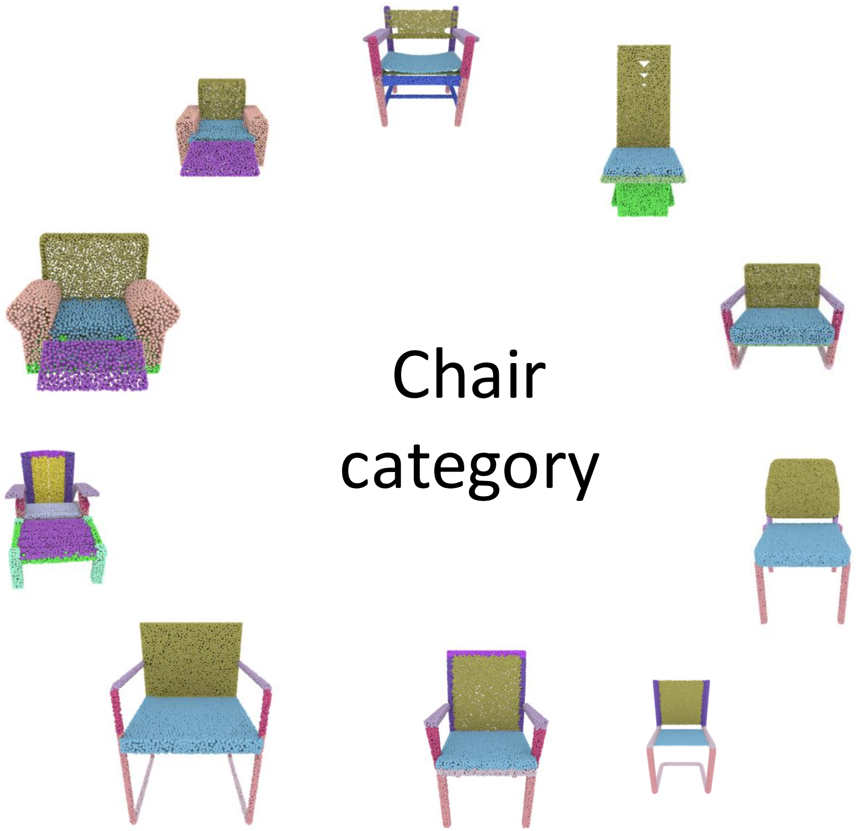
PartNet dataset



[Mo et al. 2019]

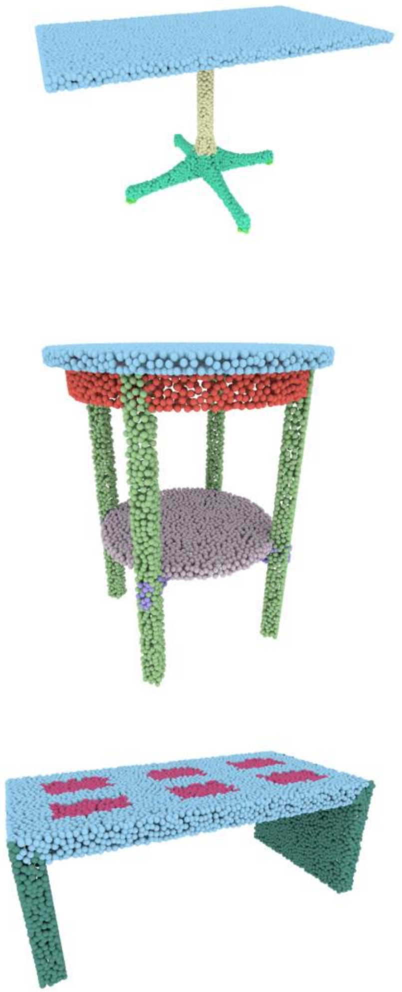
Examples of shape collections

5,707 training shapes



4,489 training shapes

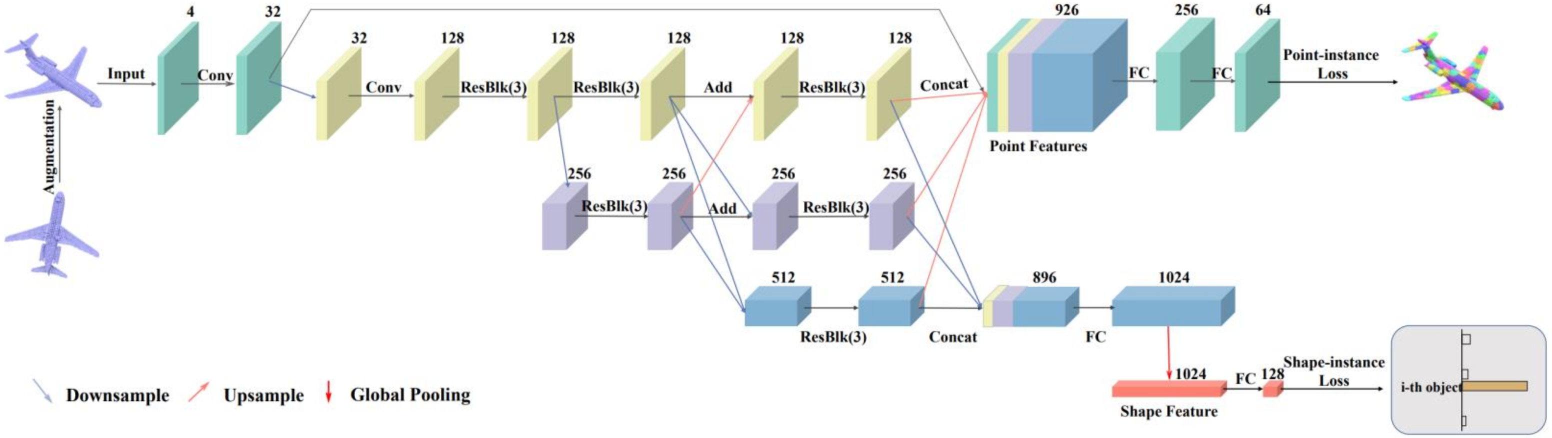
Training details: Loss


training data

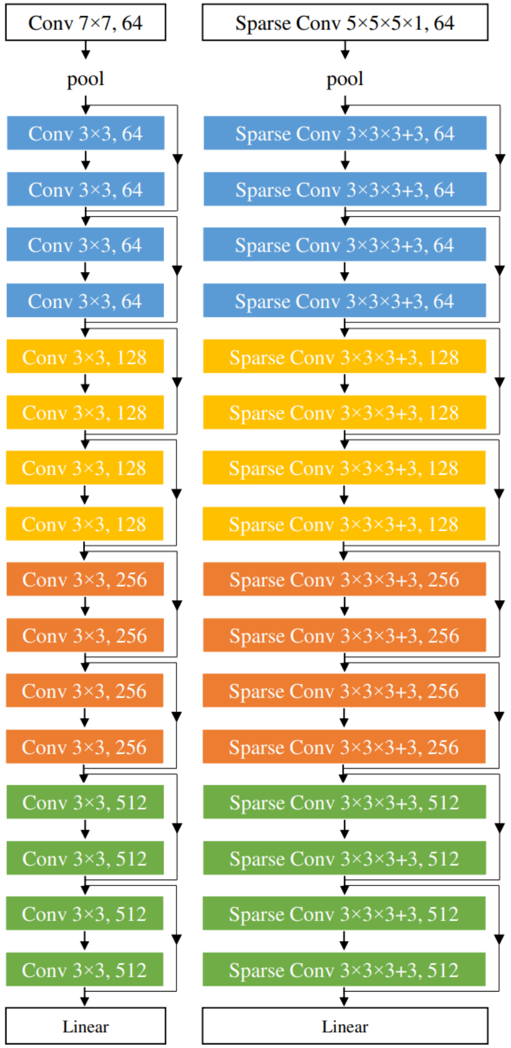
$$L_{CE} = -\sum_{\boldsymbol{p}_i \in \mathcal{S}_k} \widehat{\boldsymbol{q}}_i \log \boldsymbol{q}_i$$

$\mathcal{S}_k$: shape $k = \{\boldsymbol{p}_i\}_{i=1}^{P_k}$
$\widehat{\boldsymbol{q}}_i$: ground-truth one-hot label vector for point $\boldsymbol{p}_i$
$\boldsymbol{q}_i$: predicted label probabilities for point $\boldsymbol{p}_i$

Training details: Backbones

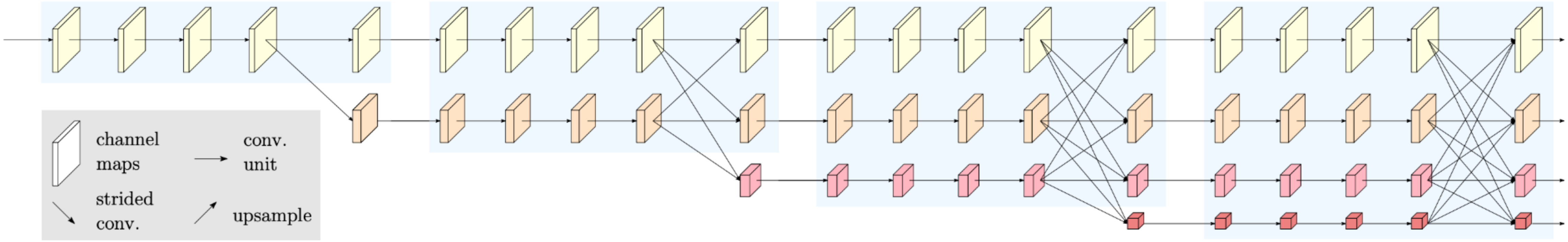


MID-FC [Wang et al. 2021]



