INCHEON NATIONAL UNIVERSITY
COMPUTER VISION LABORATORY

Emerging Properties in Self-Supervised Vision Transformers (ICCV 2021)

Undergraduate Researcher at CVLab

Lee Dohyeong

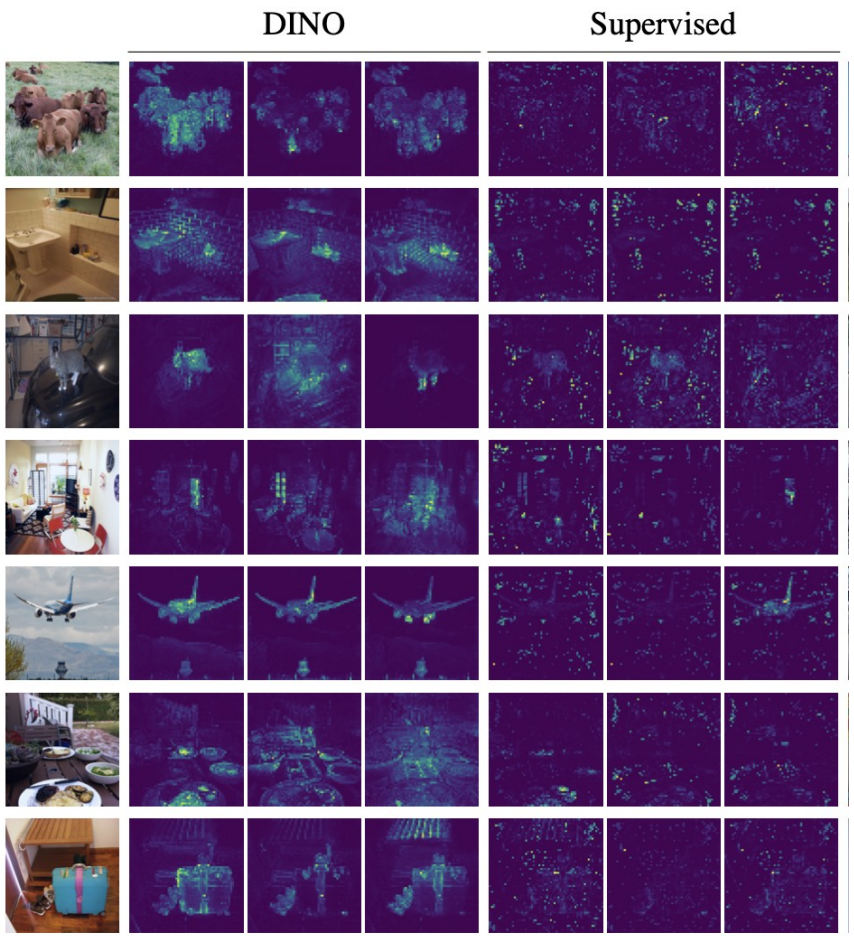
2026.01.15

Contents

- Introduction
- Background
- Method
- Experiments
- Conclusion

Introduction

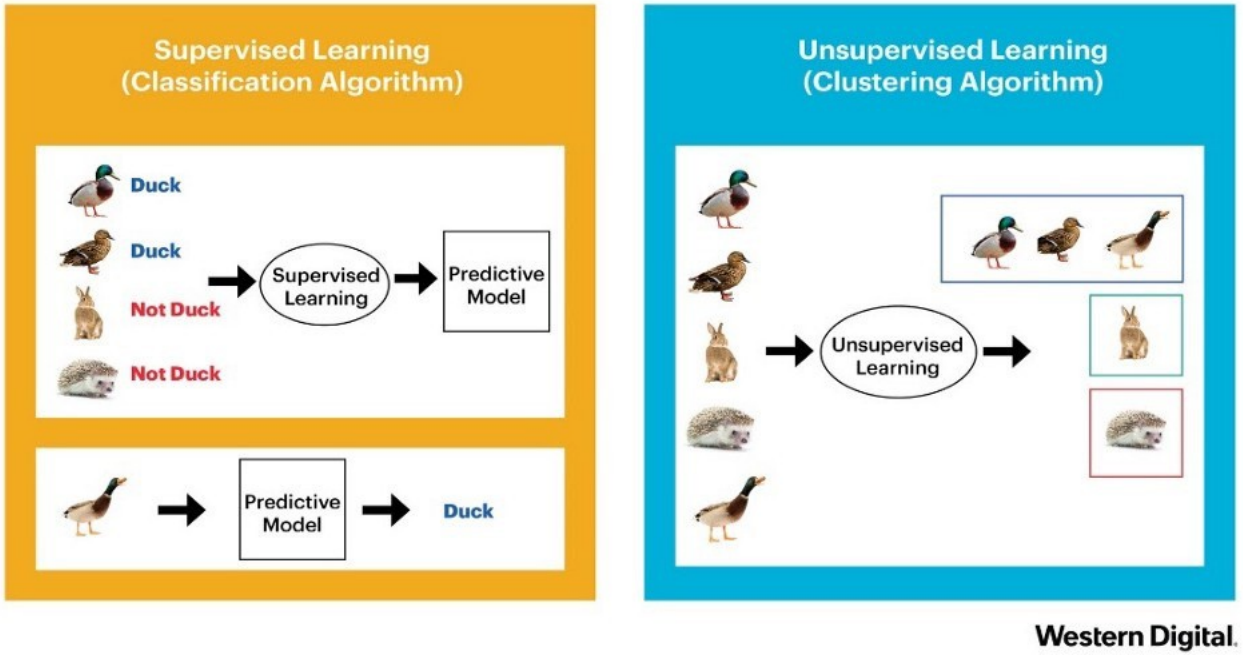
- Key method
  1. Knowledge Distillation
  2. Multi Crop
  3. EMA(Exponential Moving Average)



