

The Extra Tokens Matter: Disentangled Representation Learning with Vision Transformers

Maofeng Tang¹, Hairong Qi¹



THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

¹The University of Tennessee, Knoxville, TN, USA
Seoul, South Korea, July 6th –July 11th
ICML 2026

Motivation

The XTRA is aiming for *Disentangled features* $\longrightarrow$ *interpretability, robustness.*

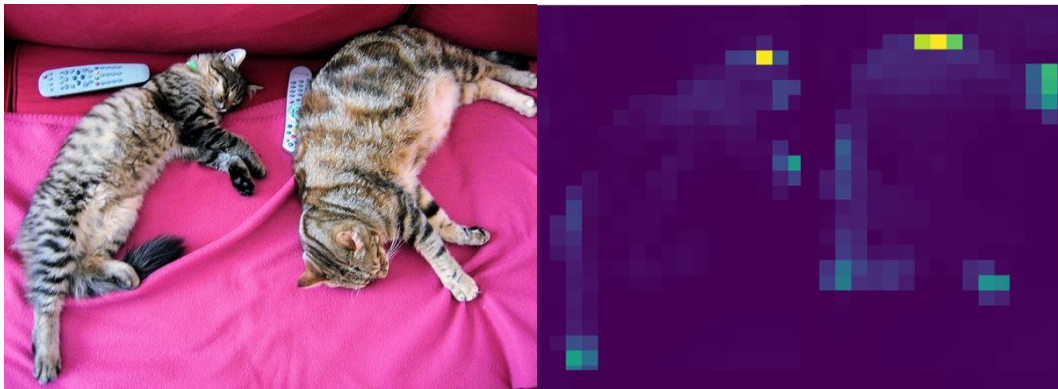


Figure.1a Vision transformer — global object representation

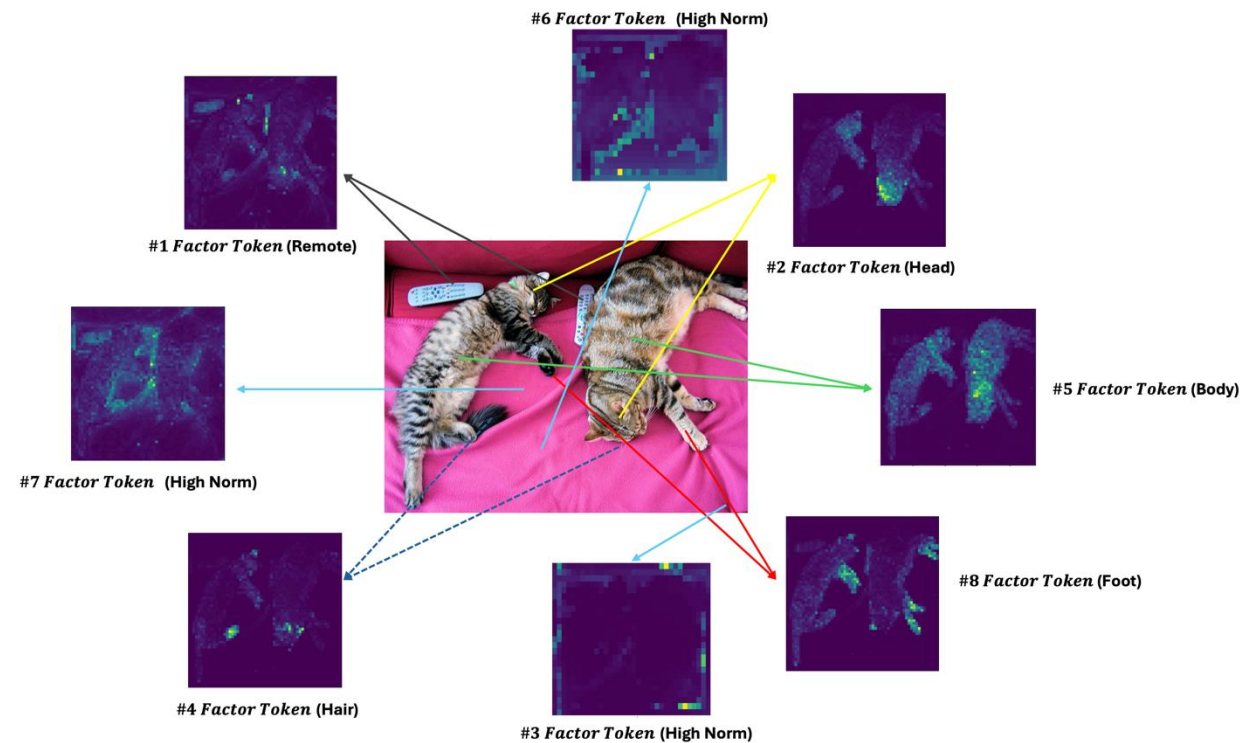


Figure.1b XTRA — explicit semantic parts

Method – Theoretical

Core Idea: Patch tokens as convex combinations of M pure semantic factors.