MinkowskiNet [Choy et al. 2019]

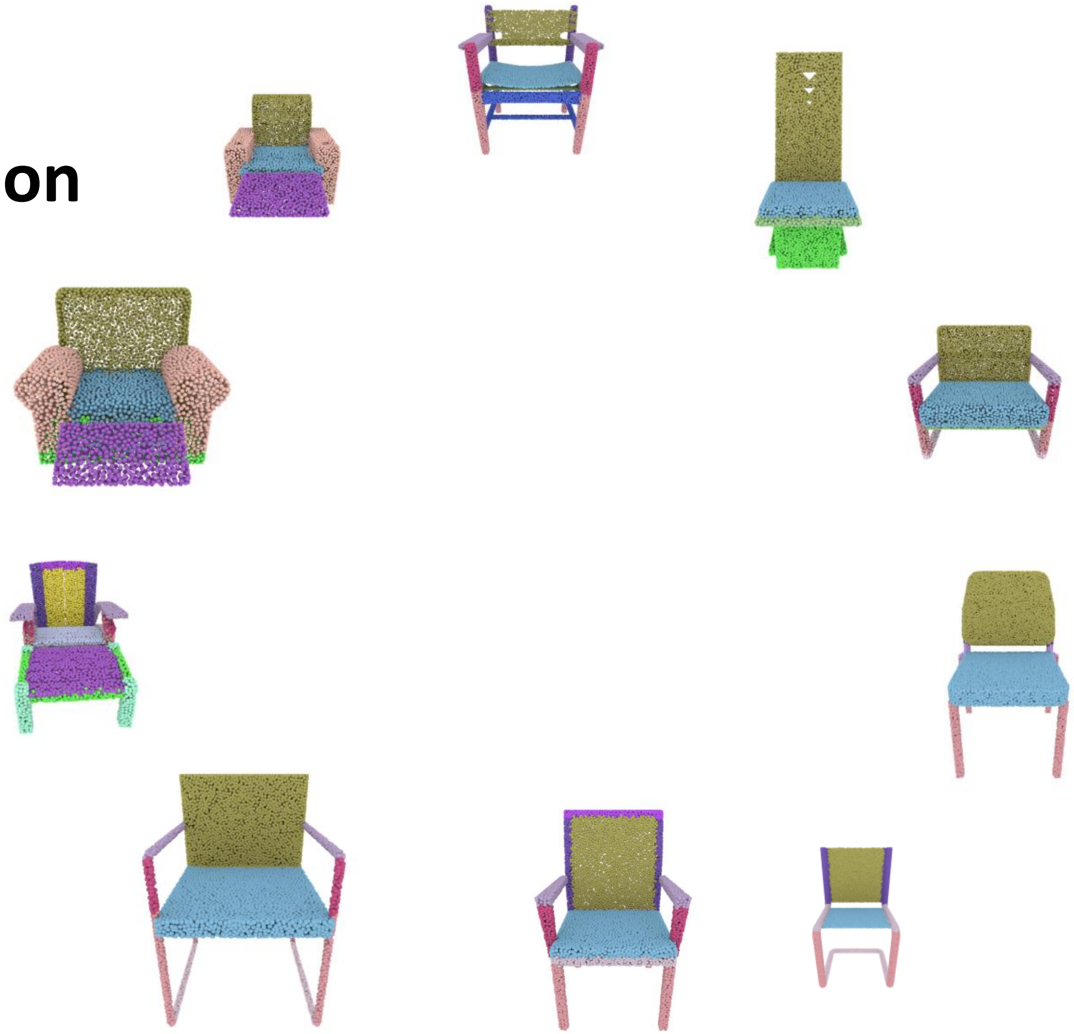
Training details: Backbones



HRNet [Wang et al. 2021]