- DINO, Supervised Self-Attention Map

Background

- Supervised Learning, Self-Supervised Learning



- Label Dependency
- Discriminative Focus

- Feature Learning (Clustering)
- Structure Understanding

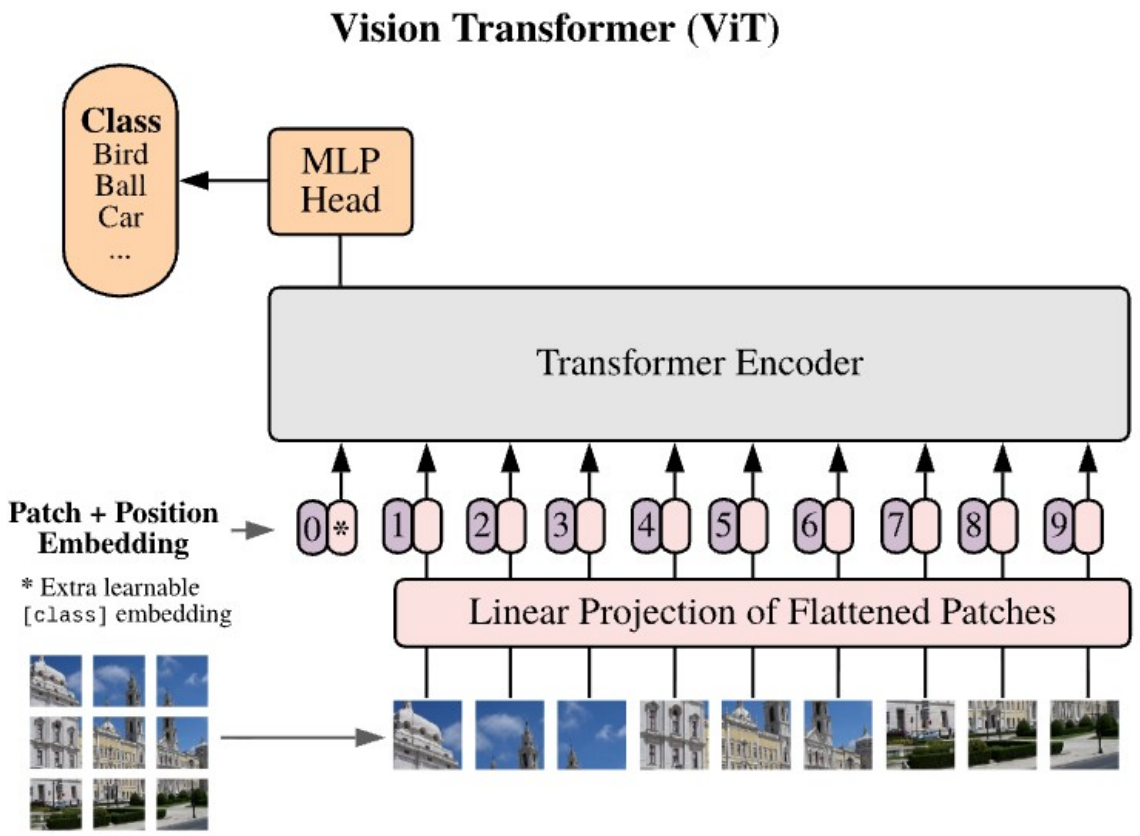
Background

- Knowledge Distillation
  - Concept: Transferring knowledge from a Large Model (Teacher) to a Small Model (Student).
