

Emerging Properties in Self-Supervised Vision Transformers

ICCV 2021

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou,
Julien Mairal, Piotr Bojanowski, Armand Joulin

Introduction

- Transformers have a clear advantage in NLP, but not in CV.
- To understand this reason, this paper combines ViT with self-supervised learning to explore its properties in depth.

- And found that **self-supervised ViT features** contain **explicit information** about the semantic segmentation of an image, which does **not emerge** as clearly with supervised ViTs, nor with convnets.
- After further research, a flexible SSL framework is proposed, which can be regarded as **self-distillation with no labels**. And achieve good performance in multiple downstream tasks.

DINO:
Self-Distillation with NO Labels

Self-supervised learning

- Data Augmentation
- Feature Extraction
- Similarity Loss

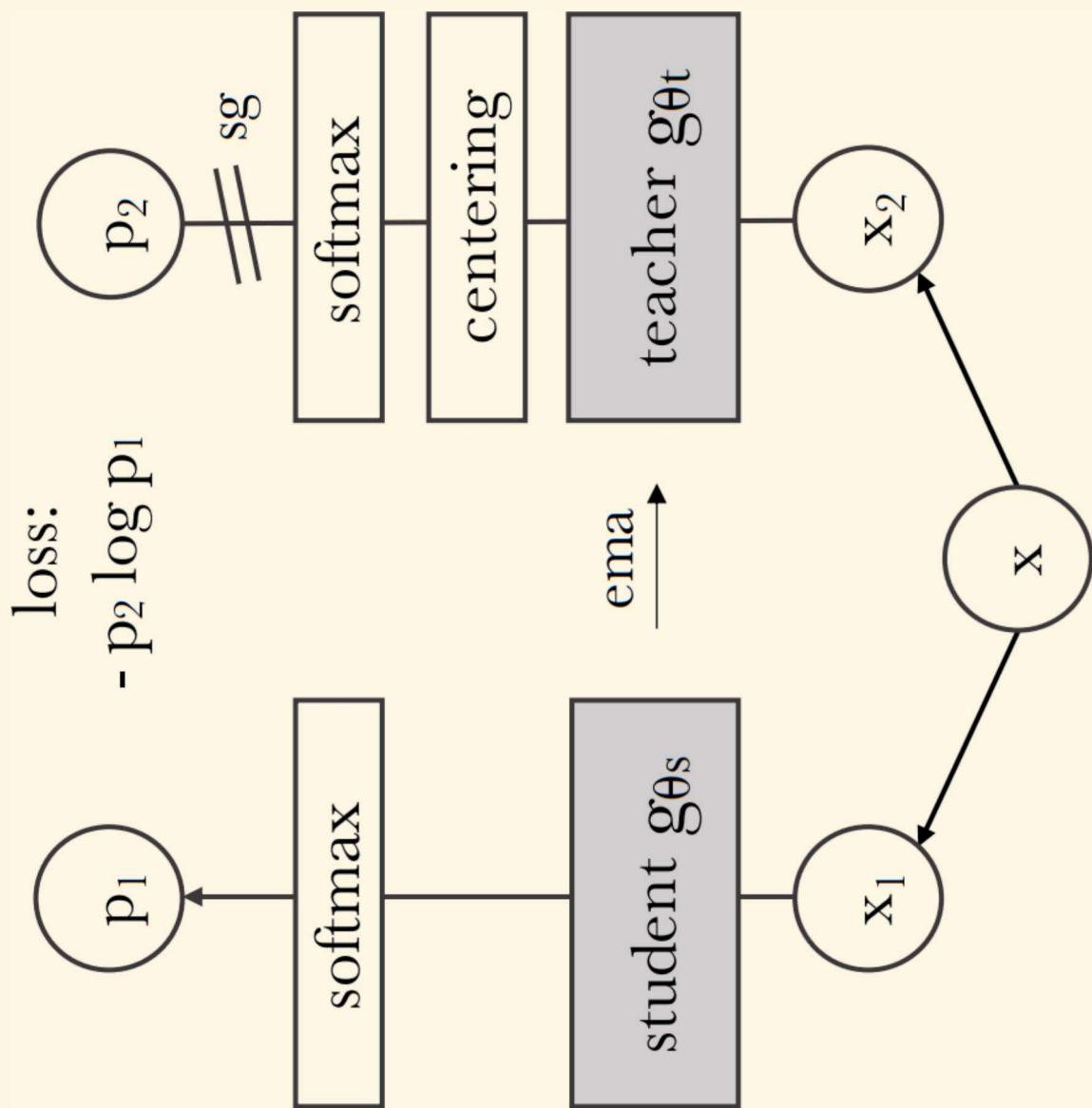
Self-supervised learning

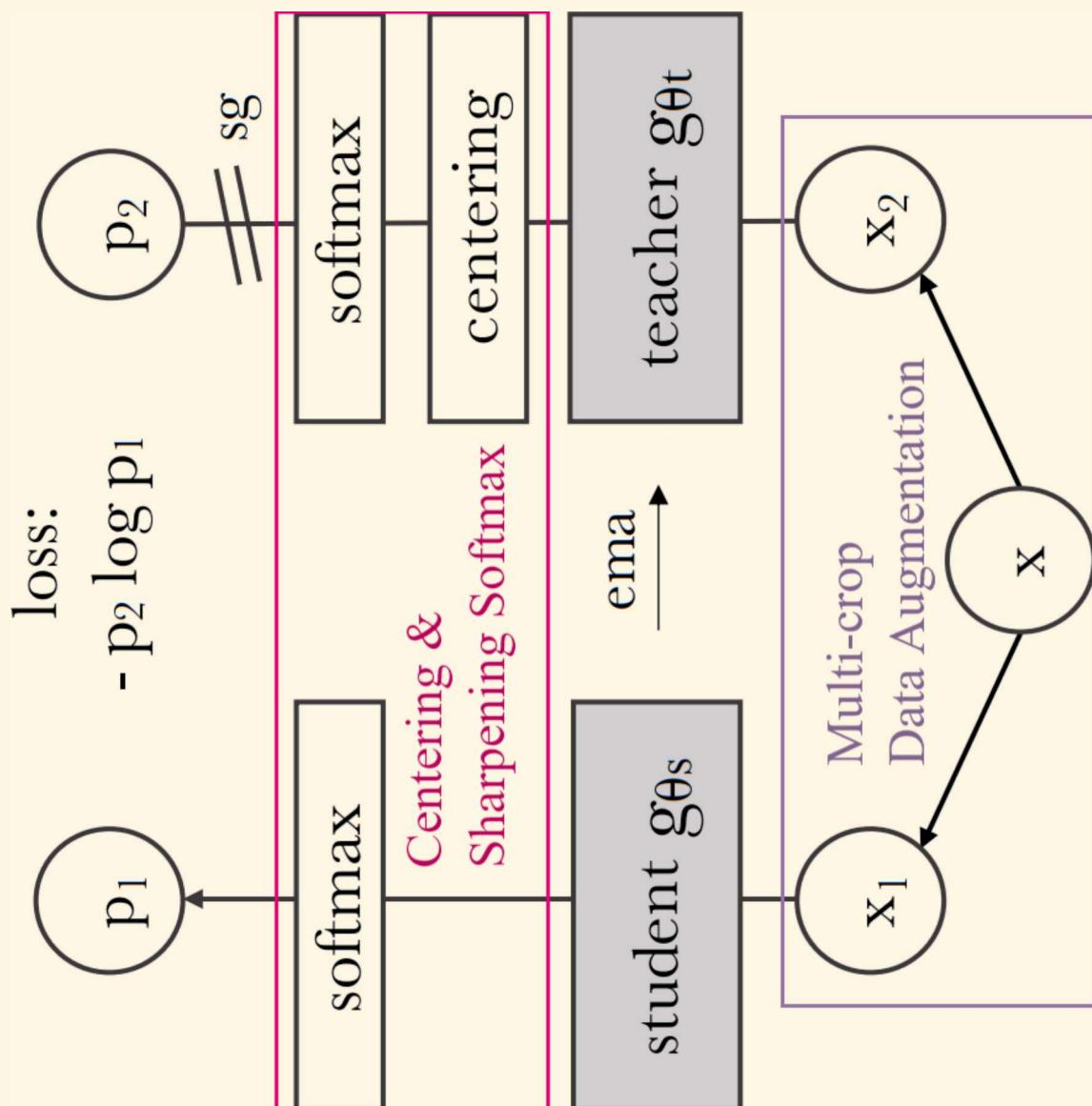
- Data Augmentation
- Feature Extraction
- ~~Similarity Loss~~

