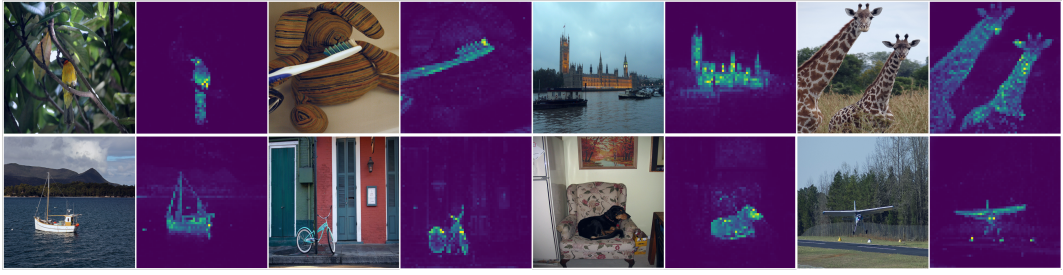
# DINO: Emerging Properties in Self-Supervised Vision Transformers

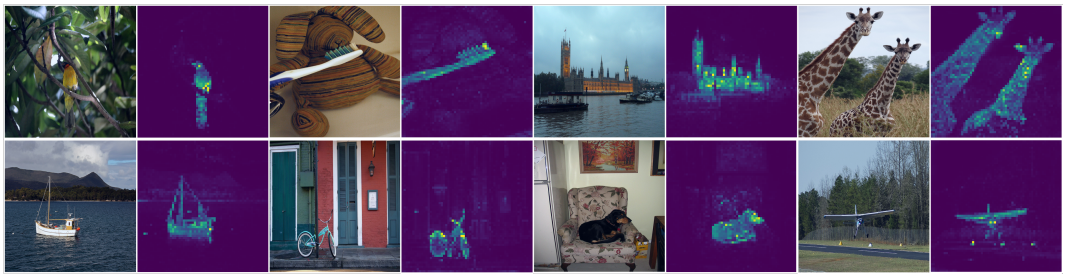
Nikita Konodyuk, Konstantin Matveev, Nikita Bashaev, Dilara Khamdeeva