  - Method: Training the Student to mimic the Teacher's outputs.
  - Objective: Achieving high performance with a lightweight model (Model Compression).

  - Dino : Self Distillation
    - Teacher <- Student

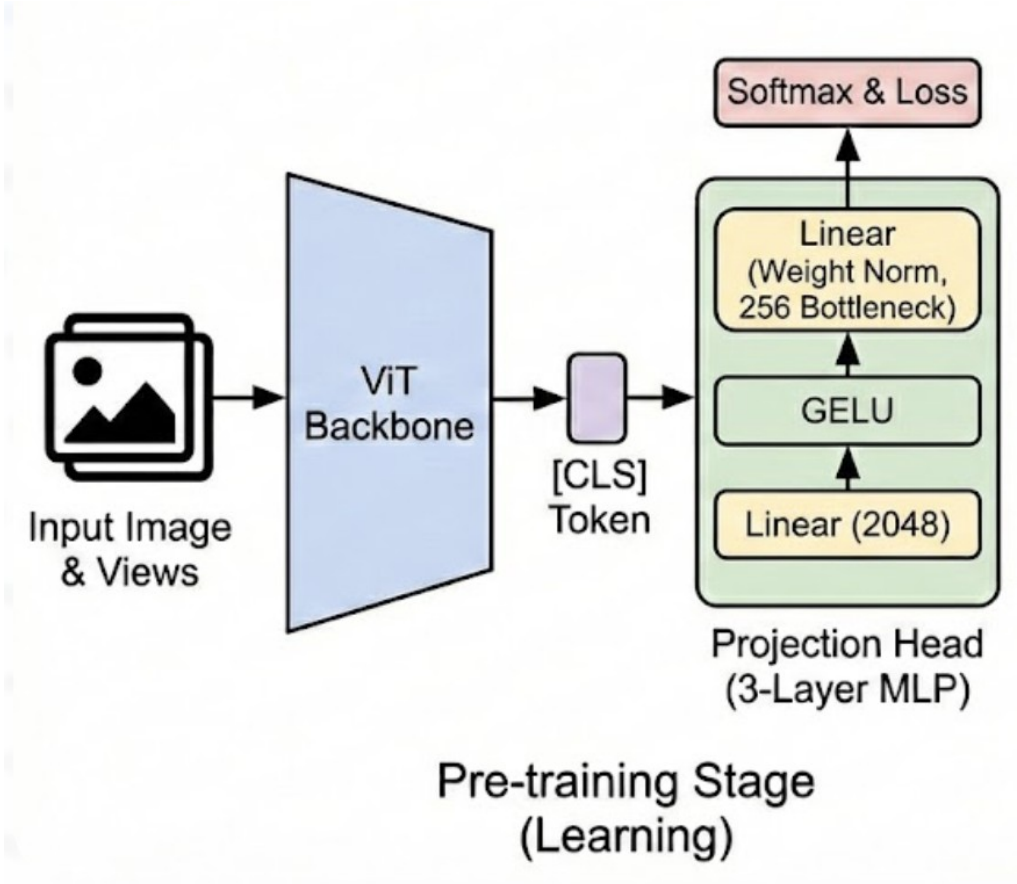
Background

- Vision Transformer(ViT)



