# CodedVTR: **Code**book-based Sparse **V**oxel **TR**ansformer with Geometric Guidance




Tianchen Zhao | Niansong Zhang | Xuefei Ning

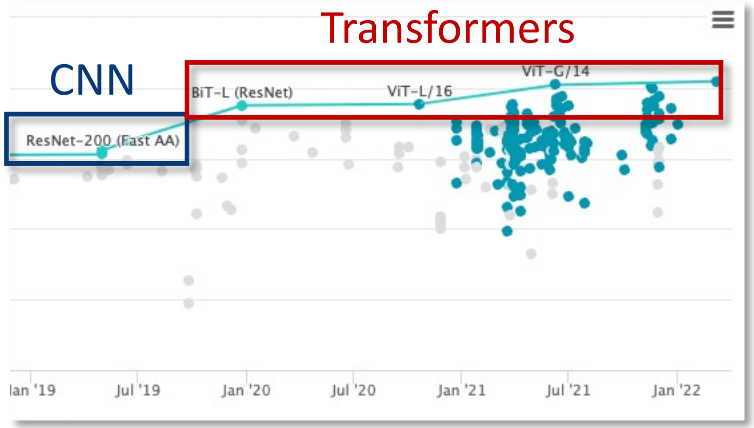


He Wang | Li Yi* | Yu Wang

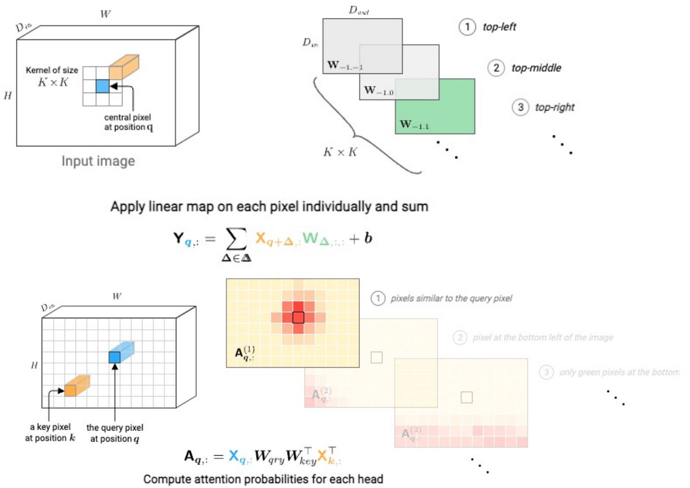
** Corresponding Author*

# Background

- **Transformers outperform CNN** and achieve SOTA in many vision tasks
- Transformer's superiority:
  - Less inductive bias -> Better representative power




Performance of ImageNet1K Classification[1]




Self-Attention could **approximate** Conv,
And it is a **more generalized form** of Conv[2]

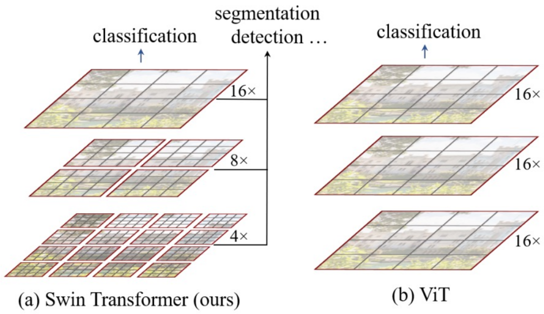

[1] ImageNet Benchmark (Image Classification) | Papers With Code
[2] Cordonnier, J., Loukas, A., & Jaggi, M. (2020). On the Relationship between Self-Attention and Convolutional Layers. ArXiv, abs/1911.03584.