Higher School of Economics

# Contributions

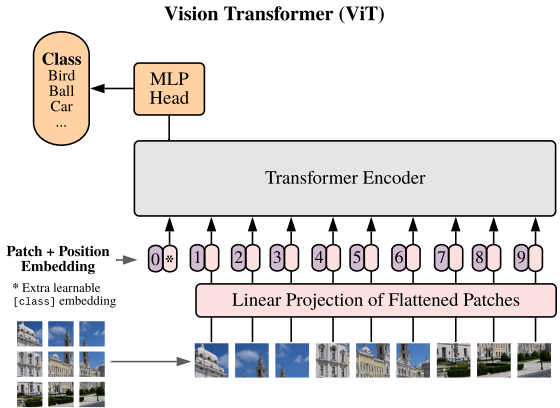


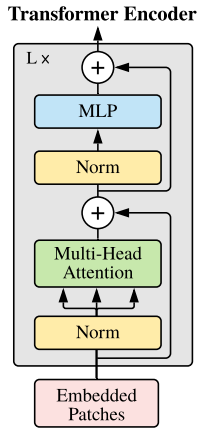
# Contributions



+ informative representations:
78% top-1 accuracy on ImageNet with just kNN

# Vision Transformer

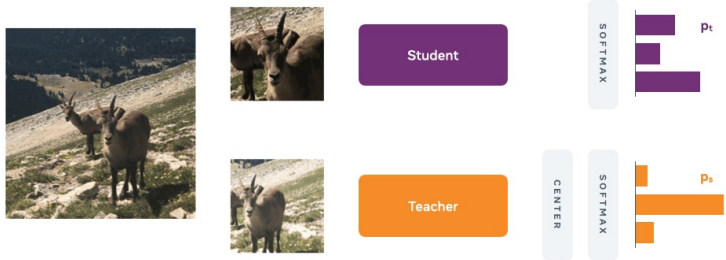


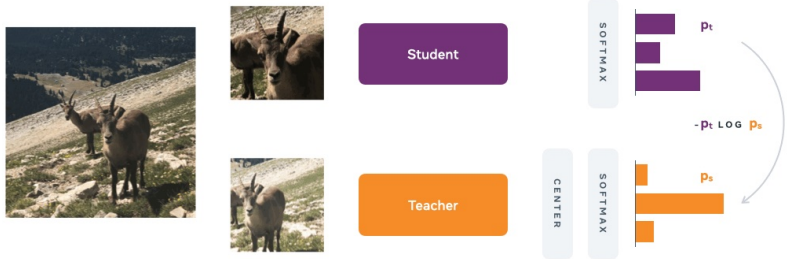
# Supervised vs Self-supervised

- Self-supervised pretraining provides richer signal
- Want something like BERT, but for CV

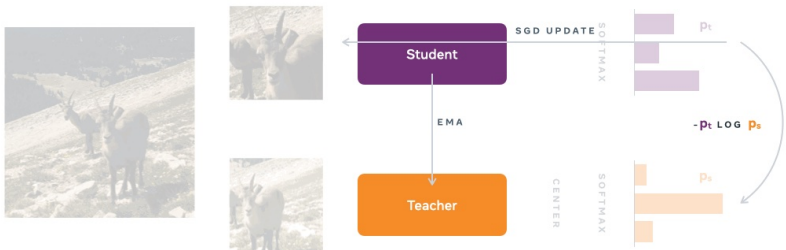
# DINO: Overview



# DINO: Overview



# DINO: Overview



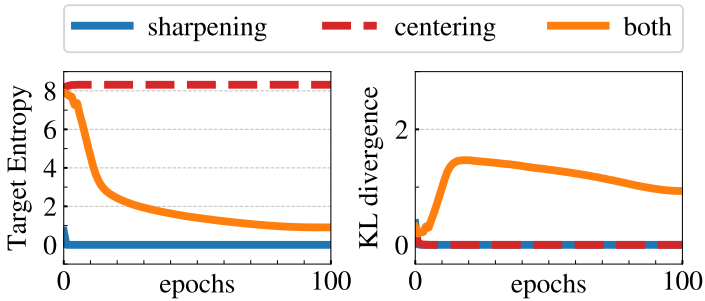
# Details

- Softmax with different temperatures
- Momentum encoder
- Avoiding collapse:
  - Centering
  - Sharpening
- Scheduling for everything:
  - Learning rate
  - Softmax temperature
  - Update rule of the teacher
- Multi-Crop

# Momentum encoder

- $\theta_t$ and $\theta_s$ are student and teacher weights
- Update rule: $\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s,$
- $\lambda$ follows a cosine schedule from 0.996 to 1
- Teacher is frozen over each epoch

# Avoiding collapse




- Teacher predictions: $g_t(x) \leftarrow g_t(x) + c$
- where $c \leftarrow mc + (1-m)\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}(x_i)$,

# Hyperparameters & Hardware

- AdamW
- $BS = 1024, lr = 0.0005 \cdot BS/256$
- $\tau_s = 0.1, \tau_t = 0.04 \rightarrow 0.07$
- ViT-S/16
- 16 GPUs for 3 days

# Benchmarks

| Method | Arch. | Param. | im/s | Linear | $k$-NN |
|---|---|---|---|---|---|
| Supervised | RN50 | 23 | 1237 | 79.3 | 79.3 |
| SCLR | RN50 | 23 | 1237 | 69.1 | 60.7 |
| MoCov2 | RN50 | 23 | 1237 | 71.1 | 61.9 |
| InfoMin | RN50 | 23 | 1237 | 73.0 | 65.3 |
| BarlowT | RN50 | 23 | 1237 | 73.2 | 66.0 |
| OBoW | RN50 | 23 | 1237 | 73.8 | 61.9 |
| BYOL | RN50 | 23 | 1237 | 74.4 | 64.8 |
| DCv2 | RN50 | 23 | 1237 | 75.2 | 67.1 |
| SwAV | RN50 | 23 | 1237 | **75.3** | 65.7 |
| DINO | RN50 | 23 | 1237 | **75.3** | **67.5** |