Method

- MLP head -> Projection Head



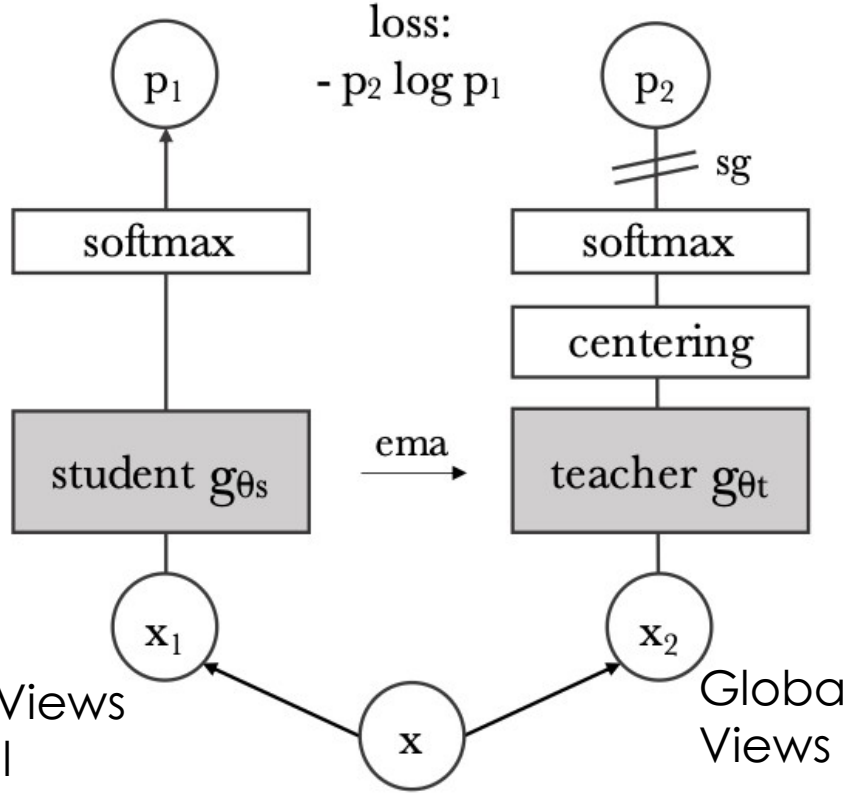
Method

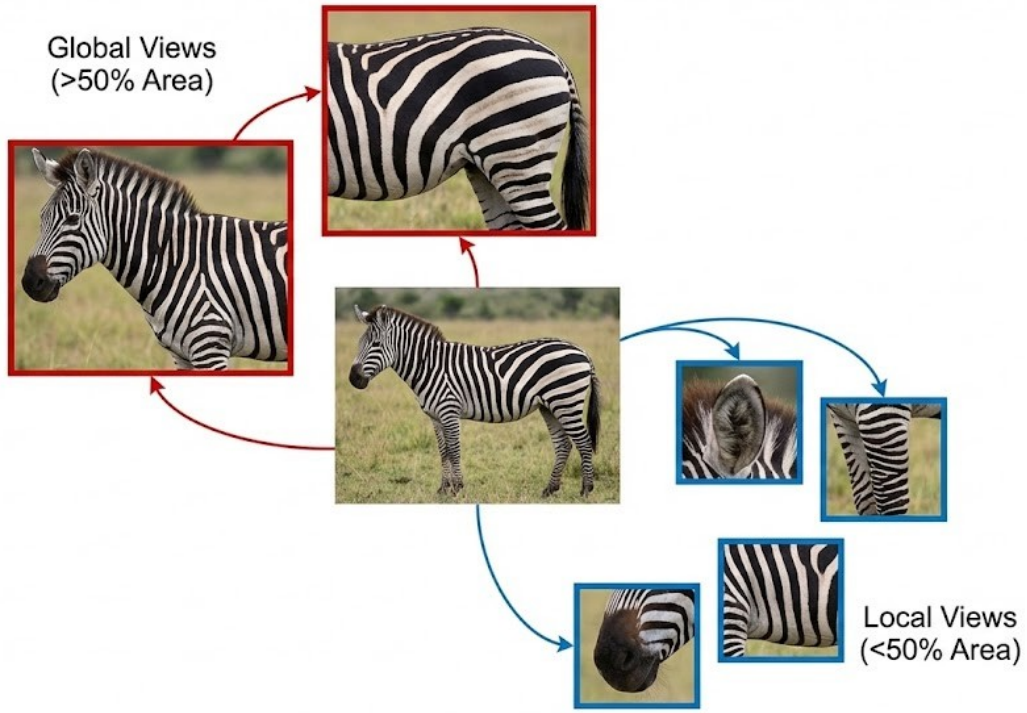
- Overview

Method

- Knowledge Distillation , Multi Crop

$$\min_{\theta_s} H(P_t(x), P_s(x)), \qquad (2)$$