# Background

- **Transformer's Problem:** Harder to optimize and generalize
  - Rely on large-scale pretraining, **Overfits** when directly trained on smaller dataset
  - Slow convergence & Sensitive to training hyperparameters (LR, initialization, DataAug…)
- **Current Solution:** Introduce domain-specific inductive bias

> *" When directly trained on the ImageNet, ViT yields modest accuracies of a few points below ResNets of comparable size " [1]*

ViT[1] requires large-scale pretraining to outperform ResNets

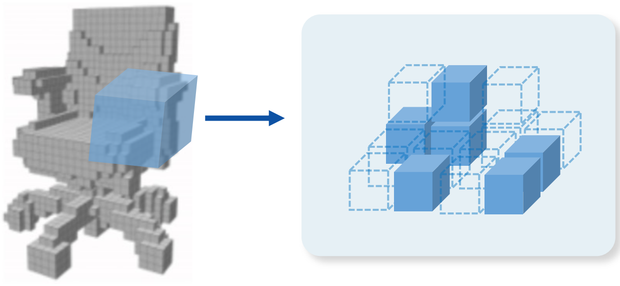



Swin Transformer[2] uses **Hierarchical Window-based local** aggregation


[1] Dosovitskiy, Alexey et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." ArXiv abs/2010.11929 (2021): n. pag.
[2] Liu, Ze et al. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2021): 9992-10002.

# Background

- Introducing Transformer into 3D Domain: The Generalization Issue is aggravated
  - 3D Data (Sparse Voxel) has unique properties (Sparse & Irregular)
  - Relatively limited data scale



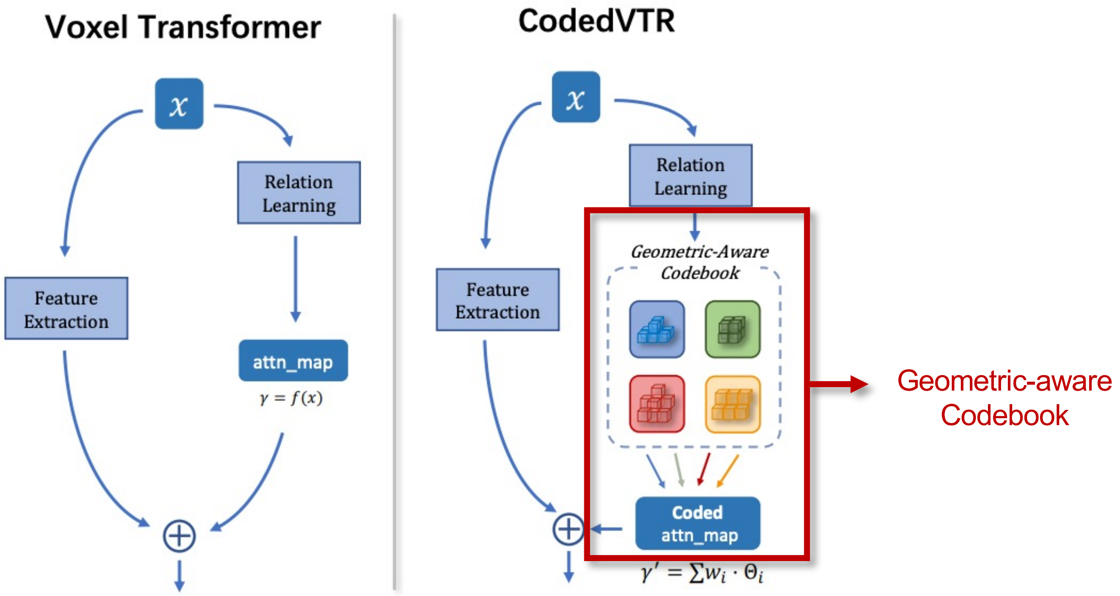
Voxel's unique Properties:
Sparse and Irregular

| Dataset | Method (Model) | | Params | mIOU |
|---|---|---|---|---|
| ScanNet | Convolution | Minkowski-M | 7M | 67.3% |
| | | Minkowski-L | 11M | 72.4% |
| | Transformer | PointTransformer | 6M | 58.6% (-8.7%) |
| | | VoTR (Mink-M) | 7M | 62.5% (-4.8%) |
| | | VoTR (Mink-L) | 11M | 66.1% (-6.3%) |
| SemanticKITTI | Convolution | Minkowski-M | 7M | 58.9% |
| | | Minkowsk-L | 11M | 61.1% |
| | Transformer | VoTR (Mink-M) † | 7M | 56.5% (-2.4%) |
| | | VoTR (Mink-L) | 11M | 58.2% (-2.9%) |