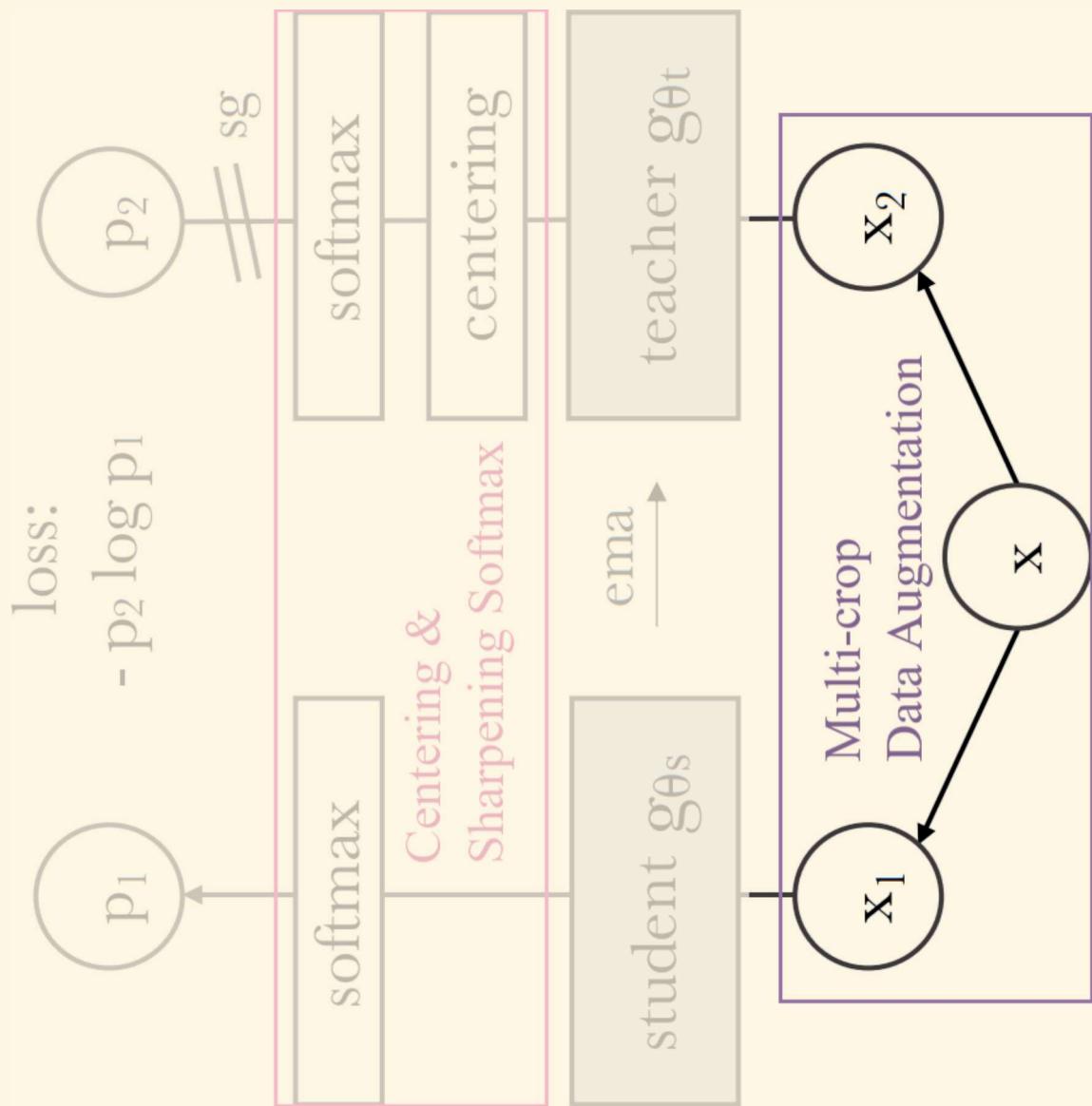
Self-supervised learning

- Data Augmentation
 - Feature Extraction
 - ~~Similarity Loss~~
- Target: dynamic pseudo labels generated by teacher

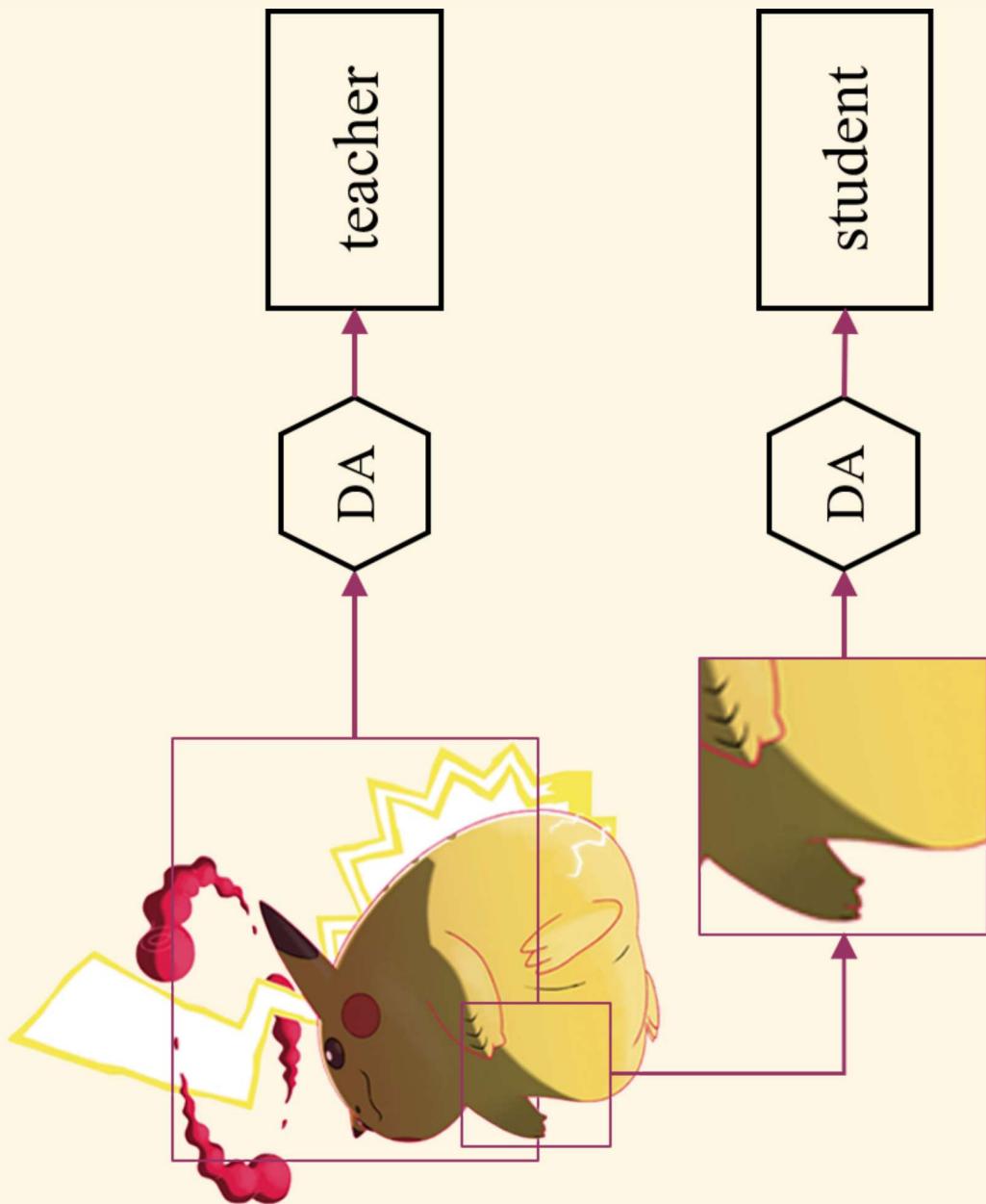
Method	MC	Center.	BN	Pred.	Mom.	Target type
BYOL			✓	✓	✓	feature
SwAV				✓		pseudo label
DINO			✓	✓	✓	pseudo label