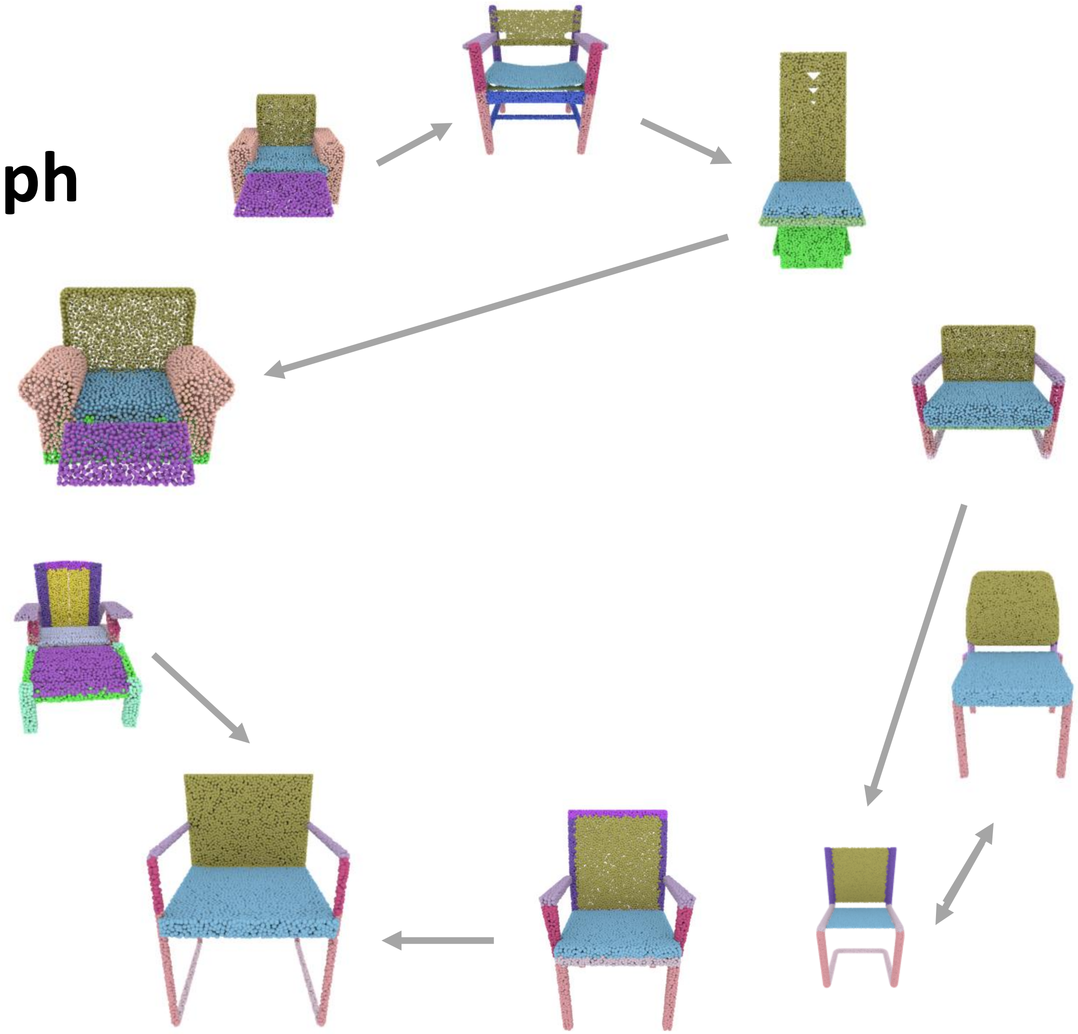
Training details: Collection graph

Shape Collection



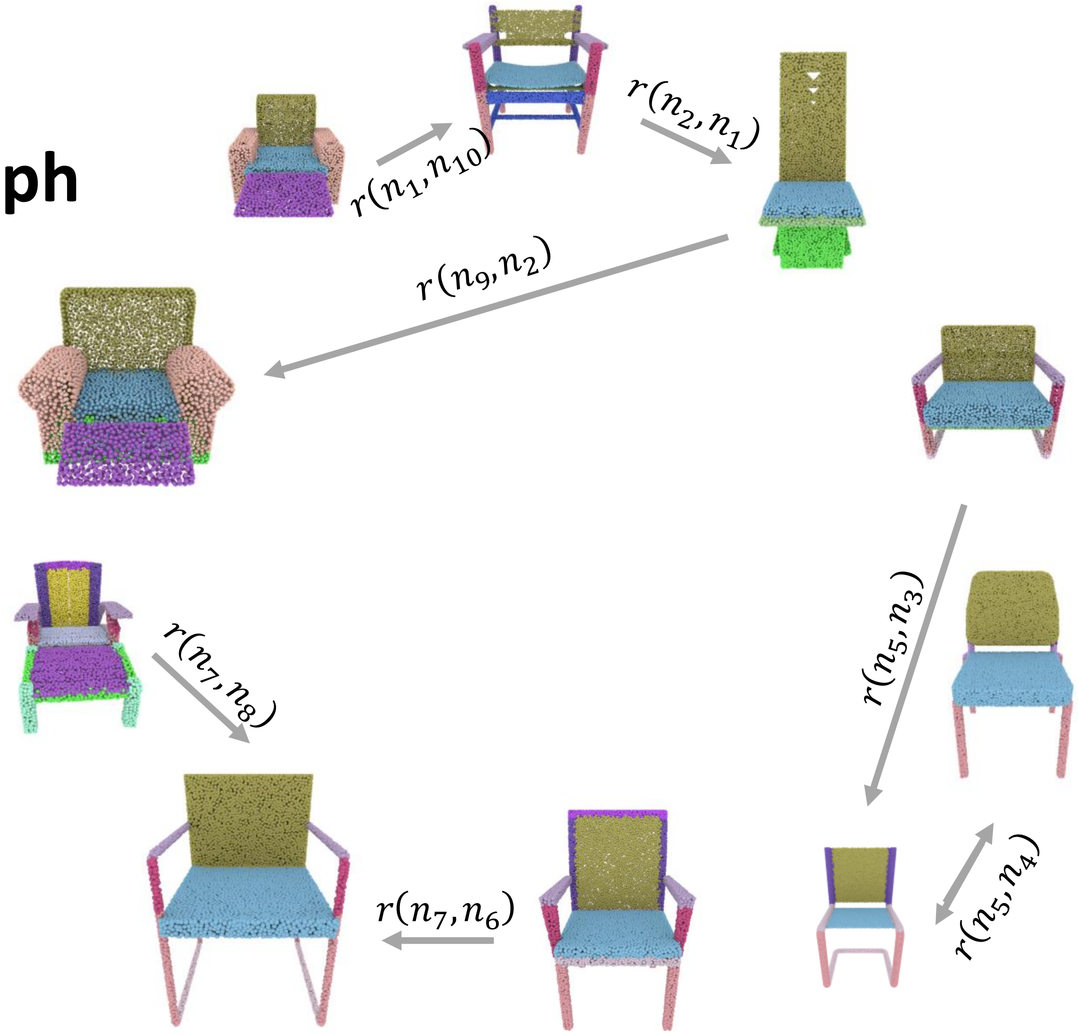
Training details: Collection graph

Collection graph



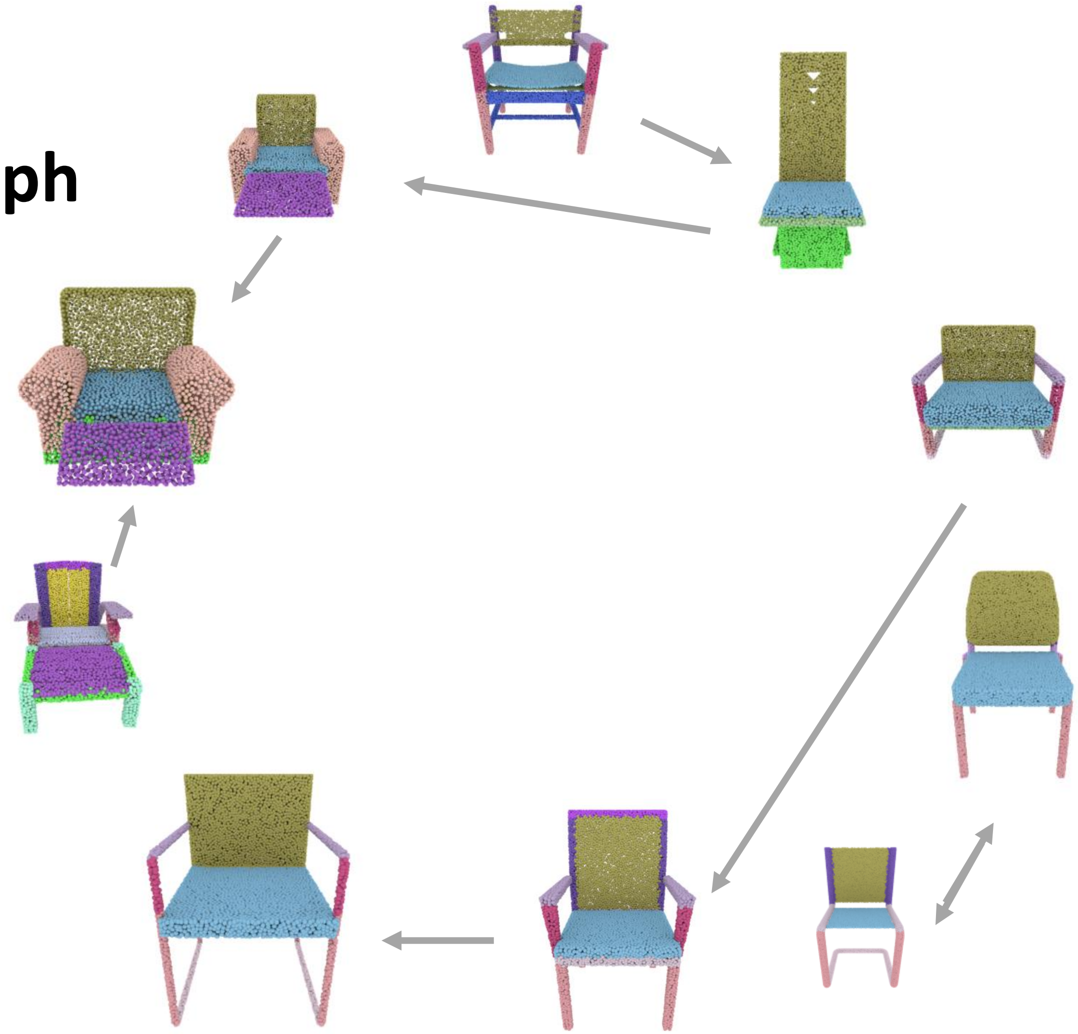
Training details: Collection graph

Collection graph



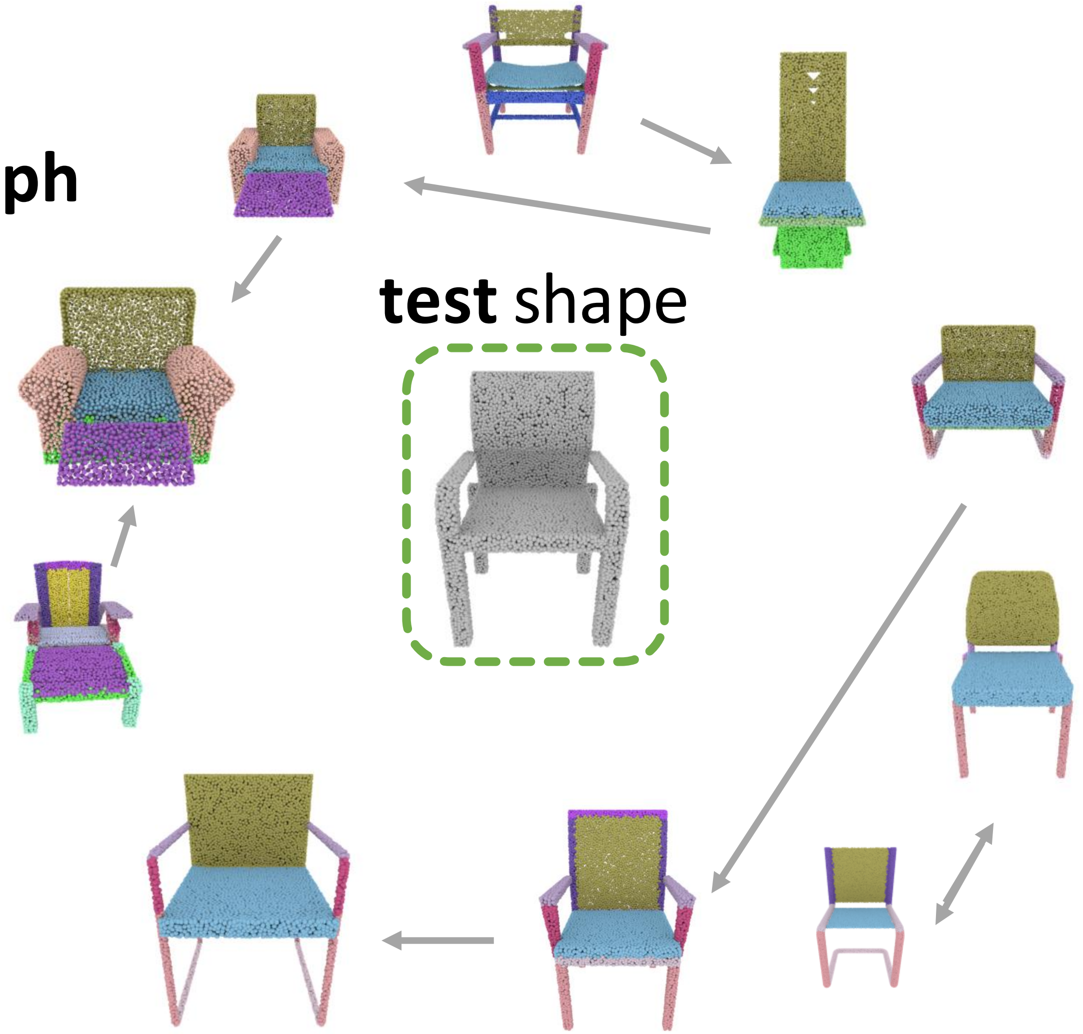
Training details: Collection graph

Collection graph

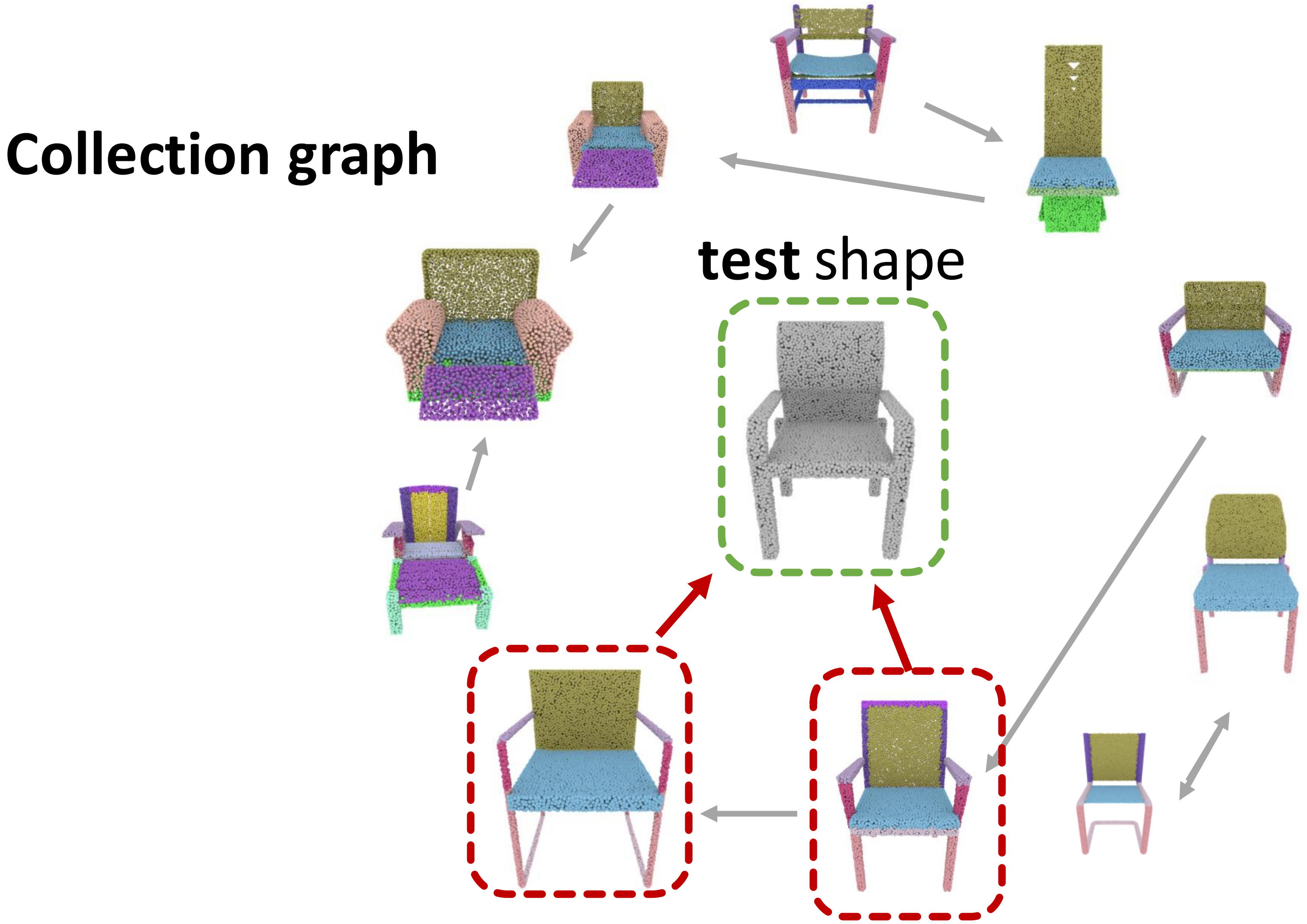