Simply employ transformer
fails to outperform CNN

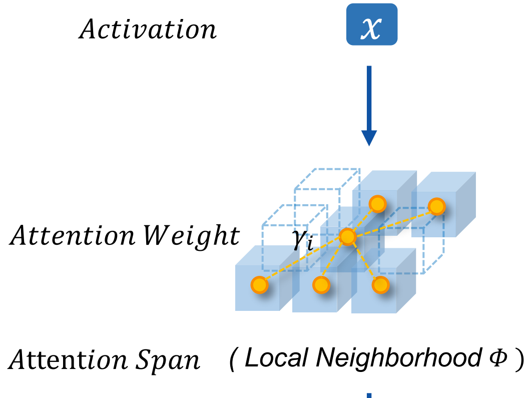
# Contribution

- **Key**: Alleviate the Generalization Issue
- **CodedVTR:** introduce **Geometric-aware Codebook**
  - Codebook-based Attention
  - Geometric-aware Attention



# Motivation

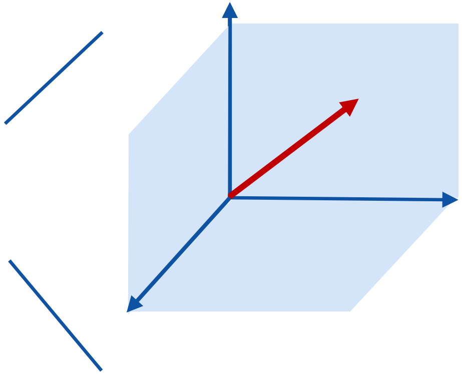
- **Comparison of Conv and Transformer** (Local Self-Attention)
  - The Attention Weight Generation:




$\gamma = f(x)$

$y = \sum_i^{\Phi} \gamma_i \cdot g(x_i)$

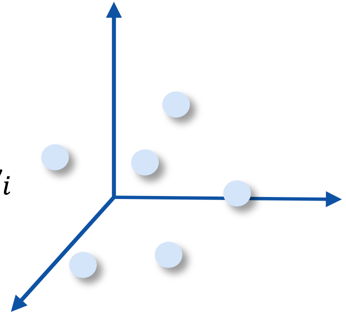
***Self-Attention***



*Learn the whole*
*Continuous space* $f(x)$
***of Dimension:*** $N(\Phi)$

***Convolution***

$\gamma = w_i$



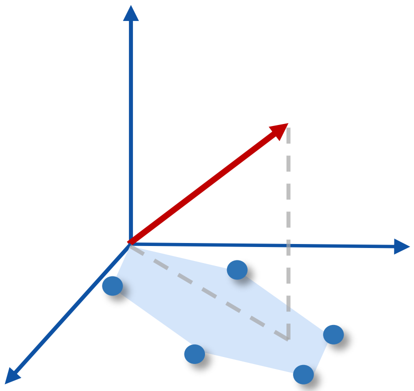
*Learn some weights* $w_i$
*on Anchor Points*

# Methodology

- **Codebook-based attention:** Encode the Attn-Weight with Codes

$$f(x) \sim f_d(x) = \sum w_i \theta_i$$

- Codes $\theta_i$ could be viewed as:
  - attention weight "Prototypes"
  - a set of basis span a subspace
- **Project** the attention learning in the subspace,
- **Regularization** – helps generalization



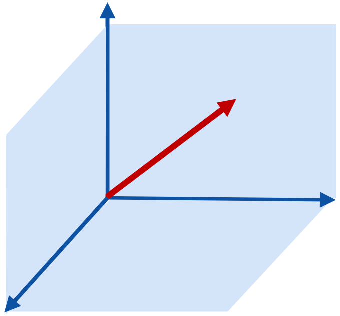
*Learn the*
*Subspace* $f_d(x)$
***of Dim:*** $N(\Theta) < N(\Phi)$

# Methodology

- Codebook-based attention is an **intermediate state** of self-attention and convolution