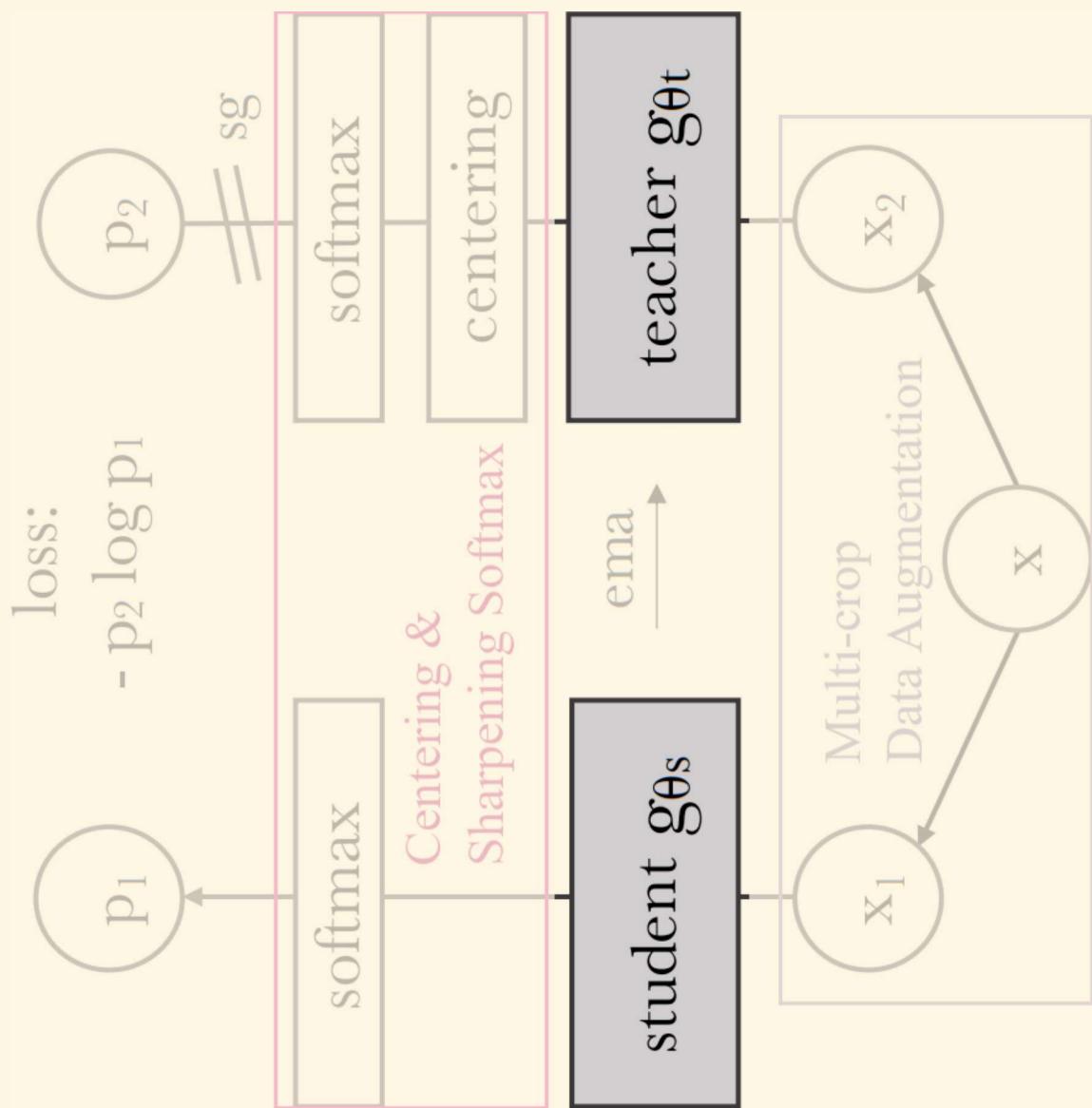






Multi-crop: local-to-global



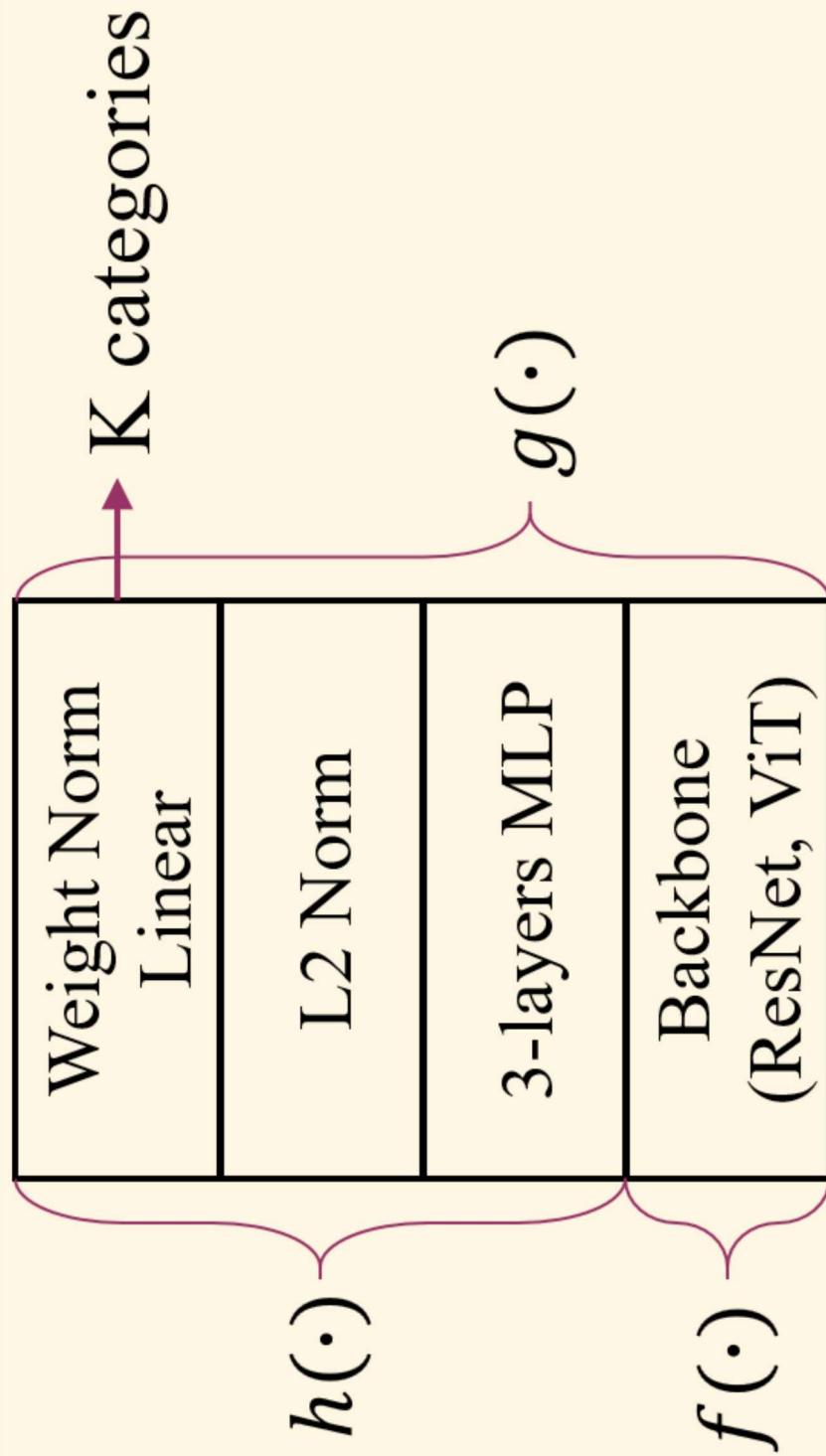


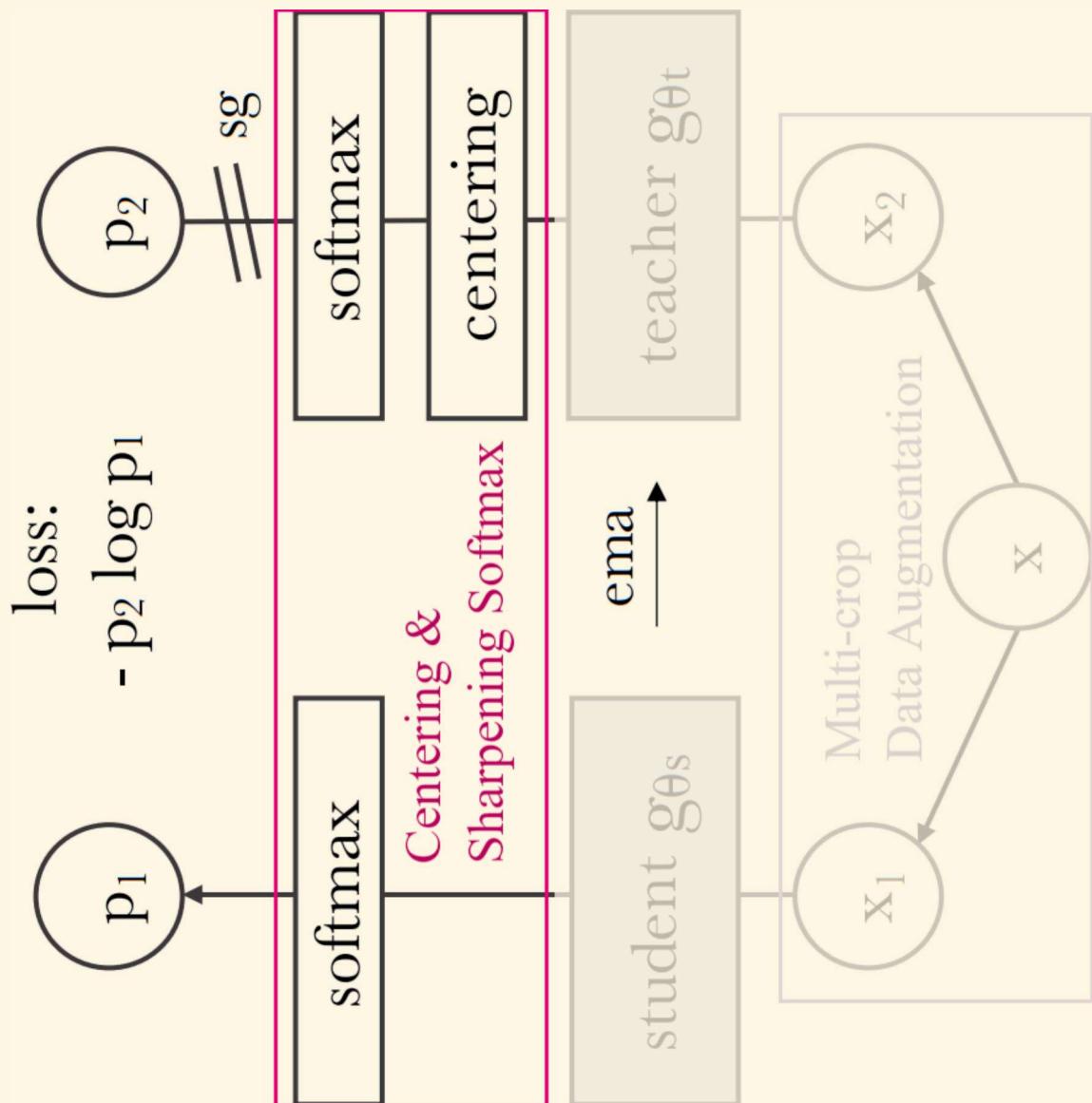
Network architecture

$$\begin{aligned}g(\cdot) &= (h \circ f)(\cdot) \\h(\cdot) &= (lin_{wn} \circ l2norm \circ mlp^N)(\cdot) \\f(\cdot) &= \text{any backbone}\end{aligned}$$

Backbone used in this paper: $\left\{ \begin{array}{l} ResNet \\ ViT (BN-free) \end{array} \right.$

Network architecture (cont.)





Avoiding collapse

- Stop Gradient
- + Momentum Encoder
- + Centering & Sharpening Cross Entropy
- Batch Norm => less dependence over the batch !

EMA: momentum encoder

when θ_s be updated