Inference: Collection graph

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

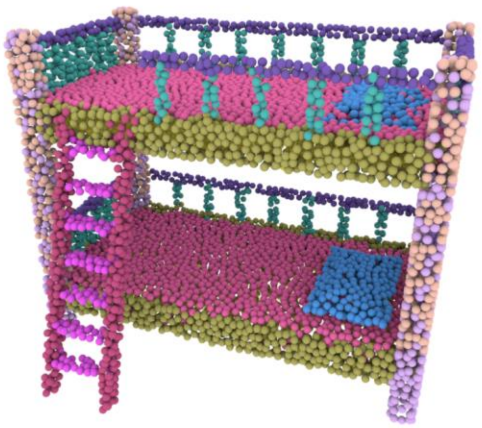
Results: MinkowskiNet variants

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

+1.2%

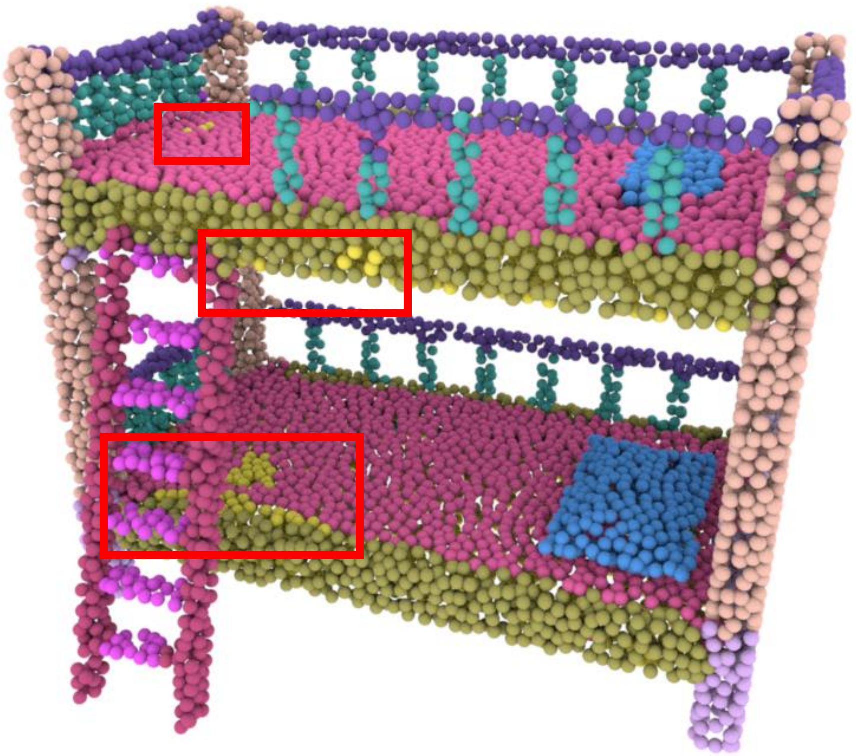
Results: MinkowskiNet variants

Ground truth

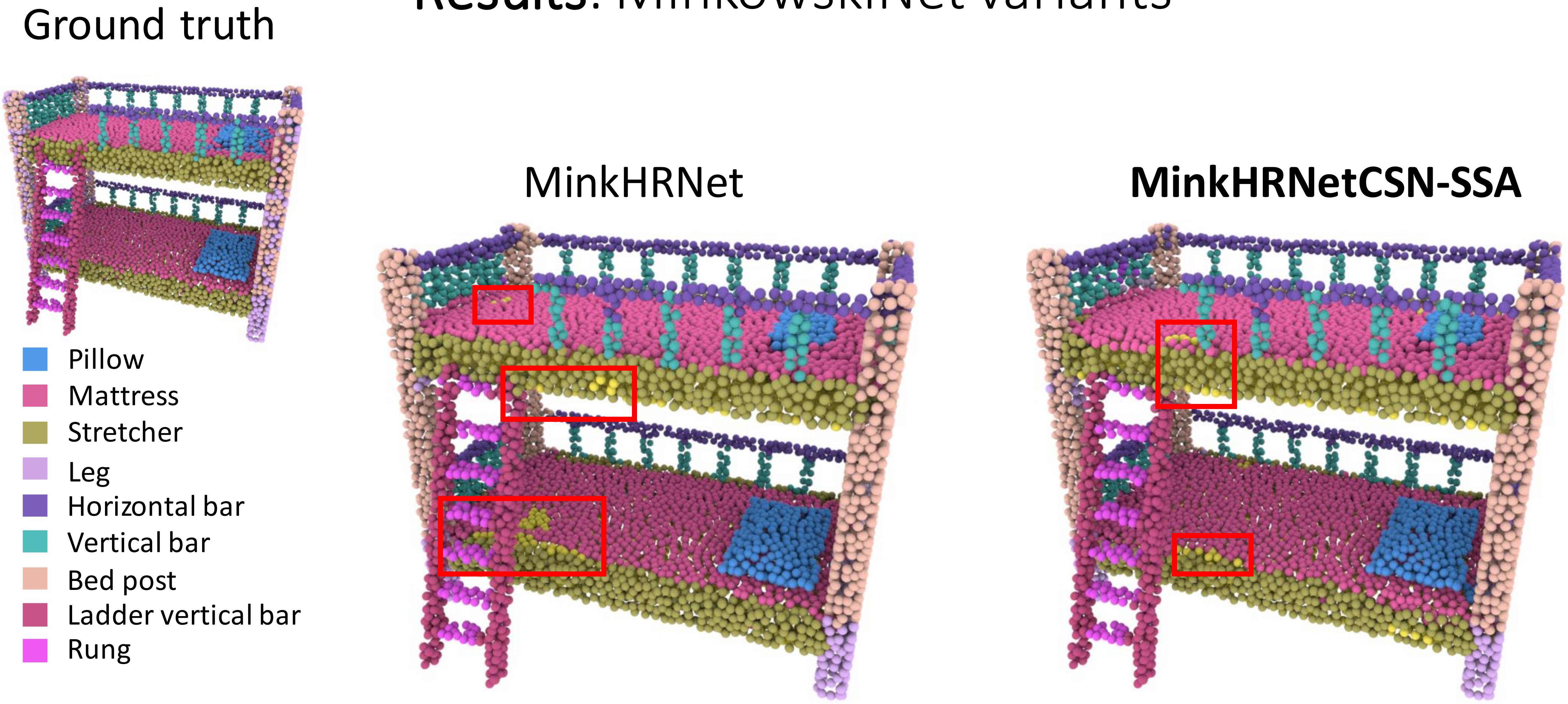


Pillow
Mattress
Stretcher
Leg
Horizontal bar
Vertical bar
Bed post
Ladder vertical bar
Rung