$$\mathbf{p}_i = F \cdot \mathbf{w}_i + \epsilon_i$$

F : Factor tokens set, P_i : Patch token, ϵ_i : noise

$J(F)$ — *orthogonality*; $\mathcal{L}_{factor}$ — *reconstruction*

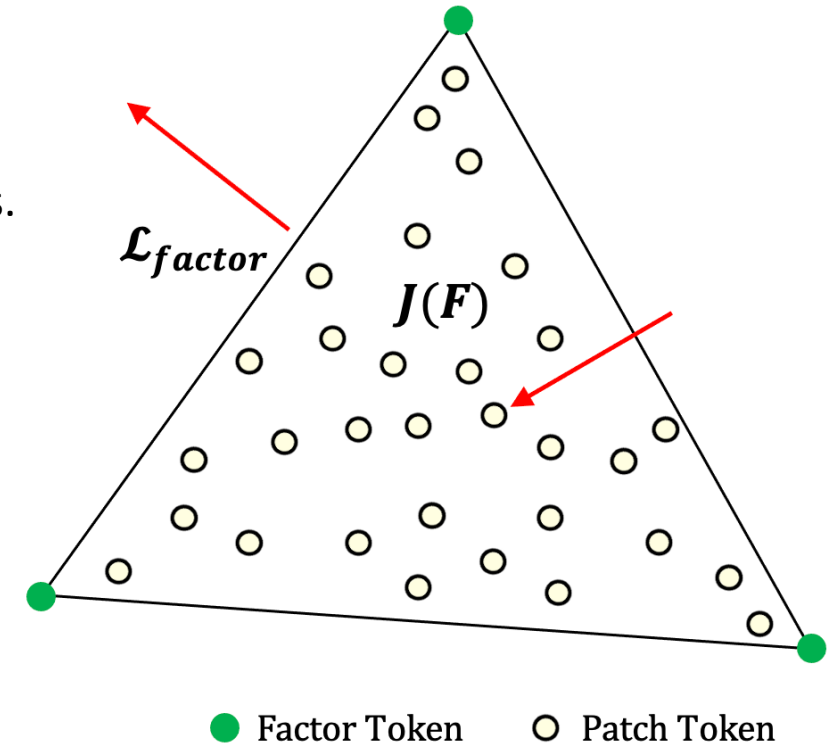


Figure.2 Illustration of the two-force geometric

Method– Overall Framework

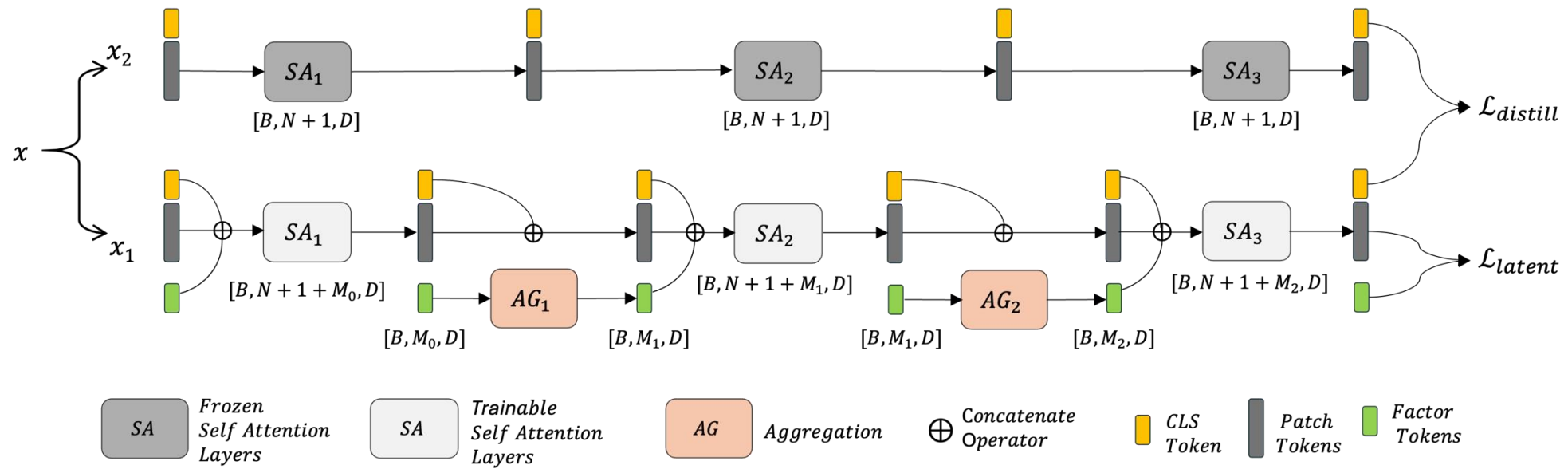


Figure.3 Illustration of the proposed XTRA architecture

Experiments

❖ Headline results

Table 1. Representation disentanglement score with SEPIN@k on ImageNet-1k

	SEPIN@1	SEPIN@10	SEPIN@100	SEPIN@all
DINOv2	0.47 ± 0.03	0.39 ± 0.02	0.28 ± 0.02	0.11 ± 0.01
DINOv2+	0.42 ± 0.02	0.35 ± 0.03	0.25 ± 0.01	0.13 ± 0.01
XTRA	3.95 ± 0.12	3.02 ± 0.09	1.54 ± 0.06	0.16 ± 0.04

Table 2. Part segmentation on PartImageNet. XTRA achieves superior part-level mIoU

Method	Part mIoU (%)	vs. DINOv2
DINOv2	42.3 ± 0.4	baseline
DINOv2+	43.1 ± 0.3	+0.8
XTRA (no MVC)	43.5 ± 0.5	+1.2
XTRA (full)	46.8 ± 0.3	+4.5

Table 3. Evaluation of representation in KNN and Linear Probing on IN-1K

	KNN	Linear
DINOv2	76.9	80.1
DINOv2+	77.3	82.1
XTRA (from scratch)	81.9	83.8
XTRA (pretrained on LVD142M)	84.2	86.0

Experiments

❖ Analysis of domain bias

Table 4. within-class patch variance of Vehicle categories vs animal categories

Category type	Within-class patch variance	XTRA SEPIN@1
Animals (quadrupeds)	0.43 ± 0.06	3.95
Vehicles (cars, planes)	0.18 ± 0.04	0.84

Table 5. within-class patch variance after Stanford Cars fine-tuning

Setting	Vehicle SEPIN@1	Animal SEPIN@1
XTRA (ImageNet pretrain only)	0.84	3.95
XTRA (+ Stanford Cars fine-tune)	2.31	3.89



Figure.4 Visualization of Factor Tokens in Other Objects

Thanks