***Self-Attention***

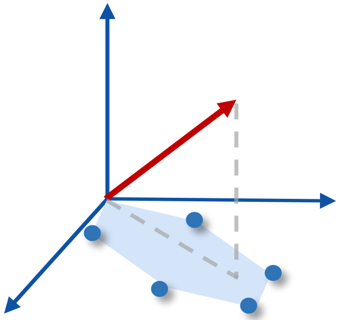
$$\gamma = f(x)$$



$$N(\Theta) = \infty$$

***Codebook-based Self-Attention***

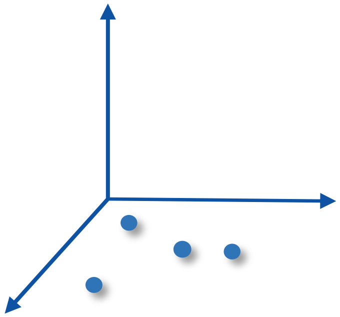
$$\gamma = f_d(x) = \sum w_i \, \theta_i$$



$$N(\Theta)$$

***Convolution***

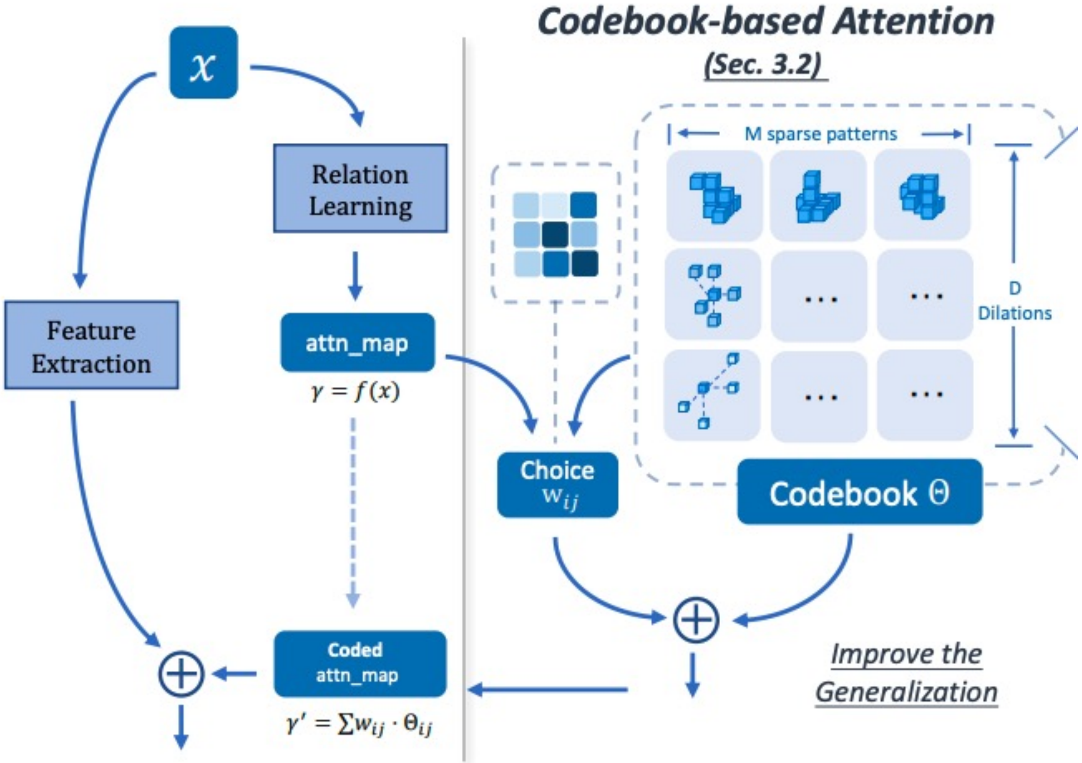
$$\gamma = w_i$$



$$N(\Theta)=1$$

# Methodology

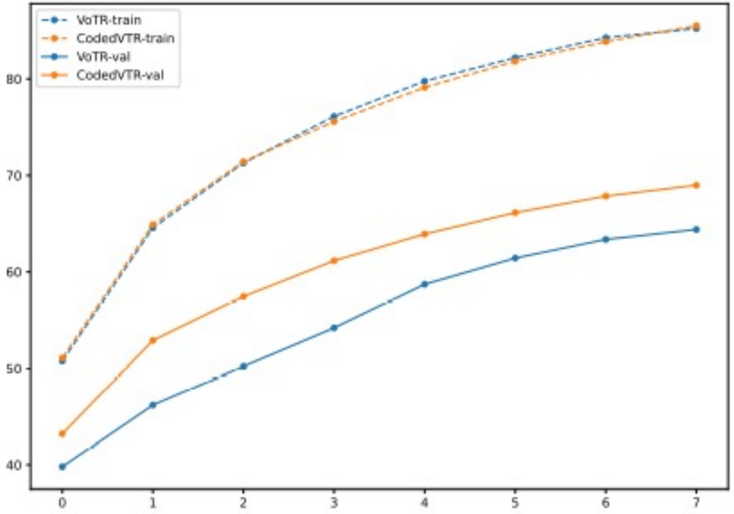
- Codebook Design



# Experimental Results

- Results of Codebook Design

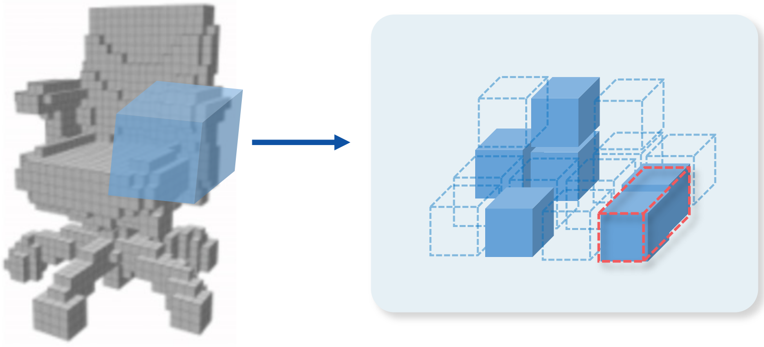
| $N(\Phi)$ | mIOU |
| --- | --- |
| 1 | 57.1% |
| 4 | 58.5% |
| 24 | **60.4%** |
| 48 | 58.1% |



