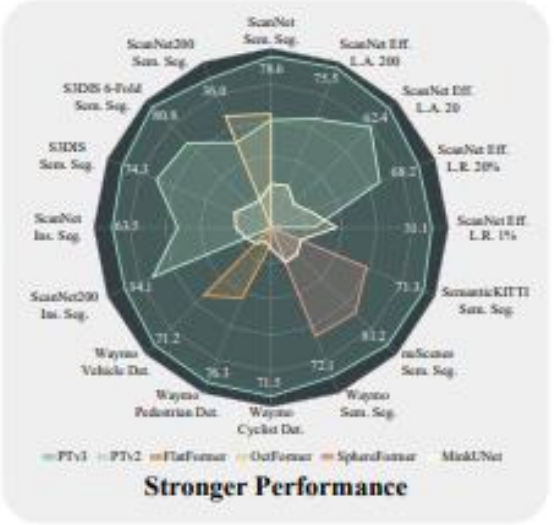
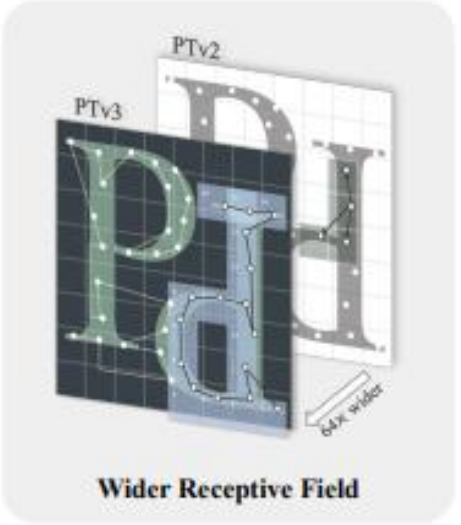
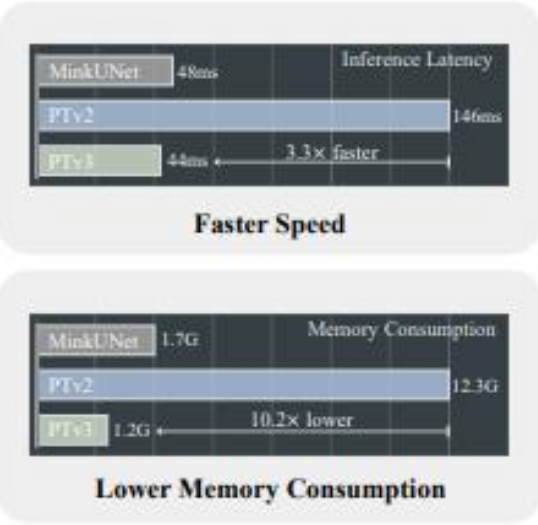
Point Transformer V3: Simpler, Faster, Stronger
(2024)