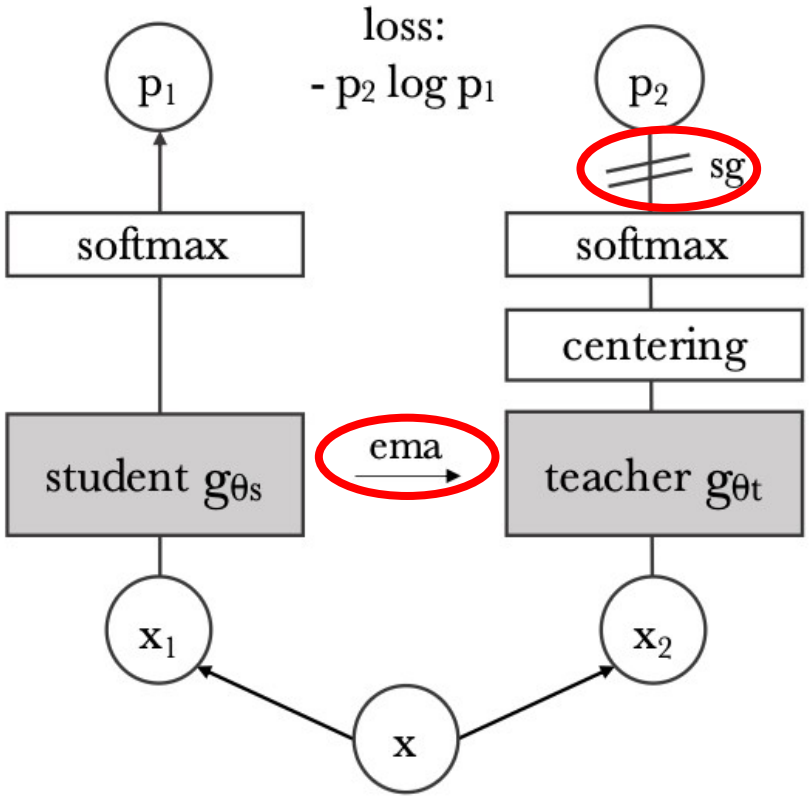


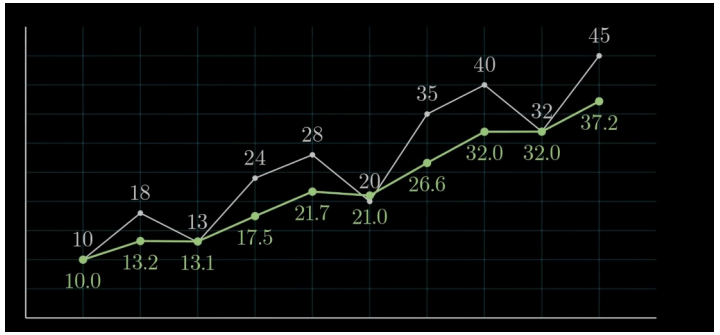


$$\min_{\theta_s} \sum_{x \in \{x_1^g, x_2^g\}} \sum_{\substack{x' \in V \\ x' \neq x}} H(P_t(x), P_s(x')). \qquad (3)$$