MinkHRNet

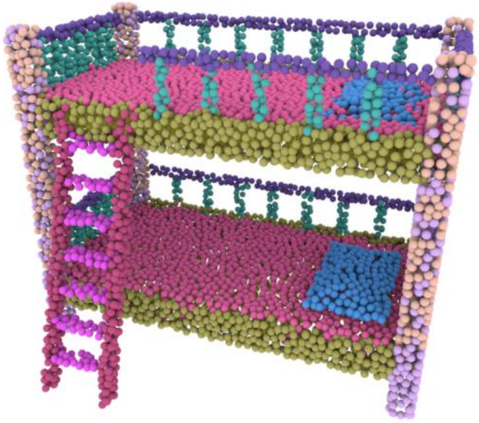


Results: MinkowskiNet variants



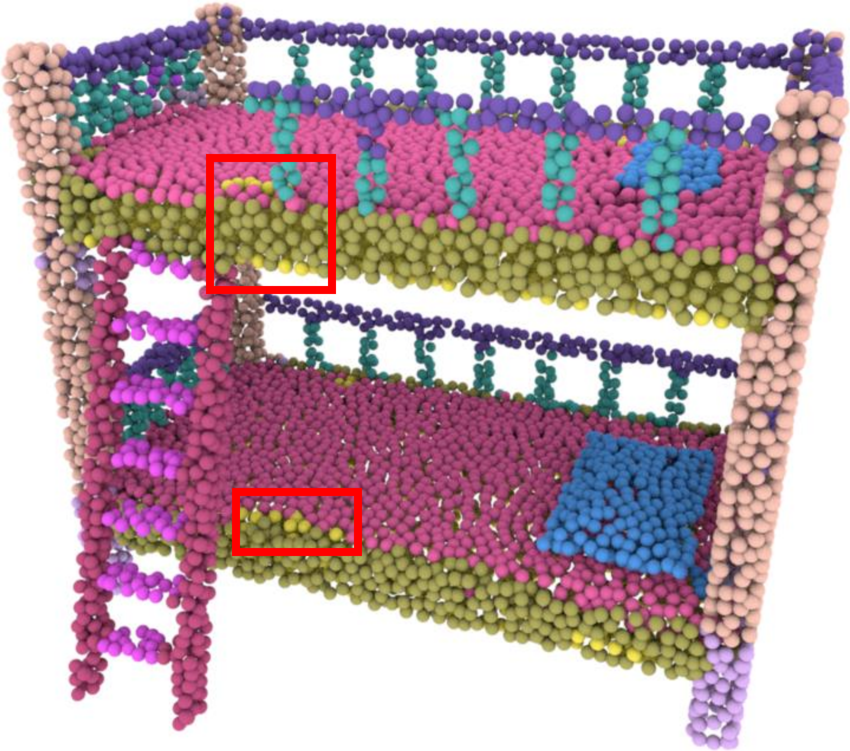
Results: MinkowskiNet variants

Ground truth

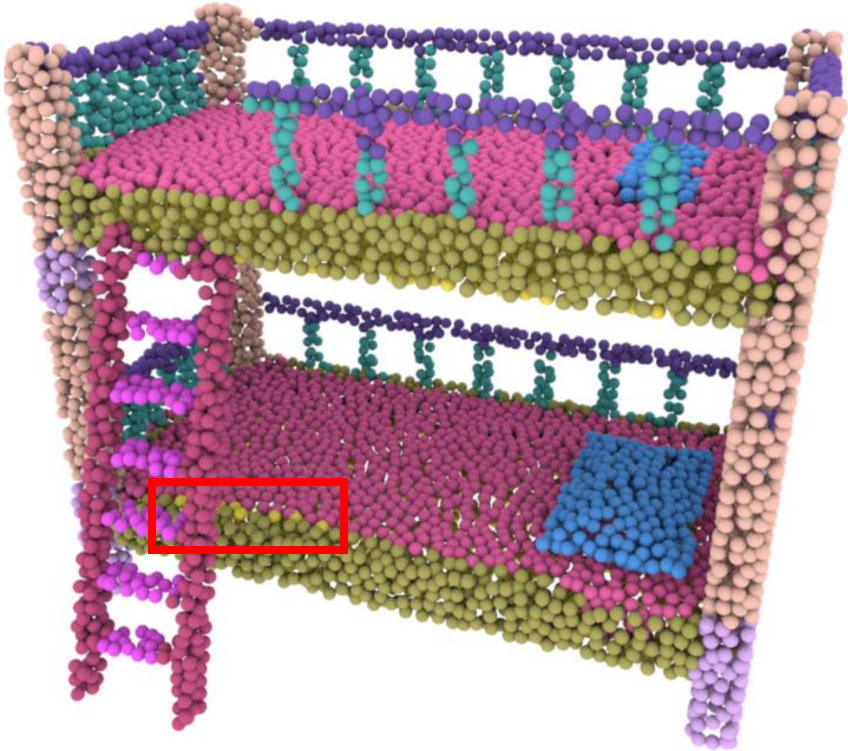


- Pillow
- Mattress
- Stretcher
- Leg
- Horizontal bar
- Vertical bar
- Bed post
- Ladder vertical bar
- Rung