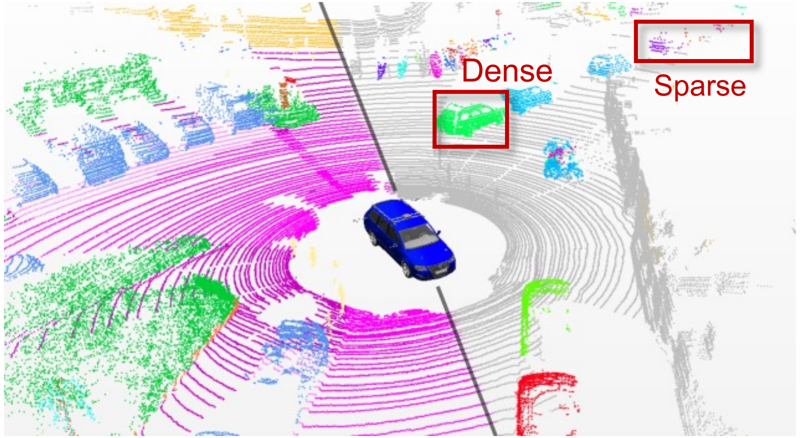

CodedVTR helps generalization

# Methodology

- 3D Data’s Unique Property -> Geometric-aware self-attention



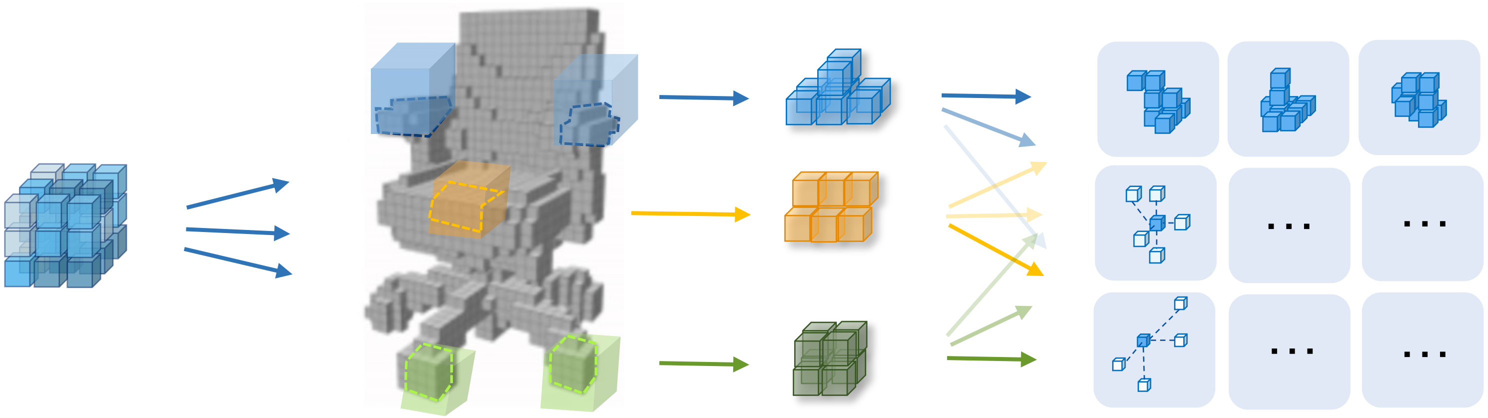
Sparse & Geometric Shape




Non-uniform Density(Outdoor)

# Methodology

- Geometric-aware self-attention
- **Geo-shape: Assign different geometric shapes for codebook elements**