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

with λ following a cosine schedule from 0.996 to 1

Centering & Sharpening

- Centering
 $P(x) \rightarrow$ uniform distribution
- Sharpening
 $P(x) \rightarrow$ one hot

Update the average probability for each category

$$c_i \leftarrow m c^{(i)} + (1 - m) \frac{1}{B} \sum_n^B g_{\theta_t}(x_n)^{(i)}, i = 1..K$$

$$\mathcal{L} = CrossEntropy(P_s(x_{v1}), SG[P_t(x_{v2})])$$

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

$$P_t^{(i)}(x) = \frac{exp((g_{\theta_s}(x)^{(i)} - c^{(i)})/\tau_t)}{\sum_{k=1}^K exp((g_{\theta_s}(x)^{(k)} - c^{(k)})/\tau_t)}$$

Main Results

Linear and k-NN classification on ImageNet

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

Linear and k-NN classification on ImageNet (cont.)

Method	Arch.	Param.	im/s	Linear	k -NN
<i>Comparison across architectures</i>					
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

Segmentations from supervised versus DINO

Supervised



DINO



Random Supervised DINO

ViT-S/16	22.0	27.3	45.9
ViT-S/8	21.8	23.7	44.7

Image retrieval

Pretrain	Arch.	Pretrain	ROx			RPar		
			M	H	M	H	M	H
Sup. [57]	RN101+R-MAC	ImNet	49.8	18.5	74.0	52.1		
Sup.	ViT-S/16	ImNet	33.5	8.9	63.0	37.2		
DINO	ResNet-50	ImNet	35.4	11.1	55.9	27.5		
DINO	ViT-S/16	ImNet	41.8	13.7	63.1	34.4		
DINO	ViT-S/16	GLDv2	51.5	24.3	75.3	51.6		

Freeze DINO's weights and perform retrieval with knn.