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%

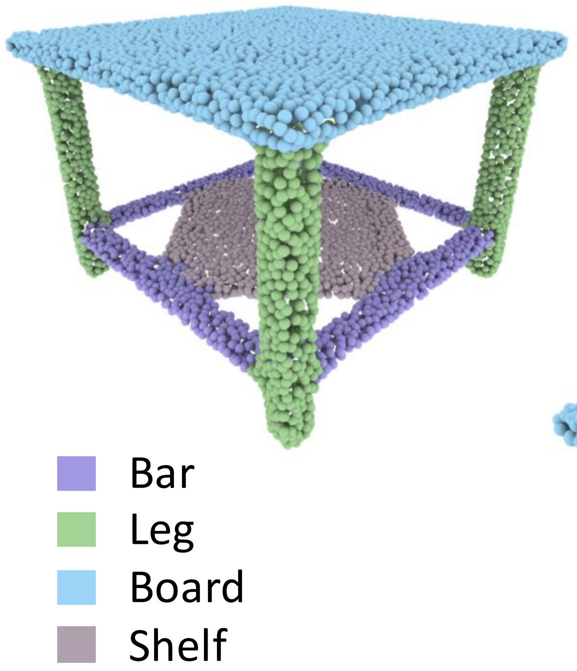
Results: MID-FC variants

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

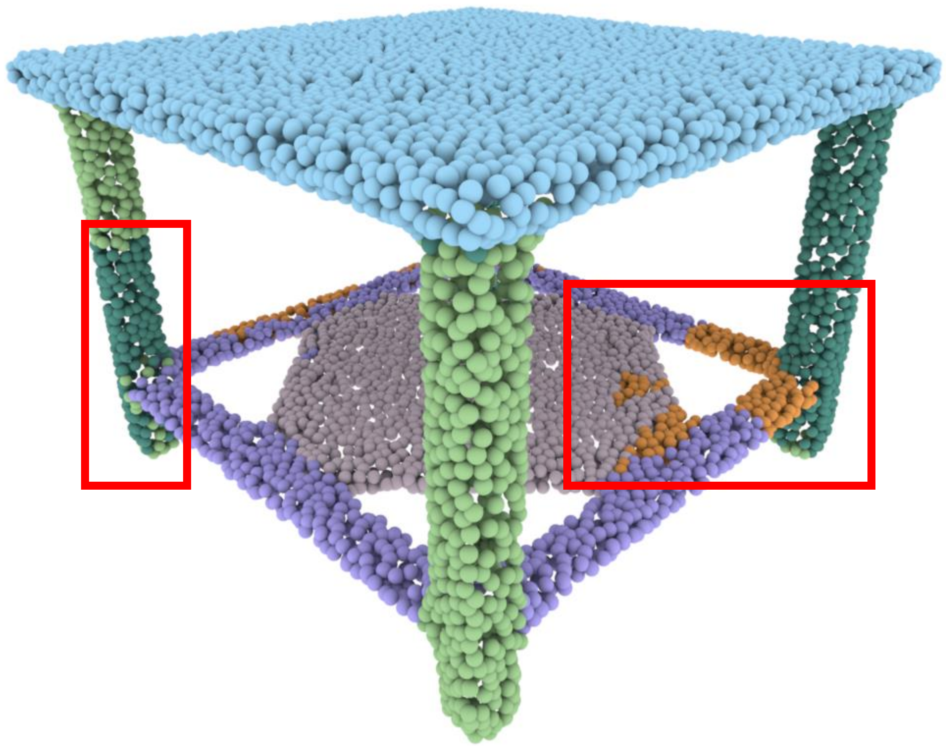
+0.3%

Results: MID-FC variants

Ground truth

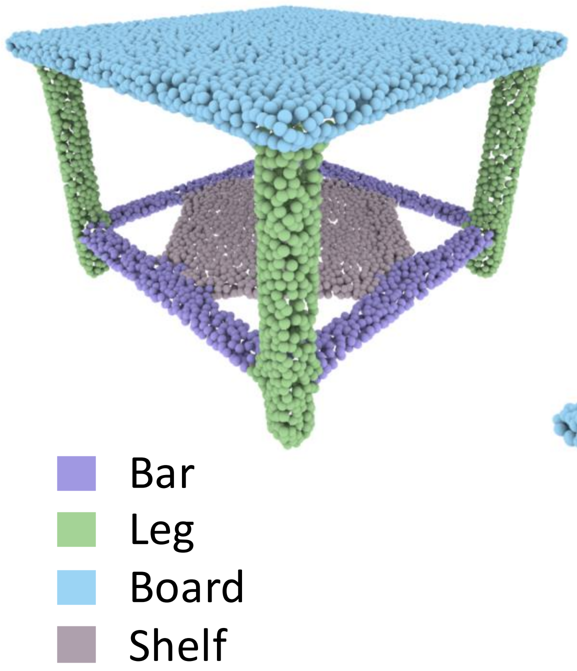


MID-FC

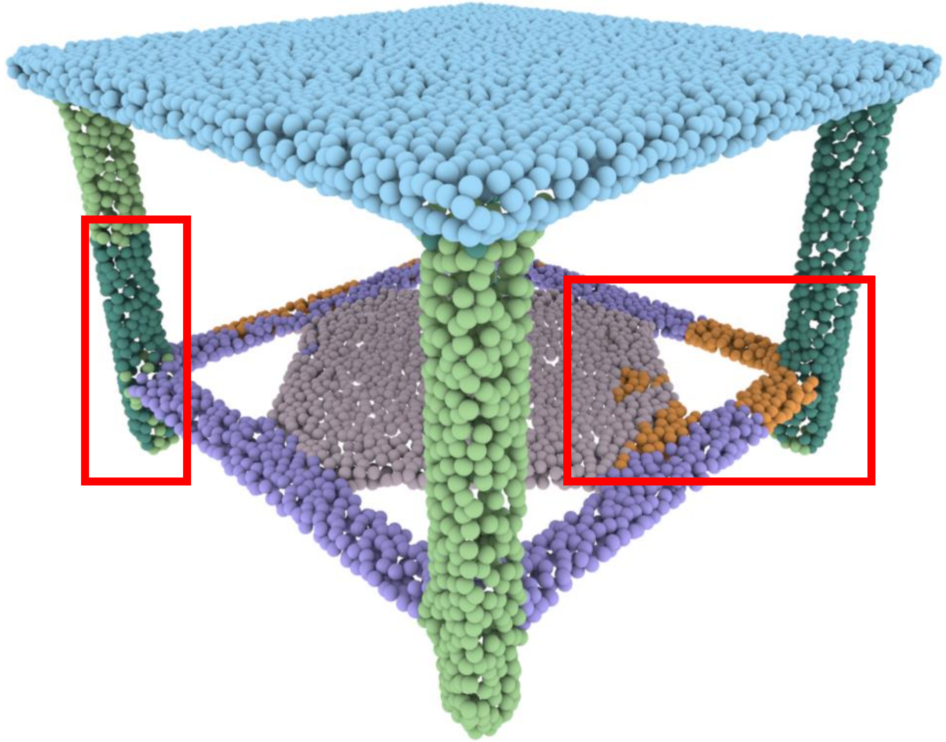


Results: MID-FC variants

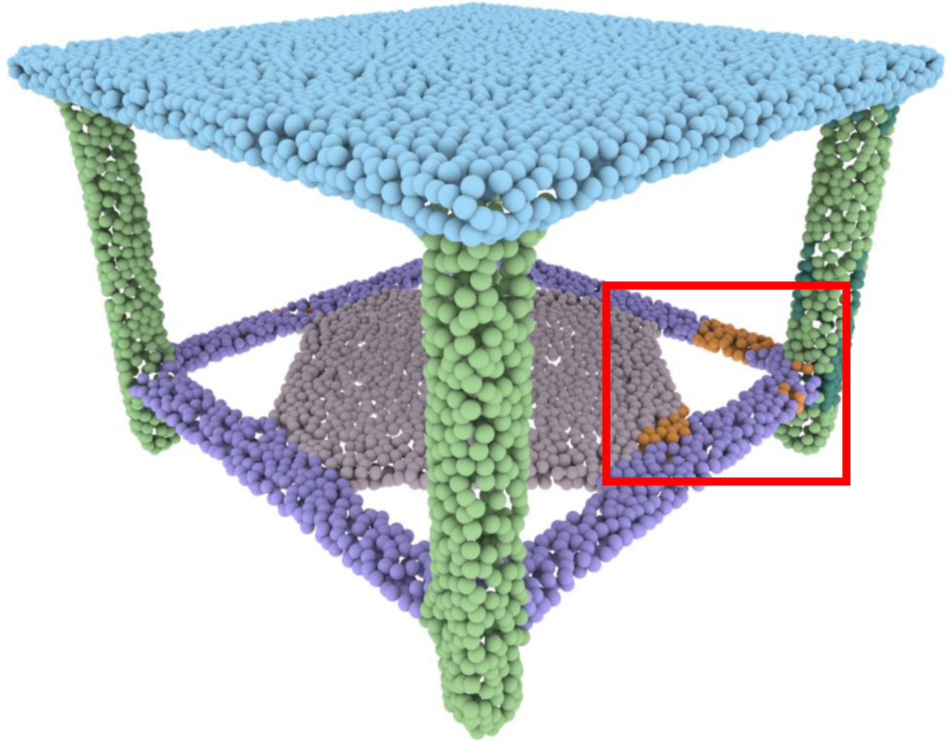
Ground truth



MID-FC

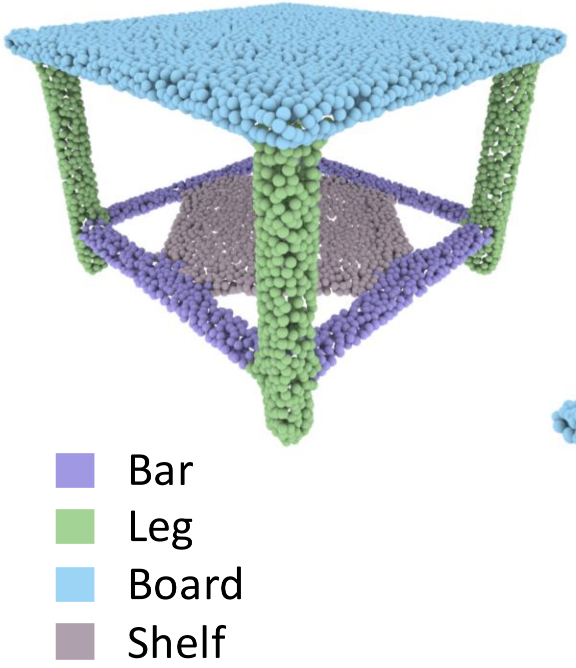


MID-FC-CSN-SSA