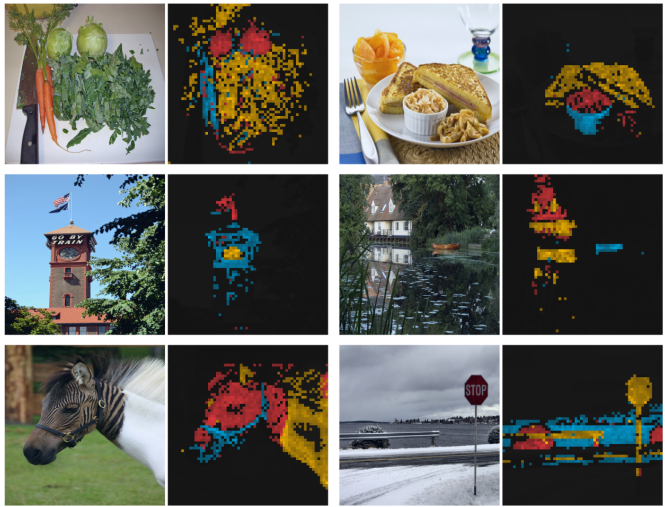
# Benchmarks

| Method | Arch. | Param. | im/s | Linear | $k$-NN |
|---|---|---|---|---|---|
| Supervised | ViT-S | 21 | 1007 | 79.8 | 79.8 |
| BYOL | ViT-S | 21 | 1007 | 71.4 | 66.6 |
| MoCov2 | ViT-S | 21 | 1007 | 72.7 | 64.4 |
| SwAV | ViT-S | 21 | 1007 | 73.5 | 66.3 |
| DINO | ViT-S | 21 | 1007 | **77.0** | **74.5** |

# Benchmarks

| Method | Arch. | Param. | im/s | Linear | $k$-NN |
|---|---|---|---|---|---|
| *Comparison across architectures* | | | | | |
| SCLR | RN50w4 | 375 | 117 | 76.8 | 69.3 |
| SwAV | RN50w2 | 93 | 384 | 77.3 | 67.3 |
| BYOL | RN50w2 | 93 | 384 | 77.4 | – |
| DINO | ViT-B/16 | 85 | 312 | 78.2 | 76.1 |
| SwAV | RN50w5 | 586 | 76 | 78.5 | 67.1 |
| BYOL | RN50w4 | 375 | 117 | 78.6 | – |
| BYOL | RN200w2 | 250 | 123 | 79.6 | 73.9 |
| DINO | ViT-S/8 | 21 | 180 | 79.7 | **78.3** |
| SCLRv2 | RN152w3+SK | 794 | 46 | 79.8 | 73.1 |
| DINO | ViT-B/8 | 85 | 63 | **80.1** | 77.4 |

# ViT Properties

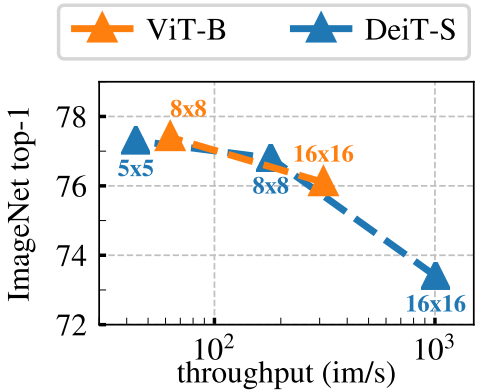


# Ablation

| | 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

# Ablation: patch size



# Ablation: batch size

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

# Other applications

- Copy detection
- Image retrieval