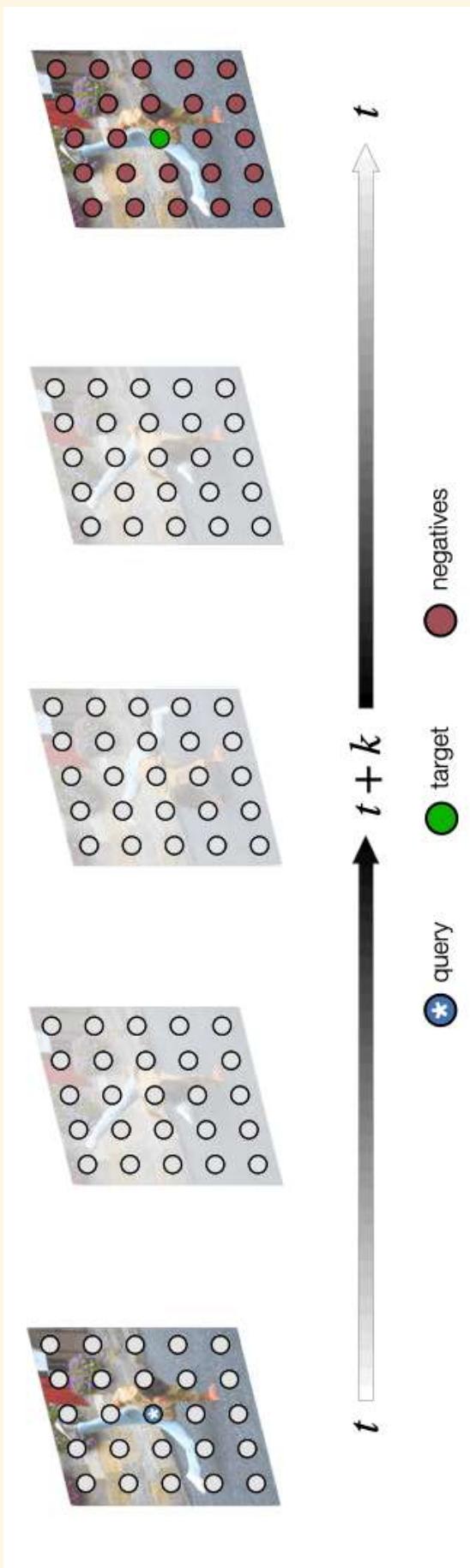
Copy detection

Method	Arch.	Dim.	Resolution	mAP
Multigrain [5]	ResNet-50	2048	224^2	75.1
Multigrain [5]	ResNet-50	2048	largest side 800	82.5
Supervised [69]	ViT-B/16	1536	224^2	76.4
DINO	ViT-B/16	1536	224^2	81.7
DINO	ViT-B/8	1536	320^2	85.5

DAVIS 2017 Video object segmentation

Method	Data	Arch.	$(\mathcal{J} \& \mathcal{F})_m$	\mathcal{J}_m	\mathcal{F}_m
<i>Supervised</i>					
ImageNet	INet	ViT-S/8	66.0	63.9	68.1
STM [48]	I/D/Y	RN50	81.8	79.2	84.3
<i>Self-supervised</i>					
CT [71]	VLOG	RN50	48.7	46.4	50.0
MAST [40]	YT-VOS	RN18	65.5	63.3	67.6
STC [37]	Kinetics	RN18	67.6	64.8	70.2
DINO	INet	ViT-S/16	61.8	60.2	63.4
DINO	INet	ViT-B/16	62.3	60.7	63.9
DINO	INet	ViT-S/8	69.9	66.6	73.1
DINO	INet	ViT-B/8	71.4	67.9	74.9

Nor finetune any weights for the task.



Space-Time Correspondence as a Contrastive Random Walk

Segment scenes with a nearest neighbor between consecutive frames.

Transfer learning

	Cifar ₁₀	Cifar ₁₀₀	INat ₁₈	INat ₁₉	Flwrs	Cars	INet
<hr/>							
<i>ViT-S/16</i>							
Sup. [69]	99.0	89.5	70.7	76.6	98.2	92.1	79.9
DINO	99.0	90.5	72.0	78.2	98.5	93.0	81.5
<hr/>							
<i>ViT-B/16</i>							
Sup. [69]	99.0	90.8	73.2	77.7	98.4	92.1	81.8
DINO	99.1	91.7	72.6	78.6	98.8	93.0	82.8

Ablation Study

Method	Mom.	SK	MC	Loss	Pred.	k -NN	Lin.
1 DINO	✓	✗	✓	✓	CE	✗	72.8
2	✗	✗	✓		CE	✗	0.1
3	✓	✓	✓		CE	✗	72.2
4	✓	✗	✗		CE	✗	67.9
5	✓	✗	✓		MSE	✗	52.6
6	✓	✗	✓		CE	✓	71.8
							75.6

