NEURAL INFORMATION
PROCESSING SYSTEMS

Expediting Large-Scale Vision Transformer for Dense Prediction without Fine-tuning

Weicong Liang
PKU

Yuhui Yuan ✉
MSRA

Henghui Ding
ETH Zurich

Xiao Luo
PKU

Weihong Lin
MSRA

Ding Jia
PKU

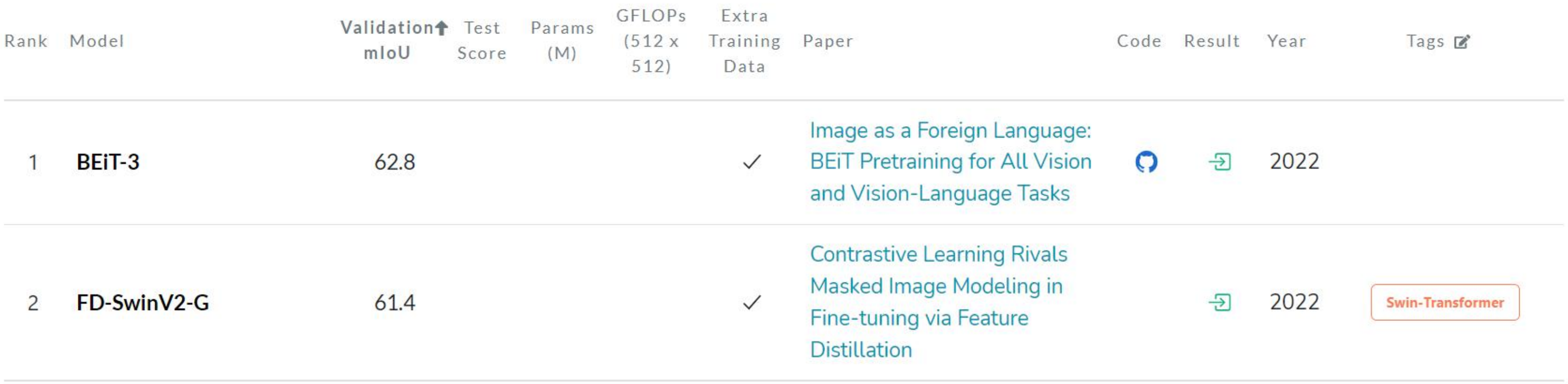
Zheng Zhang
MSRA

Chao Zhang
PKU

Han Hu
MSRA

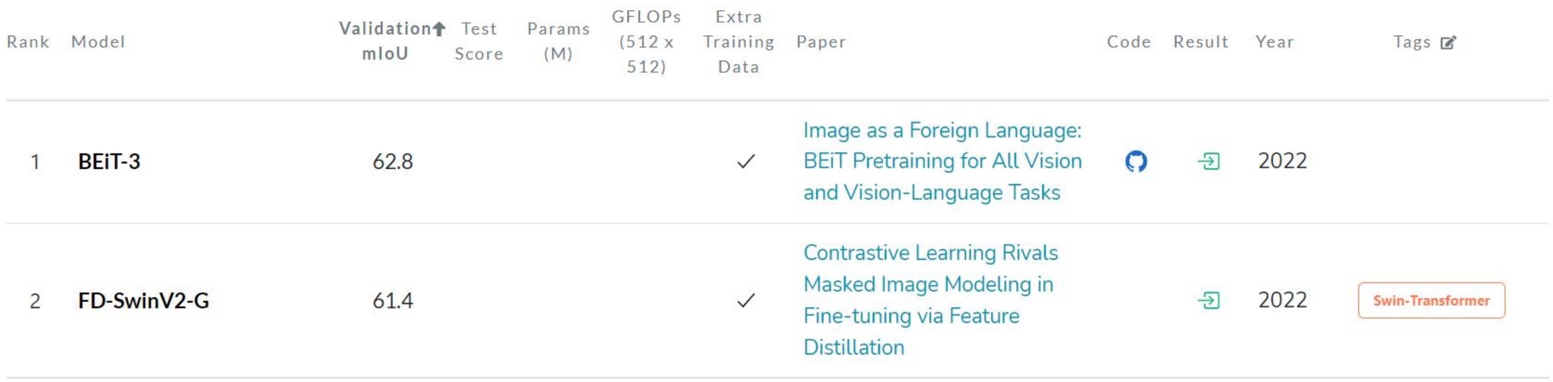
Motivation

Semantic segmentation



Rank	Model	Validation↑ mIoU	Test Score	Params (M)	GFLOPs (512 x 512)	Extra Training Data	Paper	Code	Result	Year	Tags
1	**BEiT-3**	62.8				✓	Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks			2022	
2	**FD-SwinV2-G**	61.4				✓	Contrastive Learning Rivals Masked Image Modeling in Fine-tuning via Feature Distillation			2022	Swin-Transformer