- **Geo-guide: Encourage attention to match actual sparse pattern**



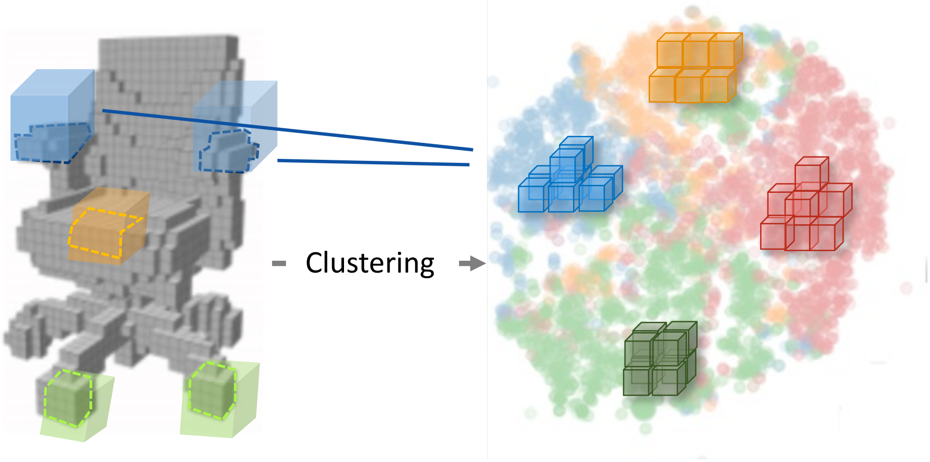

*Same convolution weight*

*different input voxel sparse pattern*

*Different Geometric Shapes Codebook-Elements*

# Methodology

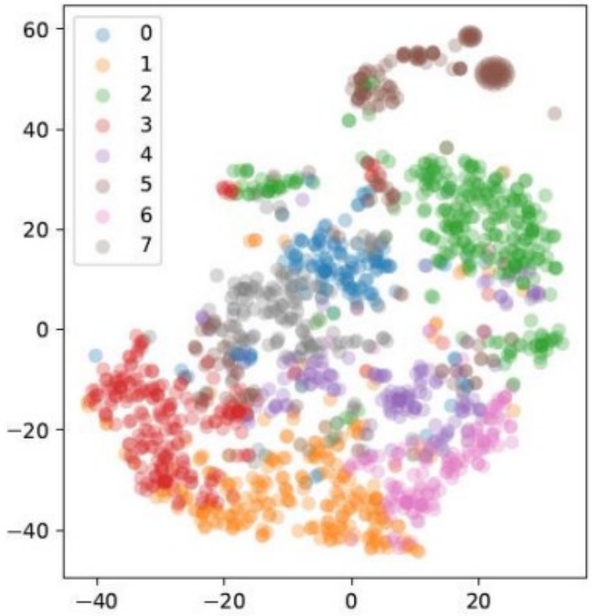
- How to determine geometric shape?
  - Adopt **K-means Clustering** to get 8 representative sparse pattern in 3 dilations