Paper Review
2025.10.13
Guebeen Lee

Contents





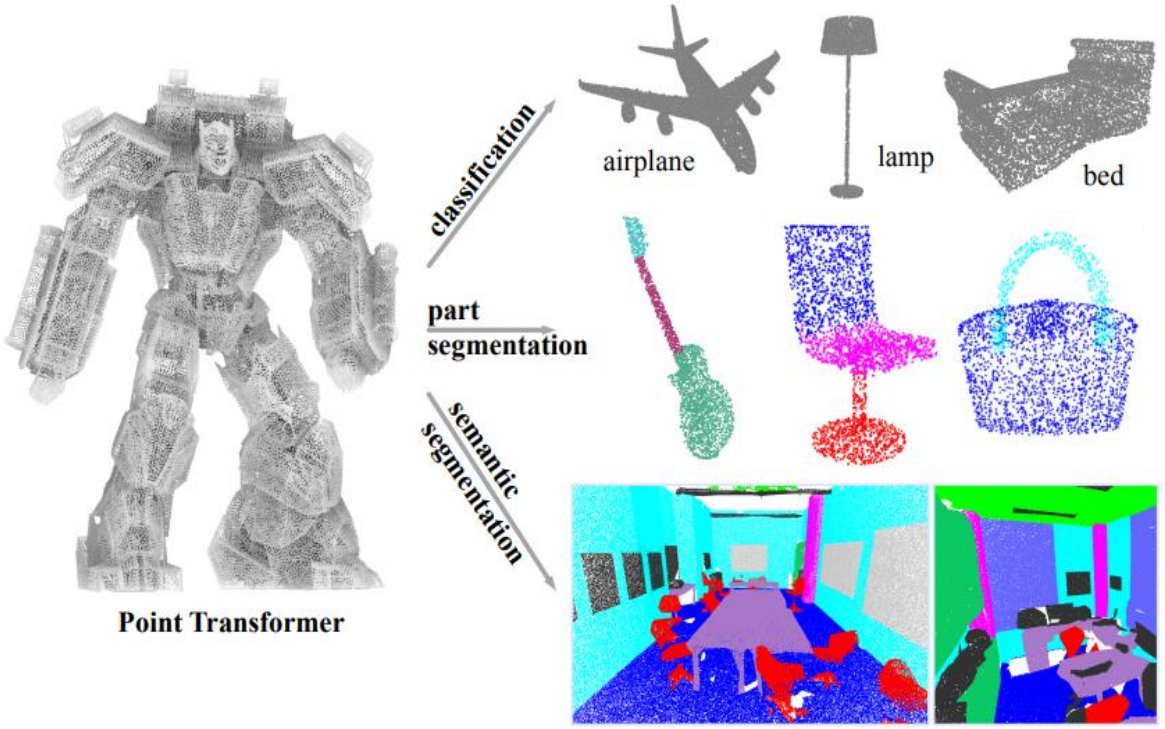


Introduction

Background(1)

>Point Transformer V1

- **Point-based** networks
- Transformer with self attention (**vector attention**)
- Position Encoding
- kNN structure
- **Residual** point transformer block



Introduction

Background(1)

>Point Transformer V1

- **Point-based** networks
- Transformer with self attention (**vector attention**)
- Position Encoding
- kNN structure
- **Residual** point transformer block

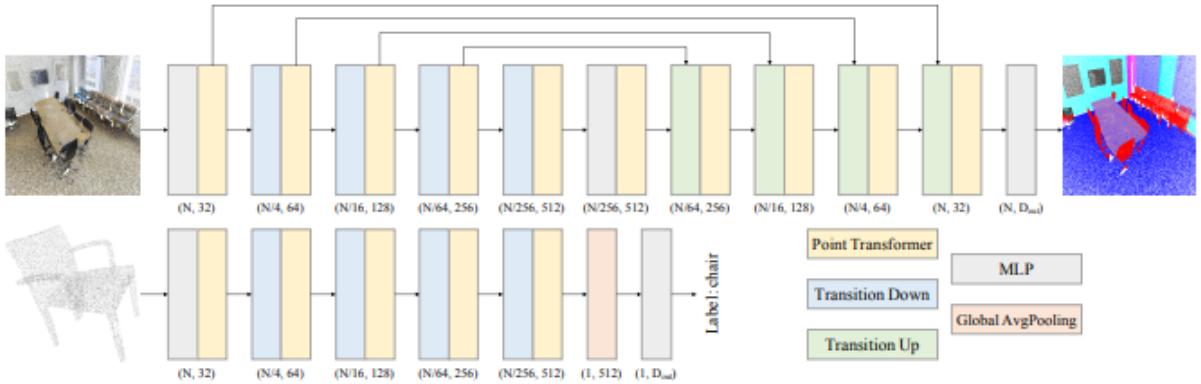
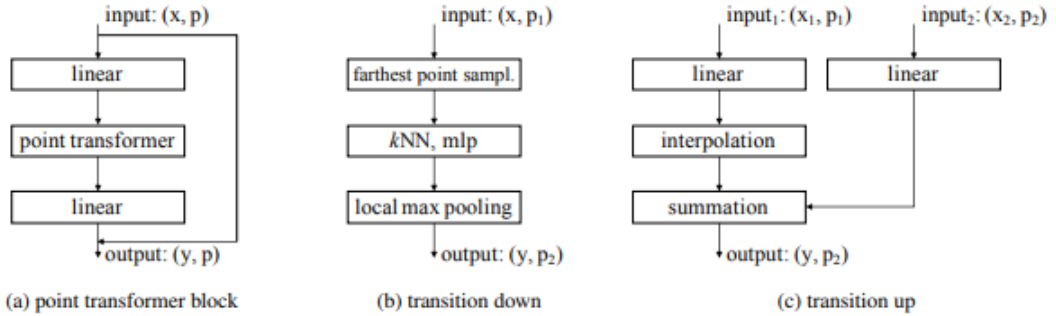


Figure 3. Point transformer networks for semantic segmentation (top) and classification (bottom).