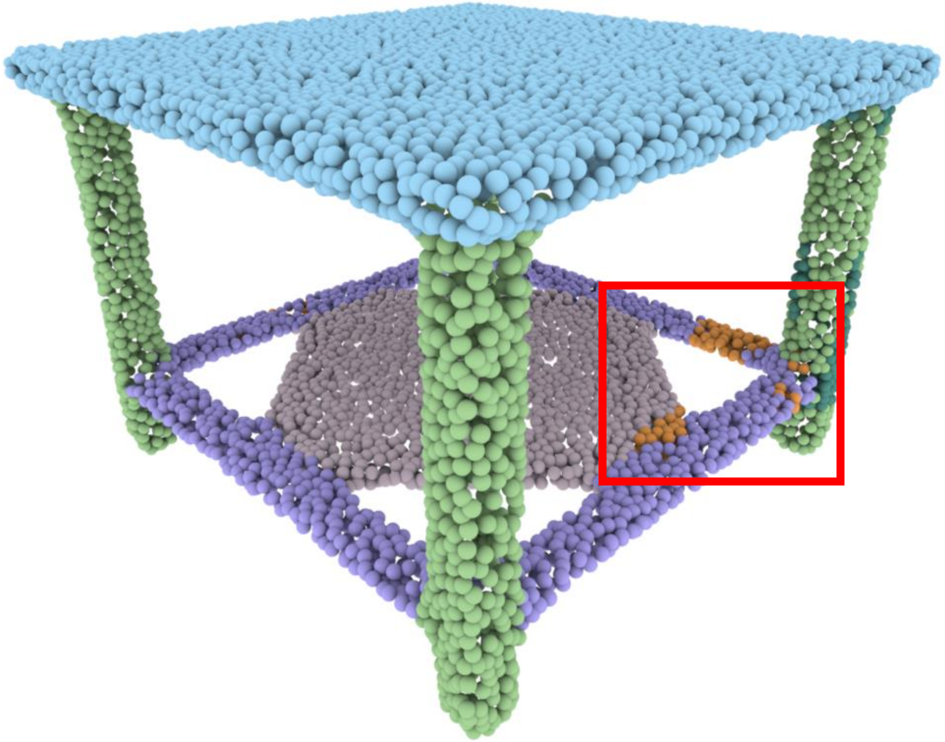


Results: MID-FC variants

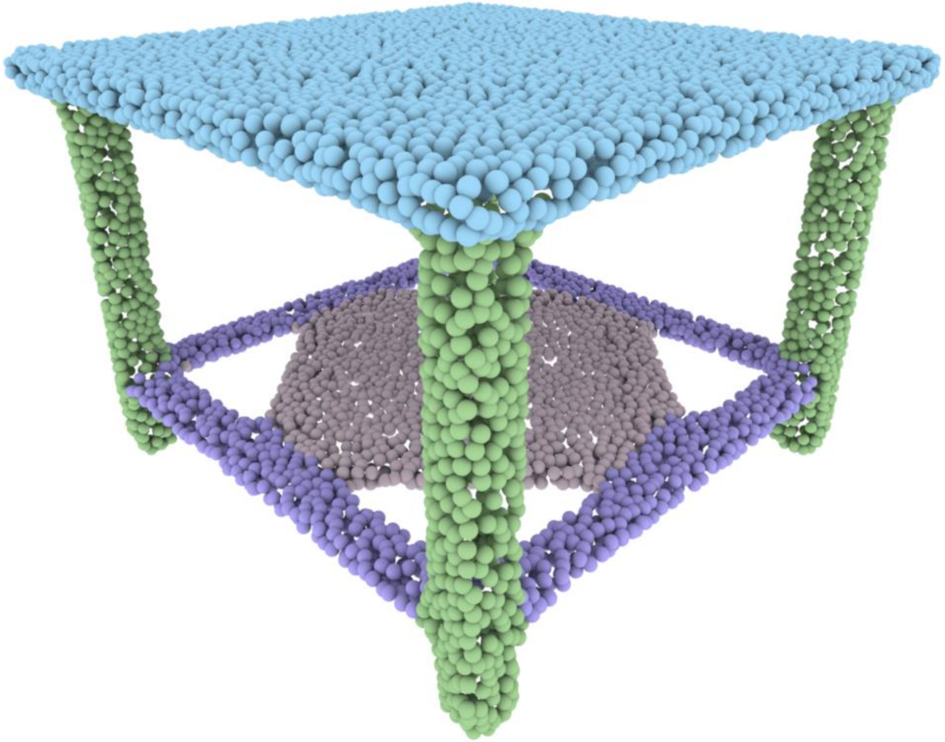
Ground truth



MID-FC-CSN-SSA



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%

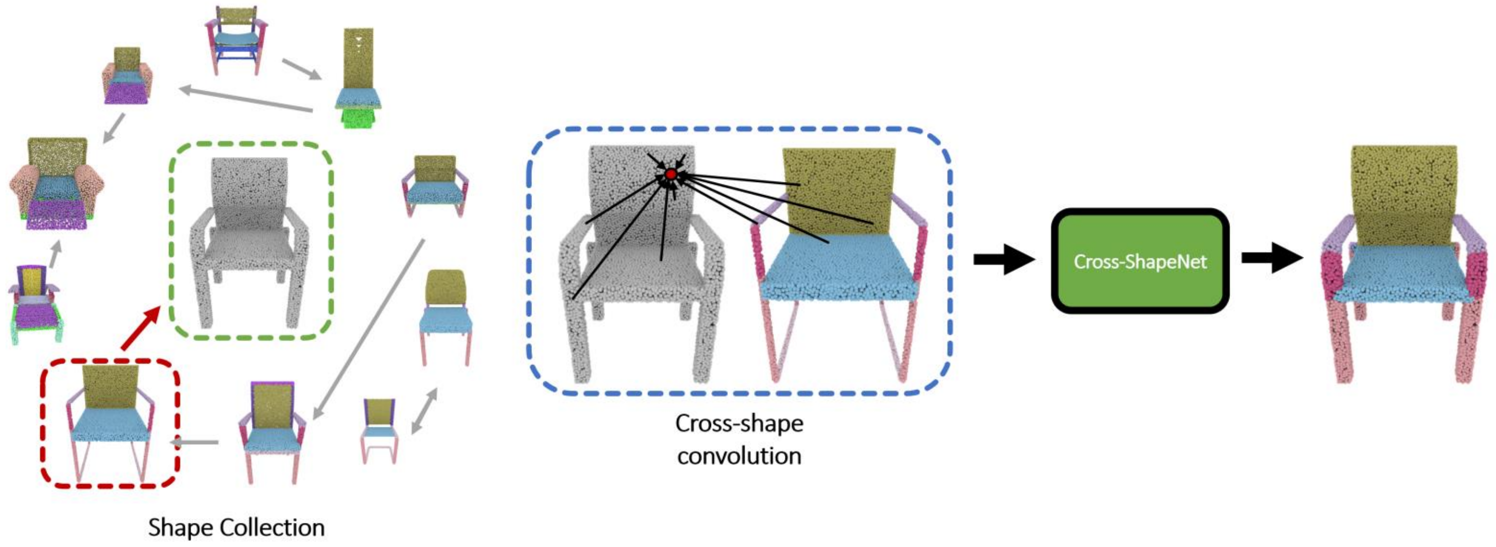
Results: Comparison with other methods

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**

+1.3%

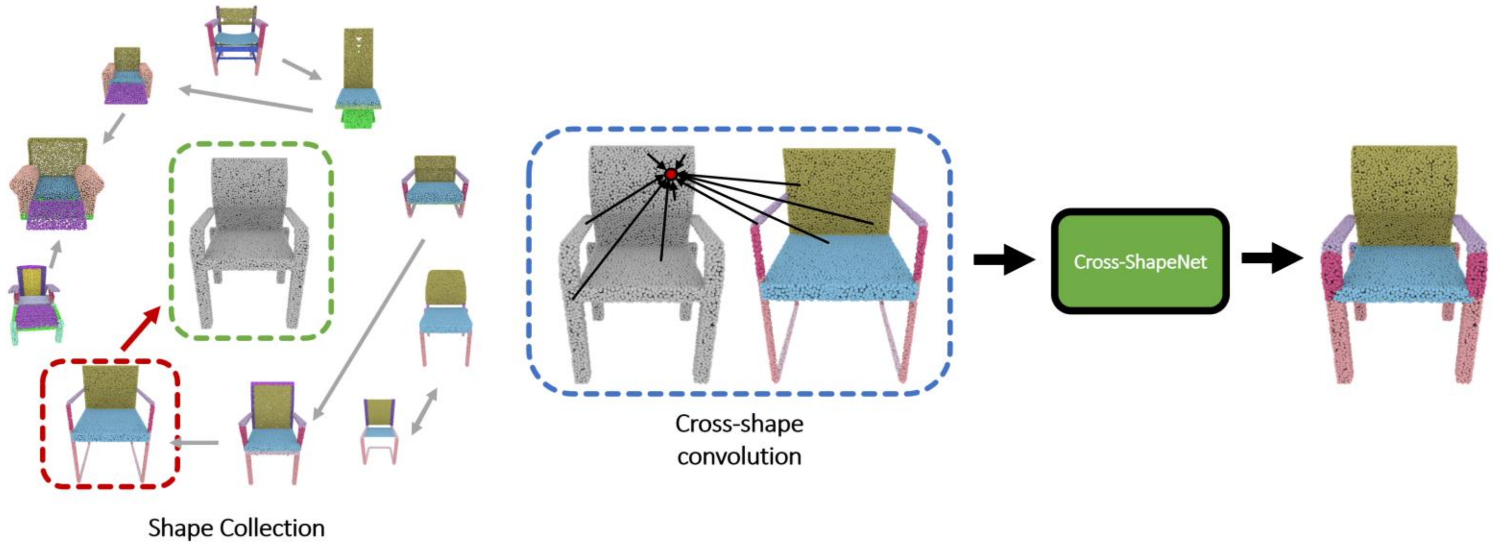
SOTA performance on the PartNet dataset

Summary



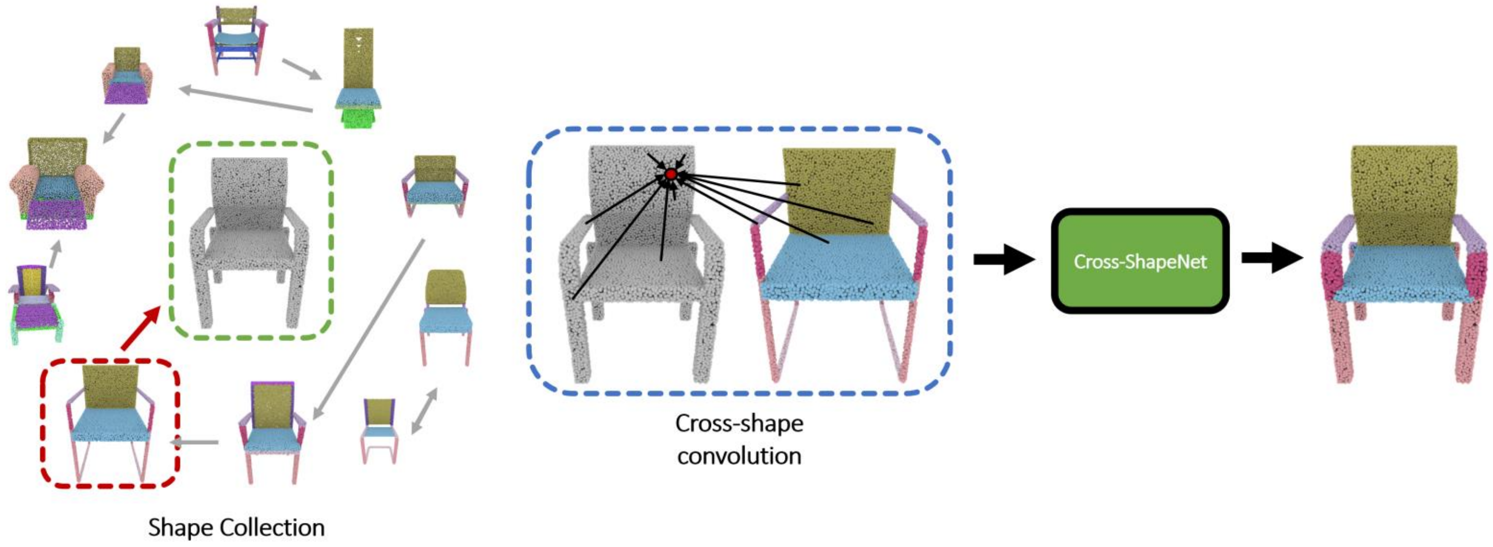
- Enable long range point feature interactions **across shapes**