Method

- EMA(Exponential Moving Average)





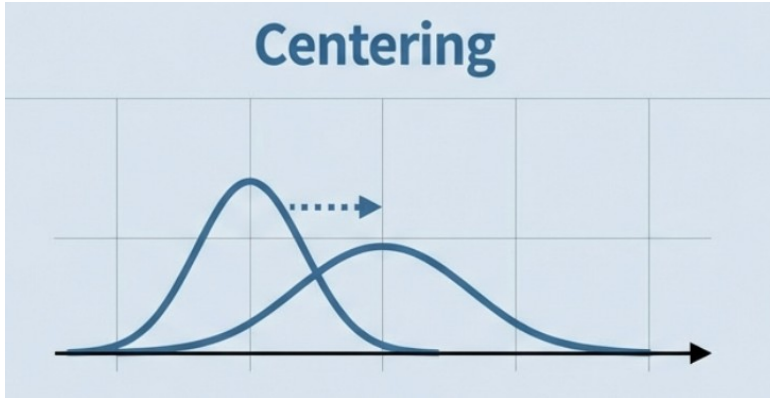
$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$$

Method

- Collapse
  - Loss를 줄일때, 제대로 학습하는 것이 아니라 지름길(shortcut)을 학습함
  - SSL(self supervised learning)의 고질적인 문제
    - 특정 차원만 지배적으로 사용하는 현상(Dominant) [해결 : Centring]
      - 입력 이미지의 무관하게 특정 차원값만 높음
      - Teacher 또한 특정 차원 값만 높음, 결국 loss는 낮아지지만 Collapse문제 발생
    - 교사 네트워크가 균등한 답만 내놓음(Uniform) [해결 : Sharpening]
      - 모든 class의 확률이 균등하게 비슷함, 결국 분포가 비슷하여 loss가 낮고 Collapse
      - Centring은 위 문제를 더 가속화
  - DINO는 Centering과 Sharpening 균형으로 해결

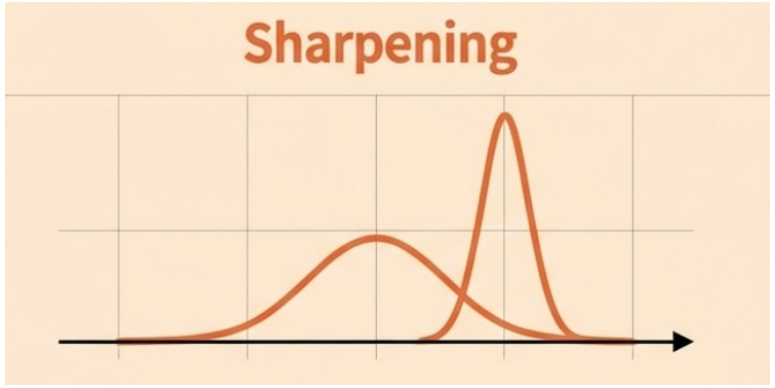
Method

- Centering, Sharpening



- Role: Prevents a specific dimension from dominating the output
- Mechanism: Computes the mean output within a batch and subtracts it from the teacher’s output $g_t(x) \leftarrow g_t(x) + c$
- Encourages the output distribution toward a **uniform distribution**

$$c \leftarrow mc + (1-m)\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}(x_i), \qquad (4)$$



- Role: Prevents the output from becoming overly uniform
- Mechanism: Sets a low Softmax temperature for the teacher model
- Effect: Sharpens the probability distribution, emphasizing strong feature

$$P_s(x)^{(i)} = \frac{\exp(g_{\theta_s}(x)^{(i)}/\tau_s)}{\sum_{k=1}^{K}\exp(g_{\theta_s}(x)^{(k)}/\tau_s)}, \qquad (1)$$