Actual Sparse Pattern

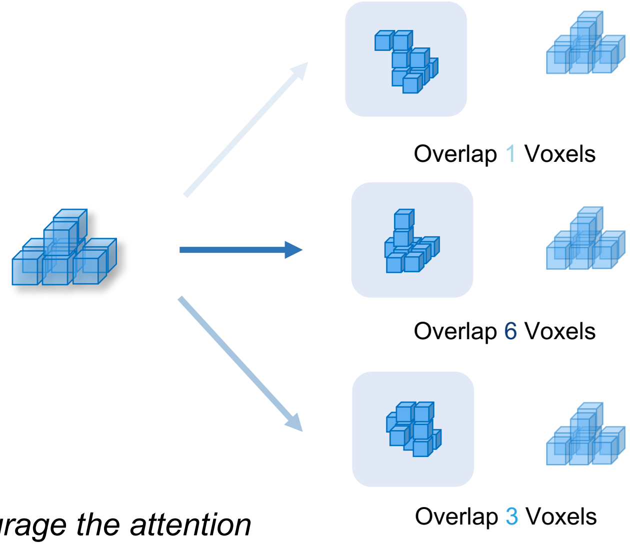
Representative Geometric Pattern

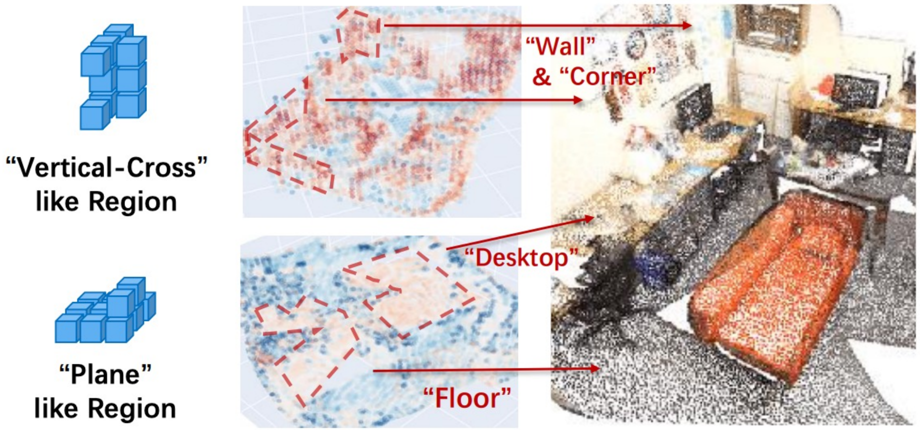
Clustering t-SNE Visualization



# Methodology

- How to Geometric guide?
  - Regularization for attention with “mismatch code”