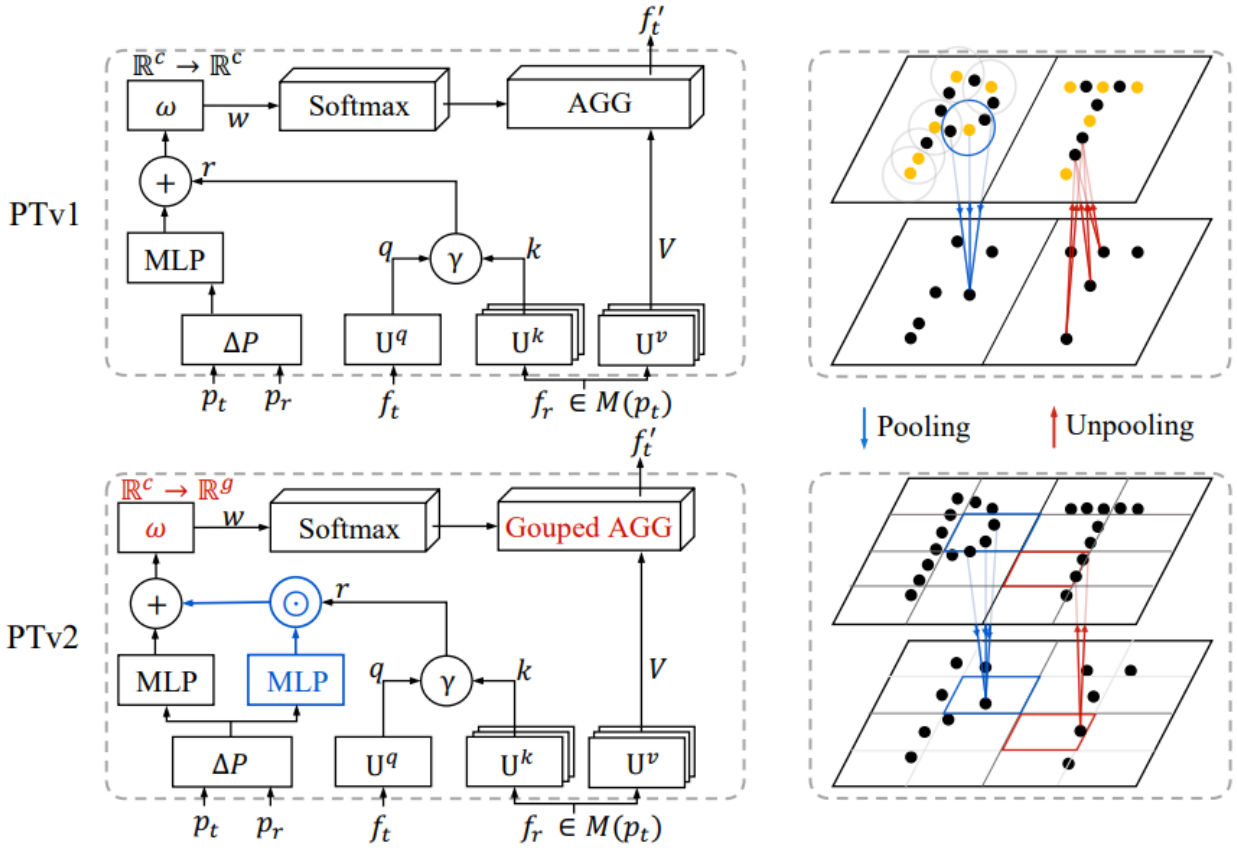
(a) point transformer block (b) transition down (c) transition up

Introduction

Background(2)

>Point Transformer V2

- Grouped Vector Attention
- Position Encoding Multiplier
- Partition-based Pooling