Summary



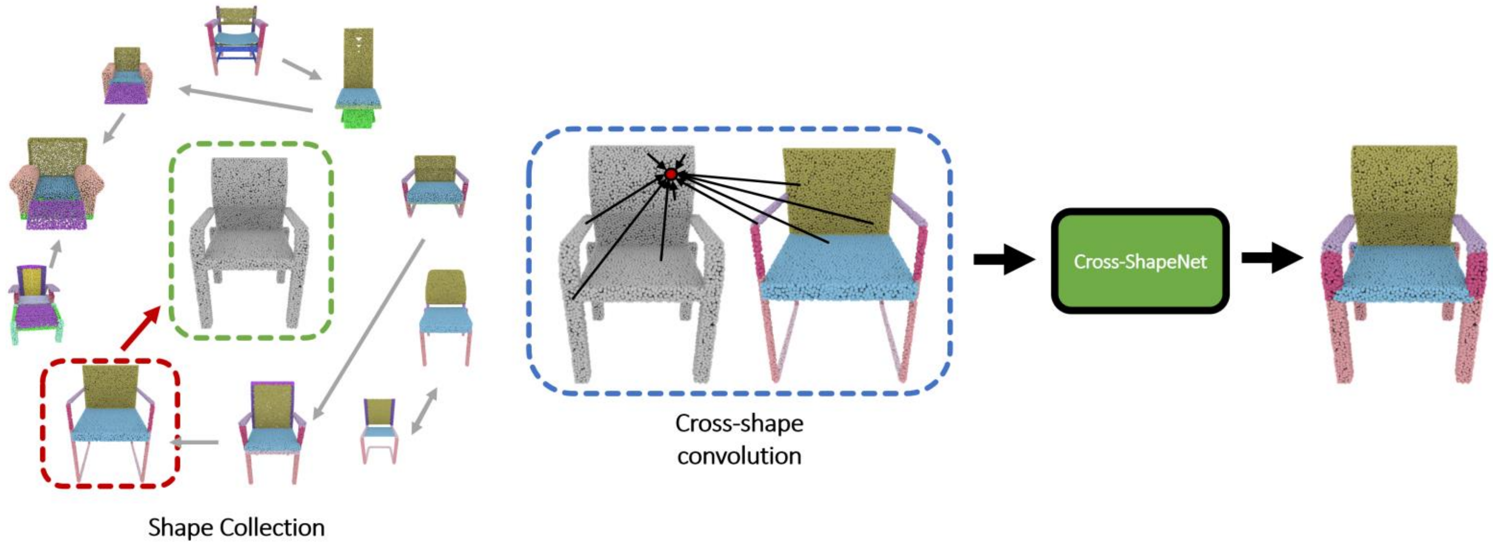
- Enable long range point feature interactions **across shapes**
- Introduce a **novel cross-shape attention** mechanism

Summary



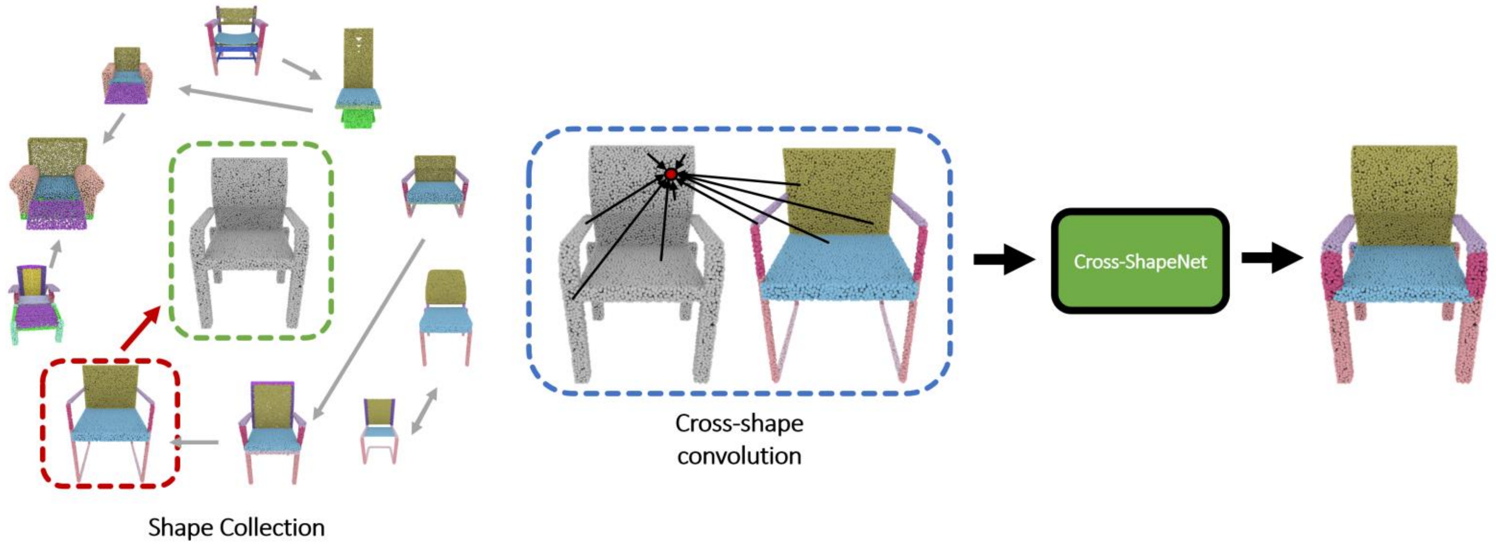
- Enable long range point feature interactions **across shapes**
- Introduce a **novel cross-shape attention** mechanism
- Retrieve **compatible shapes** for cross-shape attention

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

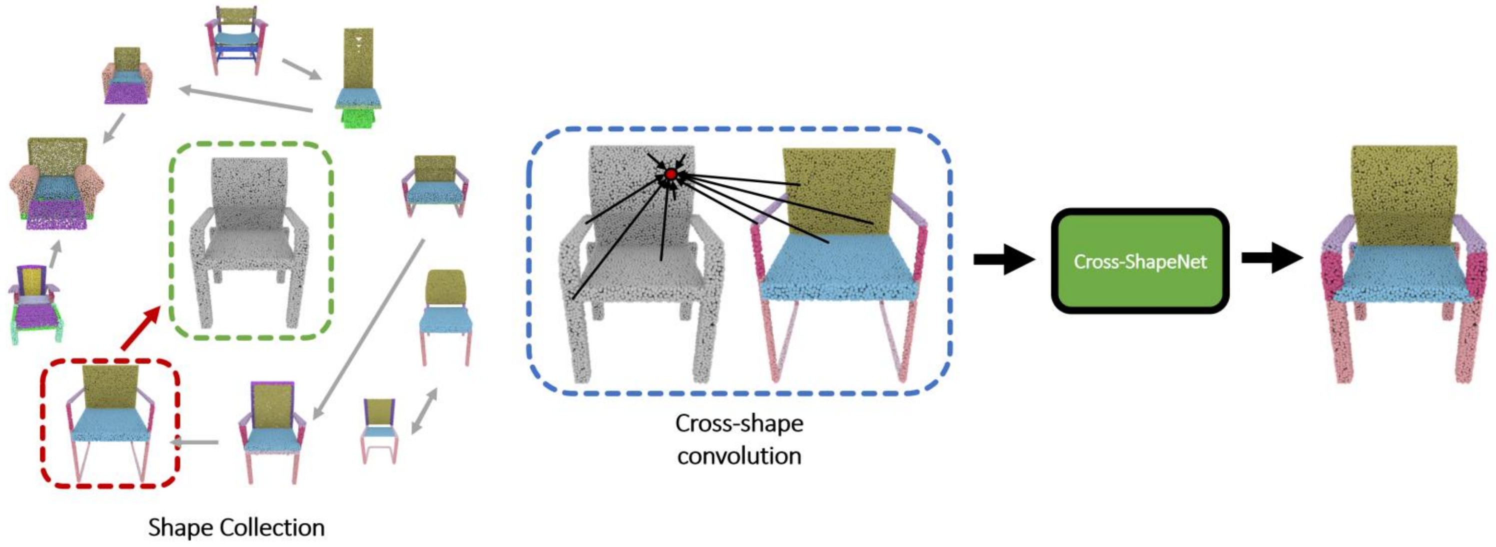
Summary



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

Adobe

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