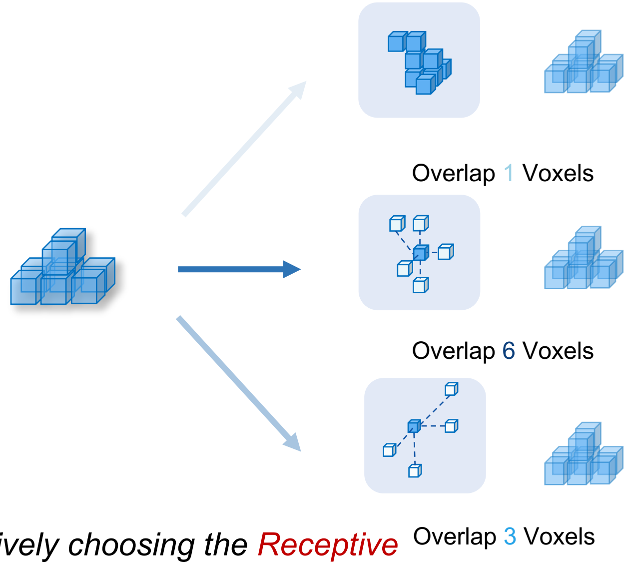


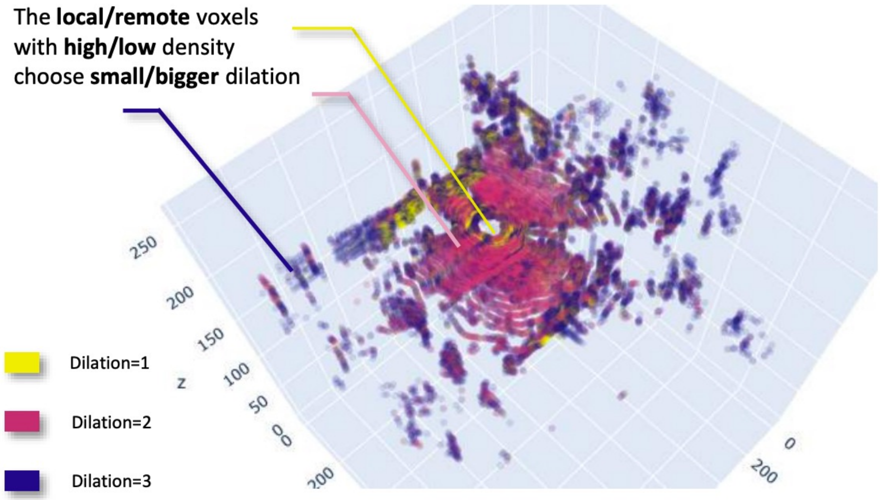

*Encourage the attention*
*To lean to matching geometric shape*

# Methodology

- How to Geometric guide?
  - Regularization for attention with “mismatch code”




*Adaptively choosing the Receptive Field for different densities*

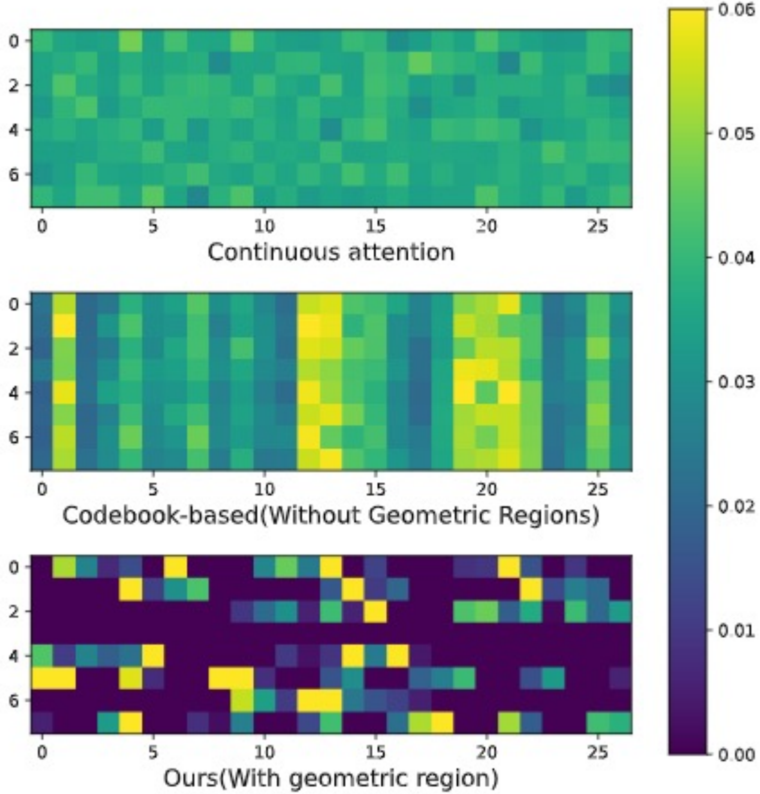


# Experimental Results

Naïve self-attention
(Uniform Attention Map)

Codebook-based self-attention
(Still Similar attention map)

Geometric-aware self-attention
(Meaningful attention map)



# Experimental Results

| Dataset | Method (Model) | | Params | mIOU |
|---|---|---|---|---|
| ScanNet | Convolution | Minkowski-M | 7M | 67.3% |
| | | Minkowski-L | 11M | 72.4% |
| | Transformer | CodedVTR (Mink-M) | 7M | **68.8**%(+1.5%) |
| | | CodedVTR (Mink-L) | 11M | **73.0**%(+0.6%) |
| SemanticKITTI | Convolution | Minkowski-M | 7M | 58.9% |
| | | Minkowsk-L | 11M | 61.1% |
| | | SPVCNN | 8M | 60.7% |
| | Transformer | CodedVTR (Mink-M) | 7M | **60.4**% (+0.5) |
| | | CodedVTR (Mink-L) | 11M | **63.2**% (+2.1%) |
| | | CodedVTR (SPVCNN) | 8M | **61.8**%(+1.1%) |
| Nuscenes | Convolution | Minkowski-M | 7M | 66.5% |
| | | Minkowsk-L | 7M | 69.4% |
| | Transformer | CodedVTR (Mink-M) | 7M | **69.9**% (+3.4%) |
| | | CodedVTR (Mink-L) | 11M | **72.5**% (+3.1%) |