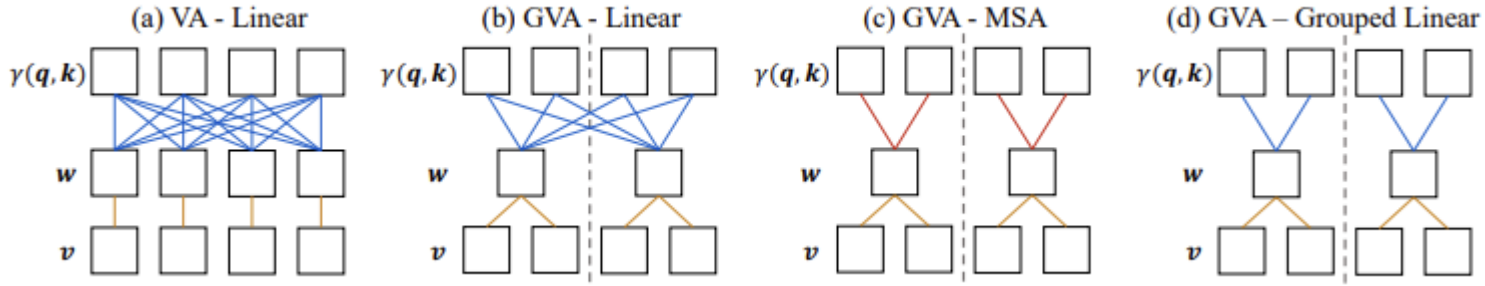


Introduction

Background(2)

>Point Transformer V2

- Grouped Vector Attention
- Position Encoding Multiplier
- Partition-based Pooling



$$\boldsymbol{w}_{ij} = \omega(\gamma(\boldsymbol{q}_i, \boldsymbol{k}_j)), \qquad \boldsymbol{f}_i^{attn} = \sum_{\boldsymbol{x}_j}^{\mathcal{M}(\boldsymbol{p}_i)} \sum_{l=1}^{g} \sum_{m=1}^{c/g} \text{Softmax}(\boldsymbol{W}_i)_{jl} \boldsymbol{v}_j^{lc/g+m},$$