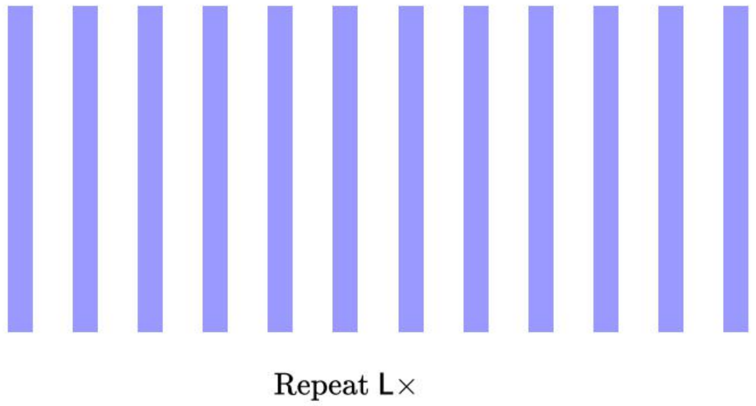
Detection



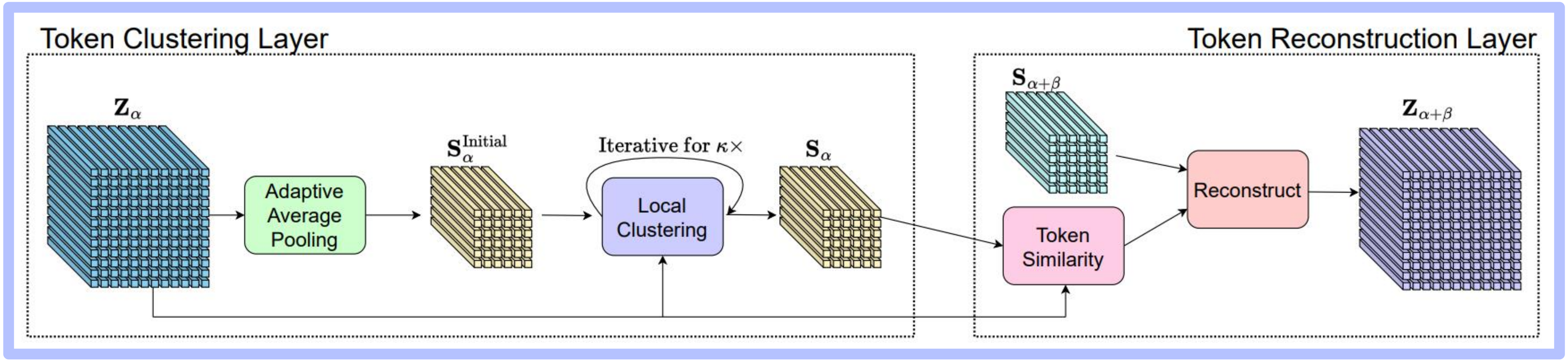
Rank	Model	Validation↑ mIoU	Test Score	Params (M)	GFLOPs (512 x 512)	Extra Training Data	Paper	Code	Result	Year	Tags
1	**BEiT-3**	62.8				✓	Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks			2022	
2	**FD-SwinV2-G**	61.4				✓	Contrastive Learning Rivals Masked Image Modeling in Fine-tuning via Feature Distillation			2022	Swin-Transformer

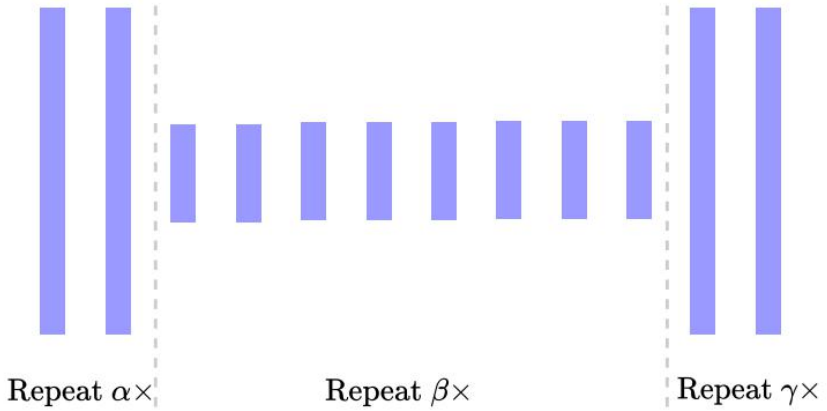
However, large-scale vision transformers suffer from **huge computation overheads** and **expensive latency**.