Mom.: Momentum Encoder, SK: Sinkhorn-Knopp

MC: Multi-Crop, CE: Cross-Entropy, Pred.: Predictor

MSE: Mean Square Error

Method	Mom.	SK	MC	Loss	Pred.	k -NN	Lin.
1 DINO	✓	✗	✓	✓	CE	✗	76.1
2	✗	✗	✓	✓	CE	✗	0.1
3	✓	✓	✓	✓	CE	✗	76.0
4	✓	✗	✗	✓	CE	✗	67.9
5	✓	✗	✓	✓	MSE	✗	72.5
6	✓	✗	✓	✓	CE	✓	52.6
							62.4
							71.8
							75.6

Mom.: Momentum Encoder, SK: Sinkhorn-Knopp

MC: Multi-Crop, CE: Cross-Entropy, Pred.: Predictor

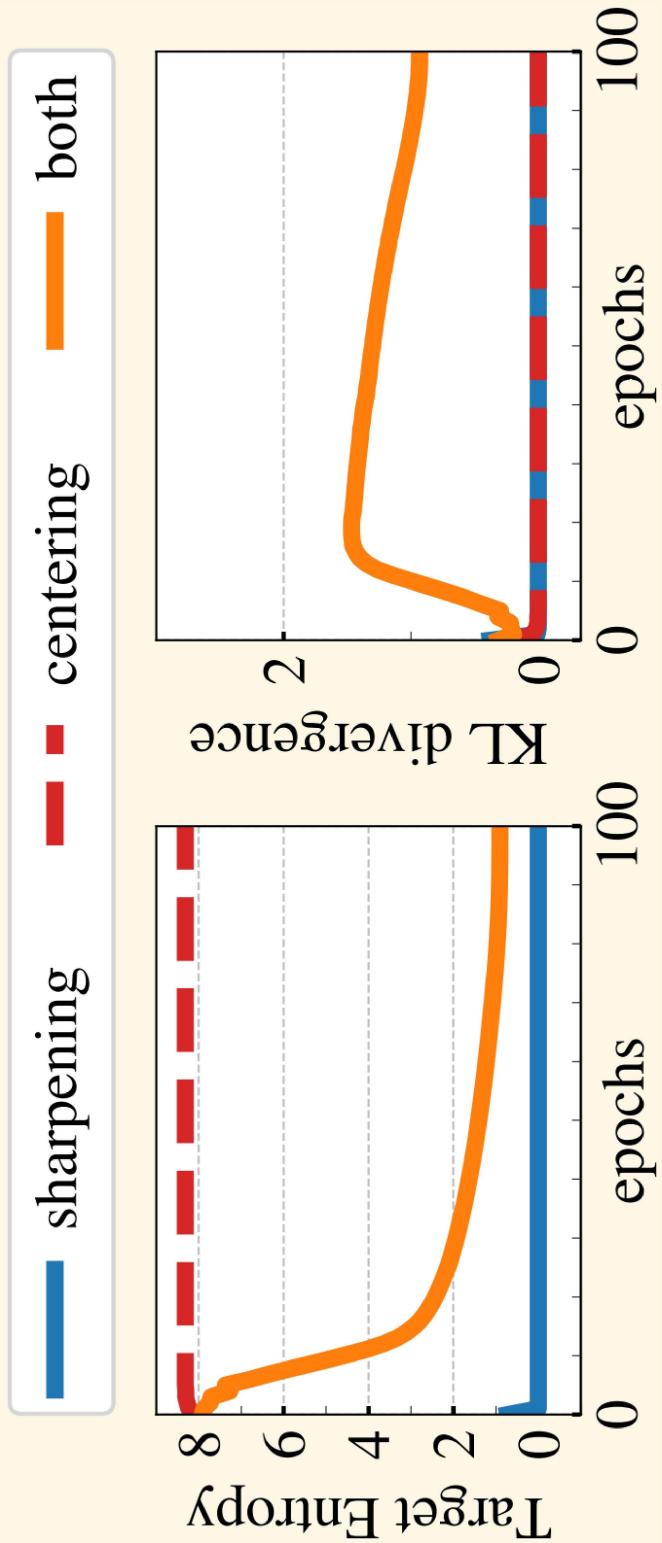
MSE: Mean Square Error

ImageNet validation with k-NN classifier.



The performance of the momentum teacher and the student.
Comparison between different types of teacher network.

$$H(P_t, P_s) = h(P_t) + D_{KL}(P_t | P_s)$$



- Only centering, the entropy $h(P_t) \rightarrow -\ln(\frac{1}{K})$.
- In contrast, $h(P_t) \rightarrow 0$.

Multi-crop in different self-supervised frameworks.

crops	2×224^2		$2 \times 224^2 + 6 \times 96^2$	
eval	k -NN	linear	k -NN	linear
BYOL	66.6	71.4	59.8	64.8
SwAV	60.5	68.5	64.7	71.8
MoCo-v2	62.0	71.6	65.4	73.4
DINO	67.9	72.5	72.7	75.9

Conclusion

- ~~KNN is all you need when you have good features.~~
- Using pseudo labels as learning targets allows the model to learn more features.
- ViT has the potential that cannot be realized with supervised learning.

Thanks for your attention.

Q&A