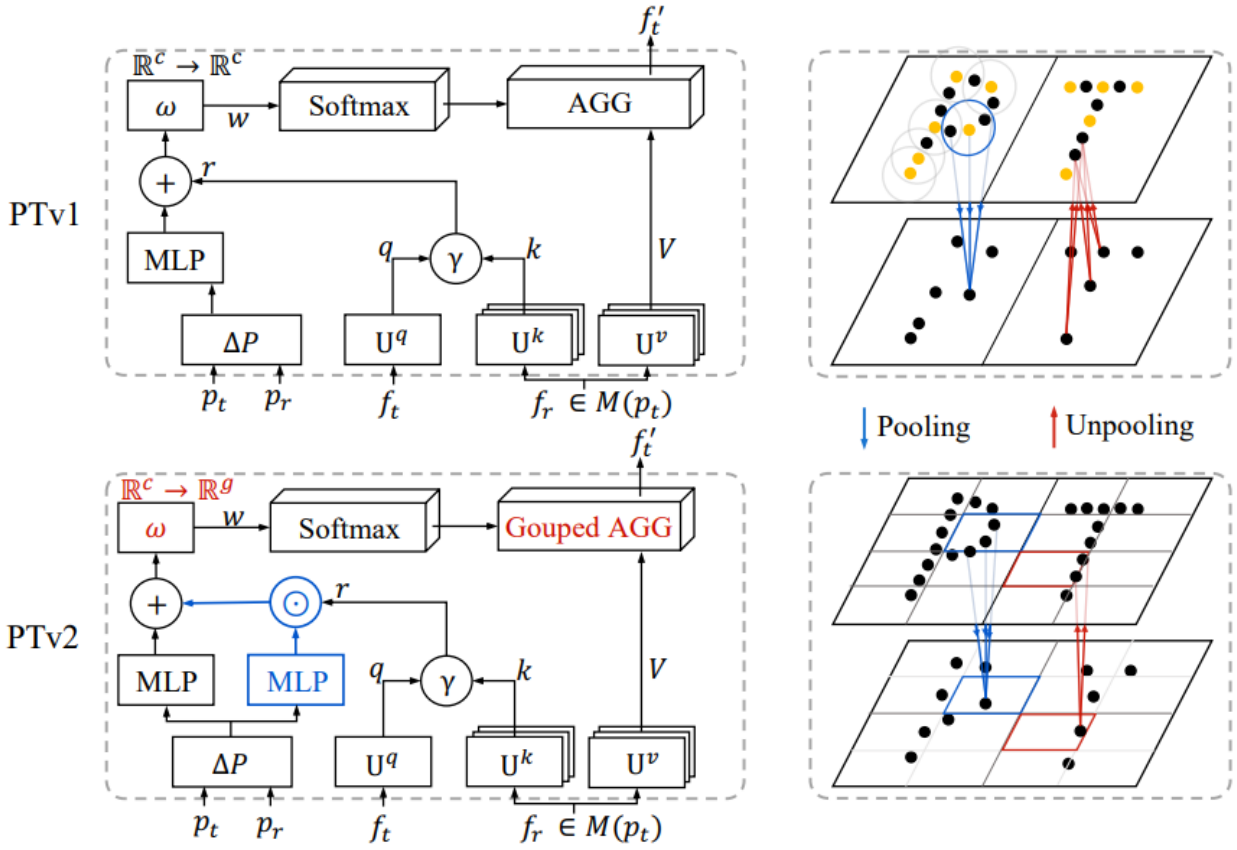
Introduction

Background(2)

>Point Transformer V2

- Grouped Vector Attention
- Position Encoding Multiplier
- Partition-based Pooling

$$\boldsymbol{w}_{ij} = \omega(\delta_{mul}(\boldsymbol{p}_i - \boldsymbol{p}_j) \odot \gamma(\boldsymbol{q}_i, \boldsymbol{k}_j) + \delta_{bias}(\boldsymbol{p}_i - \boldsymbol{p}_j)),$$



Introduction

Background(2)

>Point Transformer V2

- Grouped Vector Attention
- Position Encoding Multiplier
- Partition-based Pooling

$$\boldsymbol{f}_i' = \text{MaxPool}(\{\boldsymbol{f}_j \boldsymbol{U} \mid \boldsymbol{f}_j \in \mathcal{F}_i\}), \qquad \boldsymbol{p}_i' = \text{MeanPool}(\{\boldsymbol{p}_j \mid \boldsymbol{p}_j \in \mathcal{P}_i\}),$$

Purpose

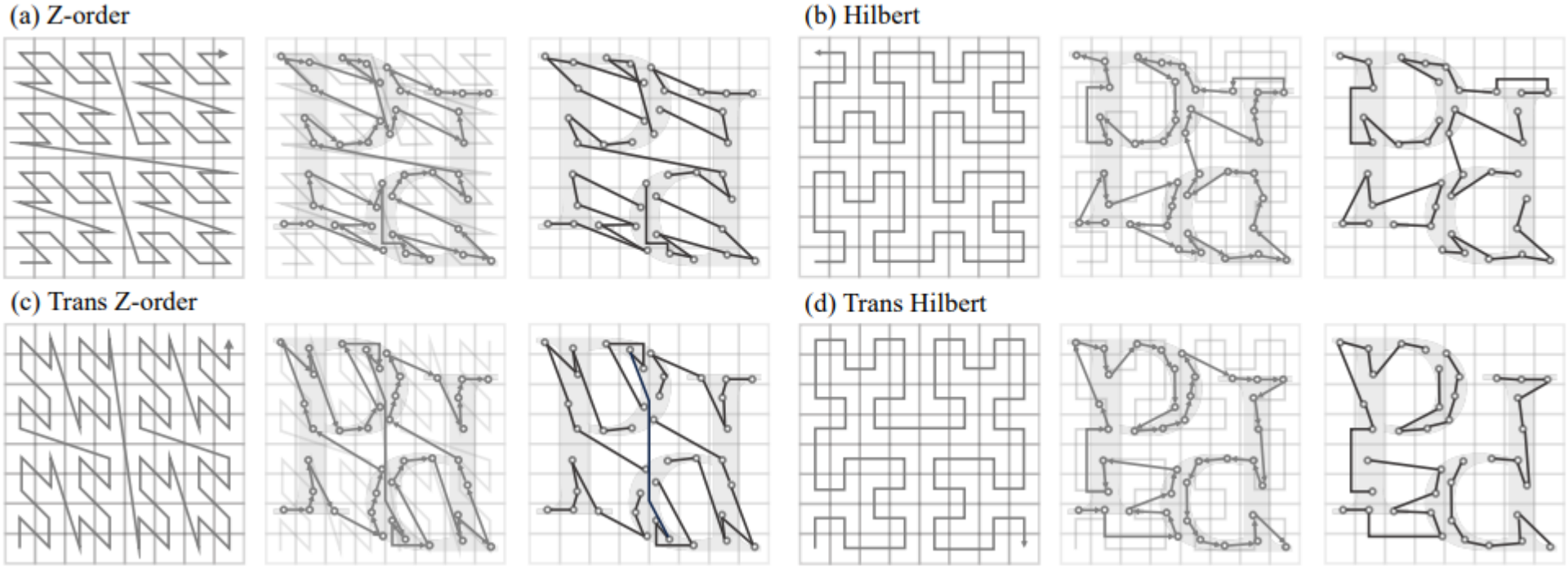
>Prioritizes simplicity and efficiency over accuracy of certain mechanisms

