Cluster to Abstract: Efficient Visual Abstraction via Cluster Distillation

THE UNIVERSITY OF BRITISH COLUMBIA

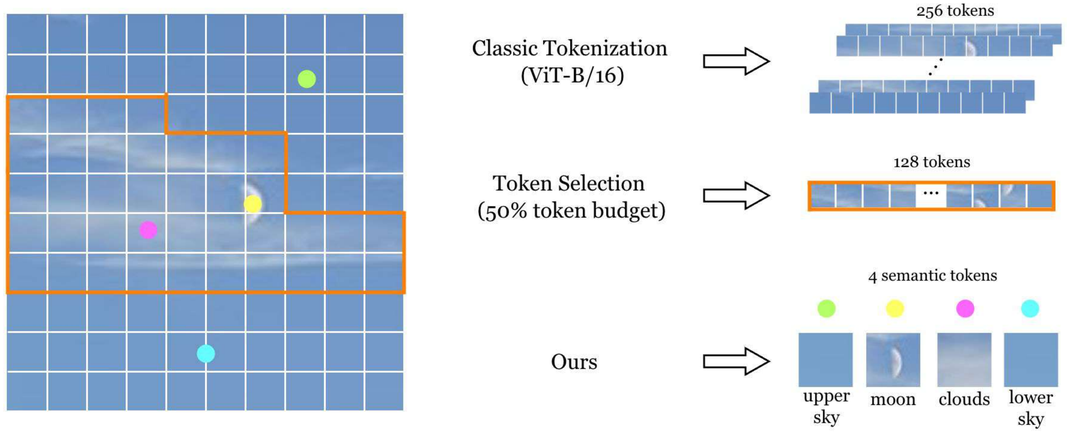
Yu-Chu Yu **Leonid Sigal** **Evan Shelhamer**

University of British Columbia Vector Institute

VECTOR INSTITUTE | INSTITUT VECTEUR

Motivation

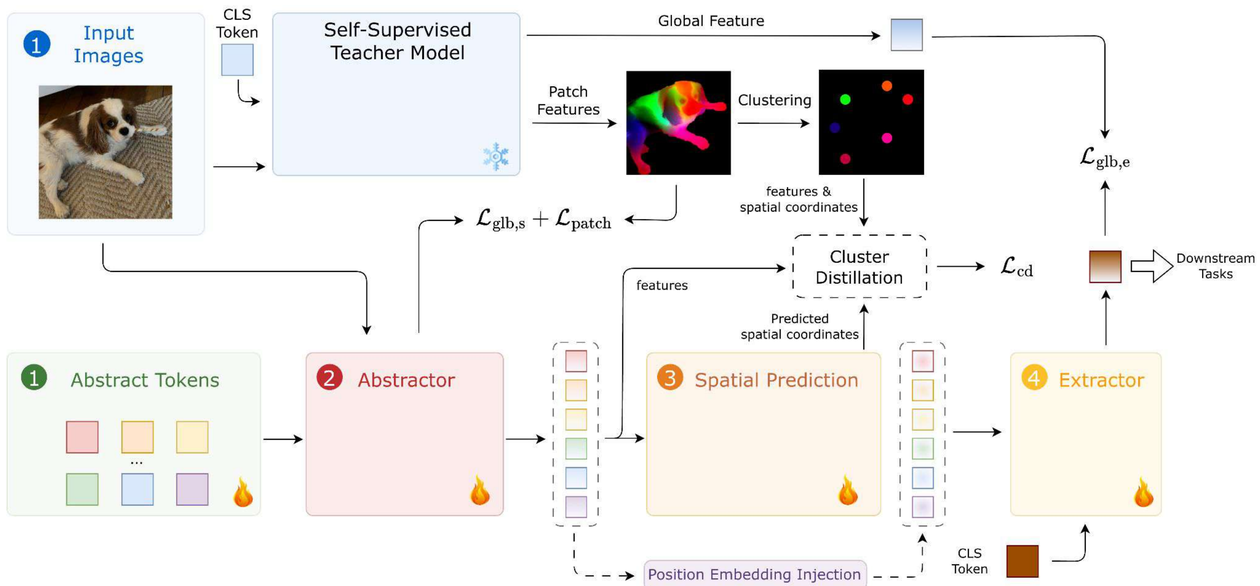
- Vision Transformers' computational costs scale quadratically with input resolution.
- Existing token selection strategies assign a fixed amount of token budget, ignoring images semantic complexity.



Contribution

- We propose abstract tokens that replace original dense patch tokens to reduce the computation for inference.
- We design a cluster distillation mechanism to learn these abstract tokens from a teacher model without label supervision.
- With only 8 abstract tokens, our method reduces the computational cost by **43%**, improves throughput by **1.81x**, and maintains **~98%** of model performance on visual recognition.

Model Architecture



Experimental Results

Method	Extractor's K	L_s	L_e	GFLOPs (↓)	Acc. (↑)
DINOv2	256	-	12	22.3 (100%)	84.9 (100%)
DTEM	94	-	12	9.2 (41.3%)	80.7 (95.3%)
ATC	169	-	12	15.3 (68.6%)	82.0 (96.6%)
LTRP	147	-	12	18.3 (82.1%)	82.8 (97.5%)
LookWhere	128	3	12	14.4 (64.8%)	83.0 (97.8%)
Ours	8	6	6	12.1 (54.3%)	83.3 (98.1%)
	8	6	12	12.7 (57.0%)	**83.6** (98.5%)