Our Approach



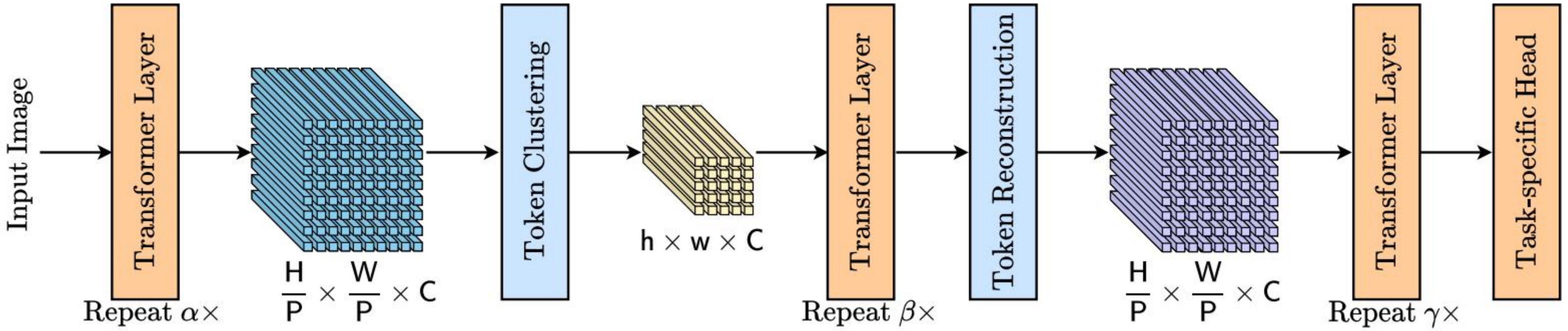
Plain high-resolution vision transformer during training

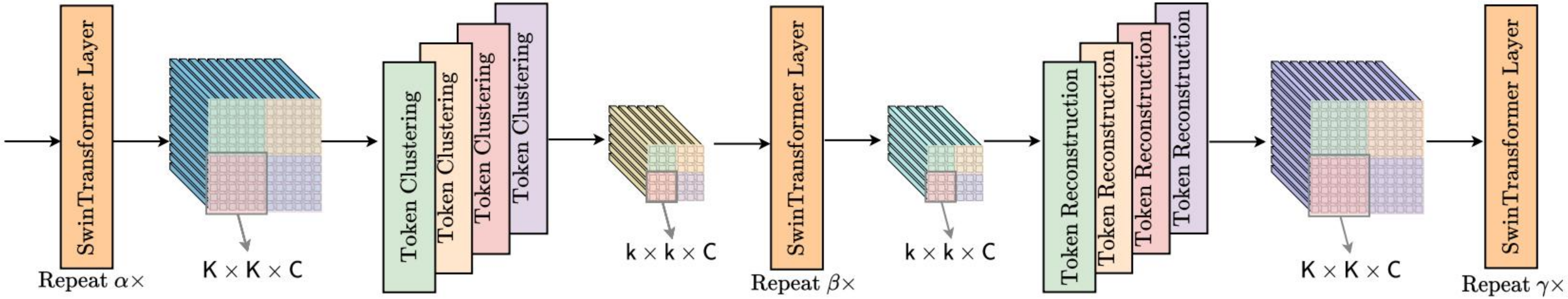

Token Clustering Layer
Token Reconstruction Layer
$\mathbf{Z}_{\alpha}$
Adaptive Average Pooling
$\mathbf{S}_{\alpha}^{\text{Initial}}$
Iterative for $\kappa\times$
Local Clustering
$\mathbf{S}_{\alpha}$
$\mathbf{S}_{\alpha+\beta}$
Token Similarity
Reconstruct
$\mathbf{Z}_{\alpha+\beta}$



U-shape high-to-low-high resolution vision transformer during evaluation

Our Approach for standard ViT and Swin Transformer





Expediting various vision tasks with our approach

Method	COCO Object Det.			COCO Instance Seg.		
	FLOPs	FPS	mAP(%)	FLOPs	FPS	mask AP(%)
SwinV2-L + HTC++	921G	2.3	**58.9**	921G	2.3	**51.2**
+ Ours	**748G**	**2.8**	57.7	**748G**	**2.8**	50.3

Method	COCO Panoptic Seg.			ADE20K Semantic Seg.			COCO Instance Seg.		
	FLOPs	FPS	PQ(%)	FLOPs	FPS	mIoU(%)	FLOPs	FPS	mask AP(%)
Mask2Former	937G	4.3	**57.8**	937G	4.3	**55.8**	937	4.3	**50.1**
+ Ours	**663G**	**5.9**	56.8	**620G**	**6.2**	55.6	**705**	**5.4**	49.1

Method	KITTI			NYUv2		
	FLOPs	FPS	RMSE	FLOPs	FPS	RMSE
DPT	810G	11.4	**2.57**	560G	17.6	**0.36**
+ Ours	**627G**	**14.8**	2.60	**404G**	**24.0**	**0.36**

Our approach saves around 25% of computation cost but keeps 98% of performance.

QR code of Paper & Code

Paper

Code