- Structuring unstructured point clouds
- Replace complex attention patch interaction(such as shift-window) and the neighborhood mechanism with streamlined approach for serialized point clouds
- Eliminates the reliance on relative positional encoding

Methodology

>Point Cloud Serialization

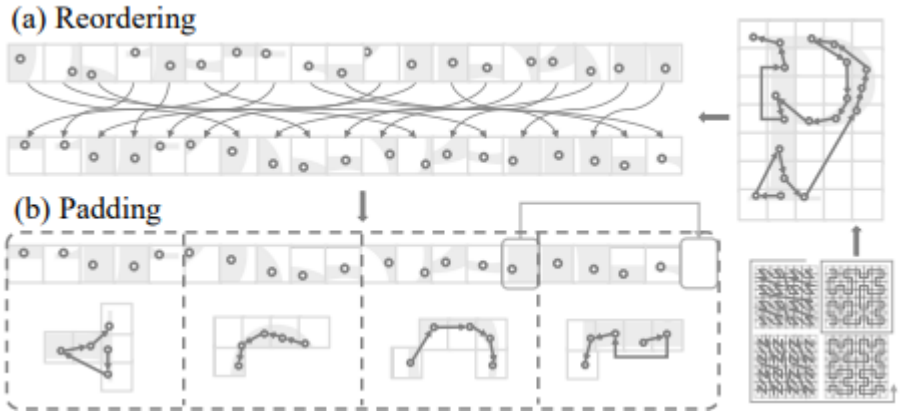
- Space filling curves
- Serialization

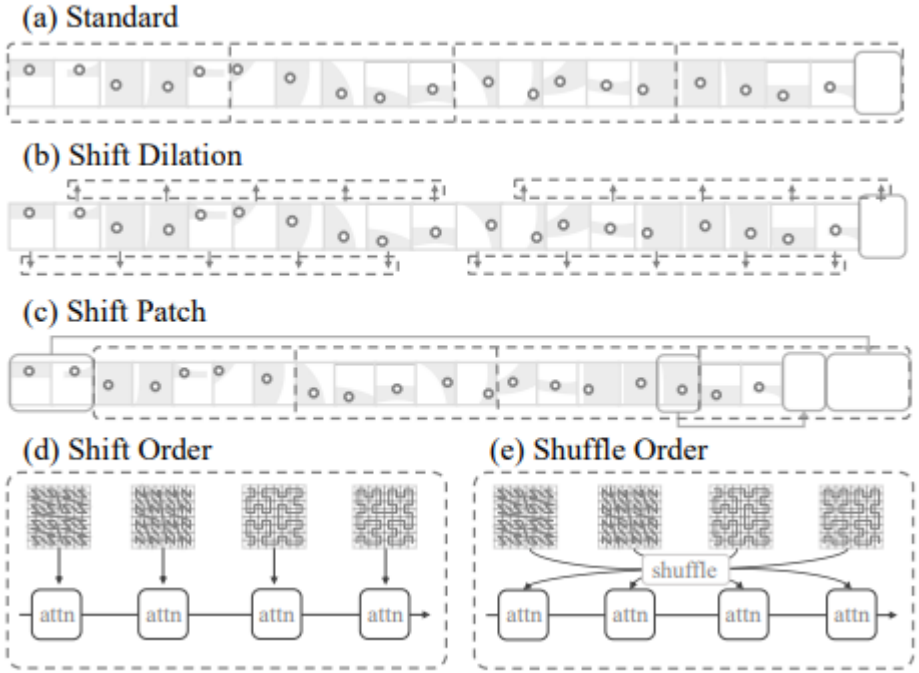


Methodology

>Serialized Attention

- Re-weigh options of attention mechanism
- Patch Grouping
- Patch Interaction
- Positional encoding





Methodology

>Network Details