Experiments

- Linear & k-NN Classification

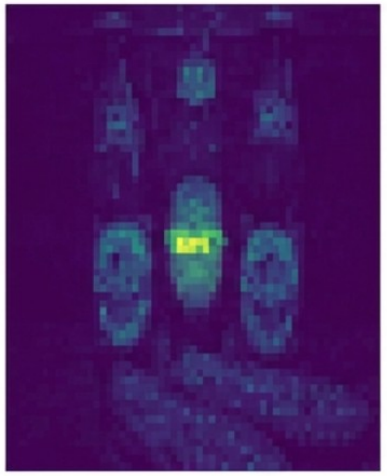
Method	Arch.	Param.	im/s	Linear	*k*-NN
Supervised	RN50	23	1237	79.3	79.3
SCLR [12]	RN50	23	1237	69.1	60.7
MoCov2 [15]	RN50	23	1237	71.1	61.9
InfoMin [67]	RN50	23	1237	73.0	65.3
BarlowT [81]	RN50	23	1237	73.2	66.0
OBoW [27]	RN50	23	1237	73.8	61.9
BYOL [30]	RN50	23	1237	74.4	64.8
DCv2 [10]	RN50	23	1237	75.2	67.1
SwAV [10]	RN50	23	1237	**75.3**	65.7
DINO	RN50	23	1237	**75.3**	**67.5**
Supervised	ViT-S	21	1007	79.8	79.8
BYOL* [30]	ViT-S	21	1007	71.4	66.6
MoCov2* [15]	ViT-S	21	1007	72.7	64.4
SwAV* [10]	ViT-S	21	1007	73.5	66.3
DINO	ViT-S	21	1007	**77.0**	**74.5**
Comparison across architectures					
SCLR [12]	RN50w4	375	117	76.8	69.3
SwAV [10]	RN50w2	93	384	77.3	67.3
BYOL [30]	RN50w2	93	384	77.4	–
DINO	ViT-B/16	85	312	78.2	76.1
SwAV [10]	RN50w5	586	76	78.5	67.1
BYOL [30]	RN50w4	375	117	78.6	–
BYOL [30]	RN200w2	250	123	79.6	73.9
DINO	ViT-S/8	21	180	79.7	**78.3**
SCLRv2 [13]	RN152w3+SK	794	46	79.8	73.1
DINO	ViT-B/8	85	63	**80.1**	77.4

Experiments

- Attention Visualization
  - Multi-head Self-Attention

Original Image

Attention Head 1:
Left Bottle



Attention Head 2:
Text Label

Attention Head 3:
Background/Shadows

Experiments

- Ablation Study

	Method	Mom.	SK	MC	Loss	Pred.	k-NN	Lin.
1	DINO	✓	✗	✓	CE	✗	72.8	76.1
2		✗	✗	✓	CE	✗	0.1	0.1
3		✓	✓	✓	CE	✗	72.2	76.0
4		✓	✗	✗	CE	✗	67.9	72.5
5		✓	✗	✓	MSE	✗	52.6	62.4
6		✓	✗	✓	CE	✓	71.8	75.6
7	BYOL	✓	✗	✗	MSE	✓	66.6	71.4
8	MoCov2	✓	✗	✗	INCE	✗	62.0	71.6
9	SwAV	✗	✓	✓	CE	✗	64.7	71.8

SK: Sinkhorn-Knopp, MC: Multi-Crop, Pred.: Predictor
CE: Cross-Entropy, MSE: Mean Square Error, INCE: InfoNCE

Experiments

- Comparison with SOTA Methods
  - Batch size

bs	128	256	512	1024
top-1	57.9	59.1	59.6	59.9

Table 9: **Effect of batch sizes.** Top-1 with k-NN for models trained for 100 epochs without multi-crop.

Conclusion

- Self-Distillation Framework
  - Proposed a simple self-distillation approach for ViT
- Emerging Properties
  - Discovered that self-supervised ViT features explicitly capture scene layouts and object boundaries.
- Superior Representation
  - Achieved state-of-the-art performance in both linear and k-NN evaluations, outperforming previous CNN-based methods.