Variants of Vision Transformer

DINO

Emerging Properties in Self-supervised Vision Transformers

https://arxiv.org/pdf/2104.14294

Vahid Mirjalili (https://vmirly.github.io)

PyML Studio
Vision Transformers Series

<u>DINO</u>: Self-**di**stillation with **no** labels

Objective: Self-supervised pre-training

A simplified way to apply self-supervised learning
Self-supervised training provides richer learning signal

Emergent Properties of Self-Supervised ViT

- Self-supervised ViT features contain explicit information for <u>semantic segmentation</u>
- Self-supervised ViT features are excellent k-NN classifiers

https://arxiv.org/pdf/2104.14294

Self-Training vs. Knowledge Distillation

Self-Training

- ➢ Using an initial set of labeled data, learn features to improve them by incorporating a larger unlabeled dataset

(aka semi-supervised learning)

Knowledge Distillation

- ➢ Transferring knowledge from a trained model (teacher) to another (student)

Noisy Student

Propagate soft pseudo-labels to an unlabeled dataset using <u>knowledge distillation</u> in a <u>self-training</u> framework

Self-Supervised Learning (SSL) Approaches

Instance Classification

Treats each image as a different class and train a model to discriminate between classes
➔
Drawback: does not scale well with the number of images

BYOL

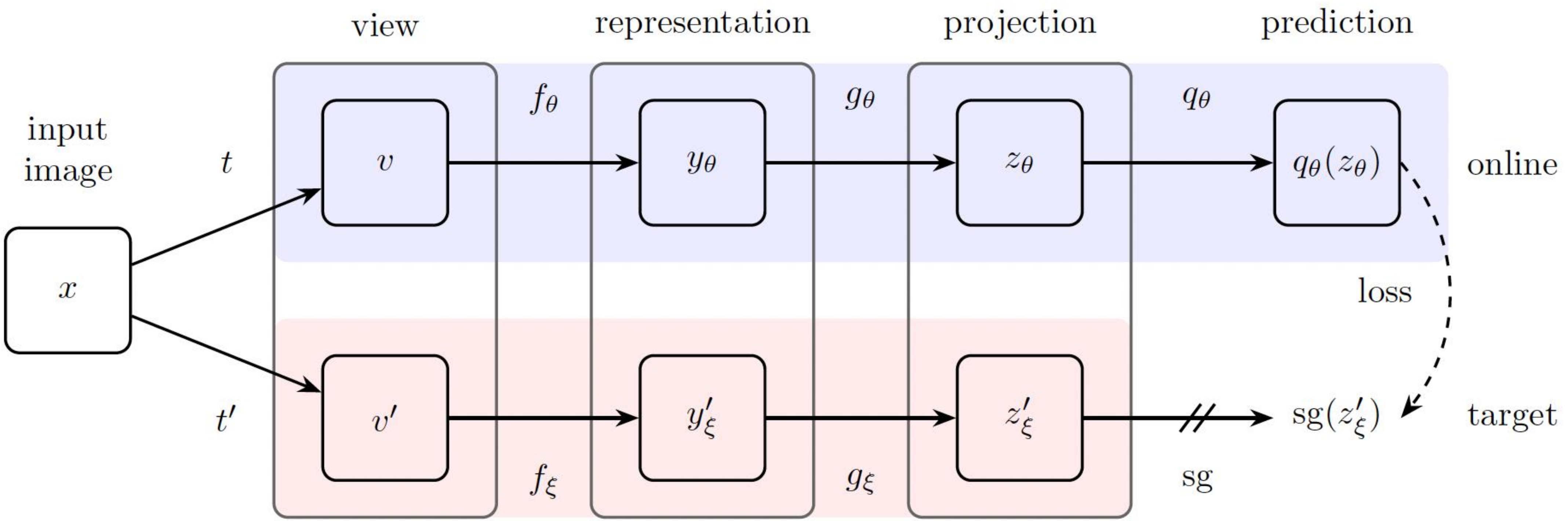
A metric-learning formulation with two networks: online & target
➔

- Generate two different augmented view of the input image and train the online network to match the target representations
- Refine target network using exponential moving-average of online network

DINO

Inspired by BYOL, but using a different similarity loss, and simultaneous training of student and teacher

BYOL

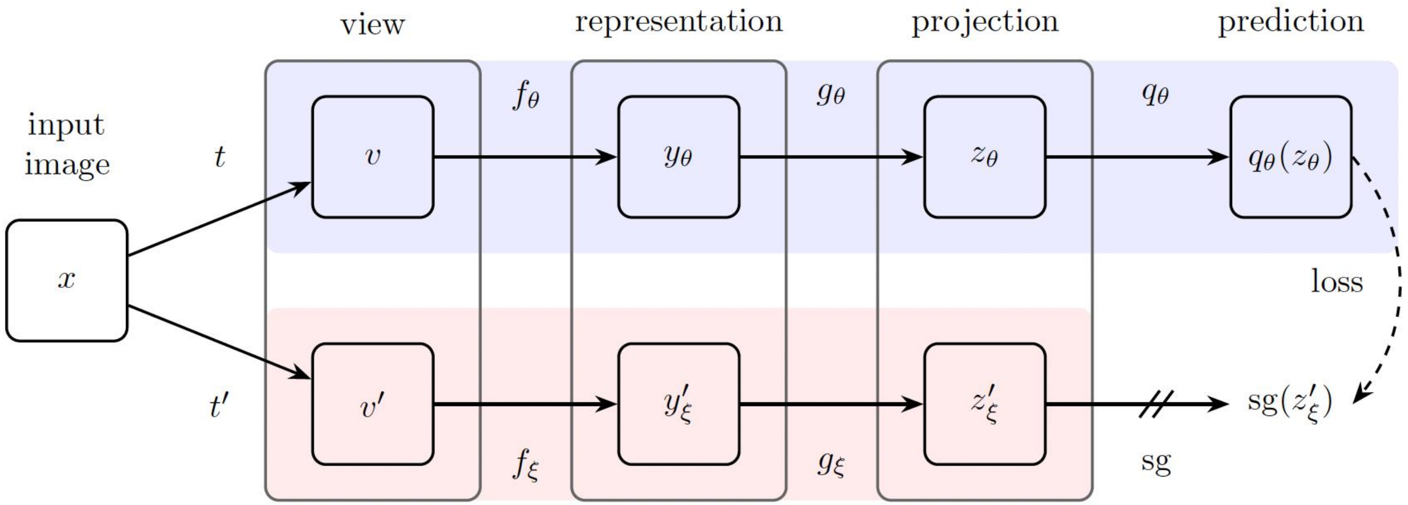


Online network: encoder f_θ, projector g_θ, predictor q_θ

Target network: encoder f_ζ, projector g_ζ

Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, 2020, https://arxiv.org/pdf/2006.07733.pdf

BYOL



Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning, 2020, https://arxiv.org/pdf/2006.07733.pdf

Loss function for the online network:

Mean square error between
l2-normalized $\overline{q_\theta}(z_\theta)$ and
l2-normalized $\overline{z'_\zeta}$

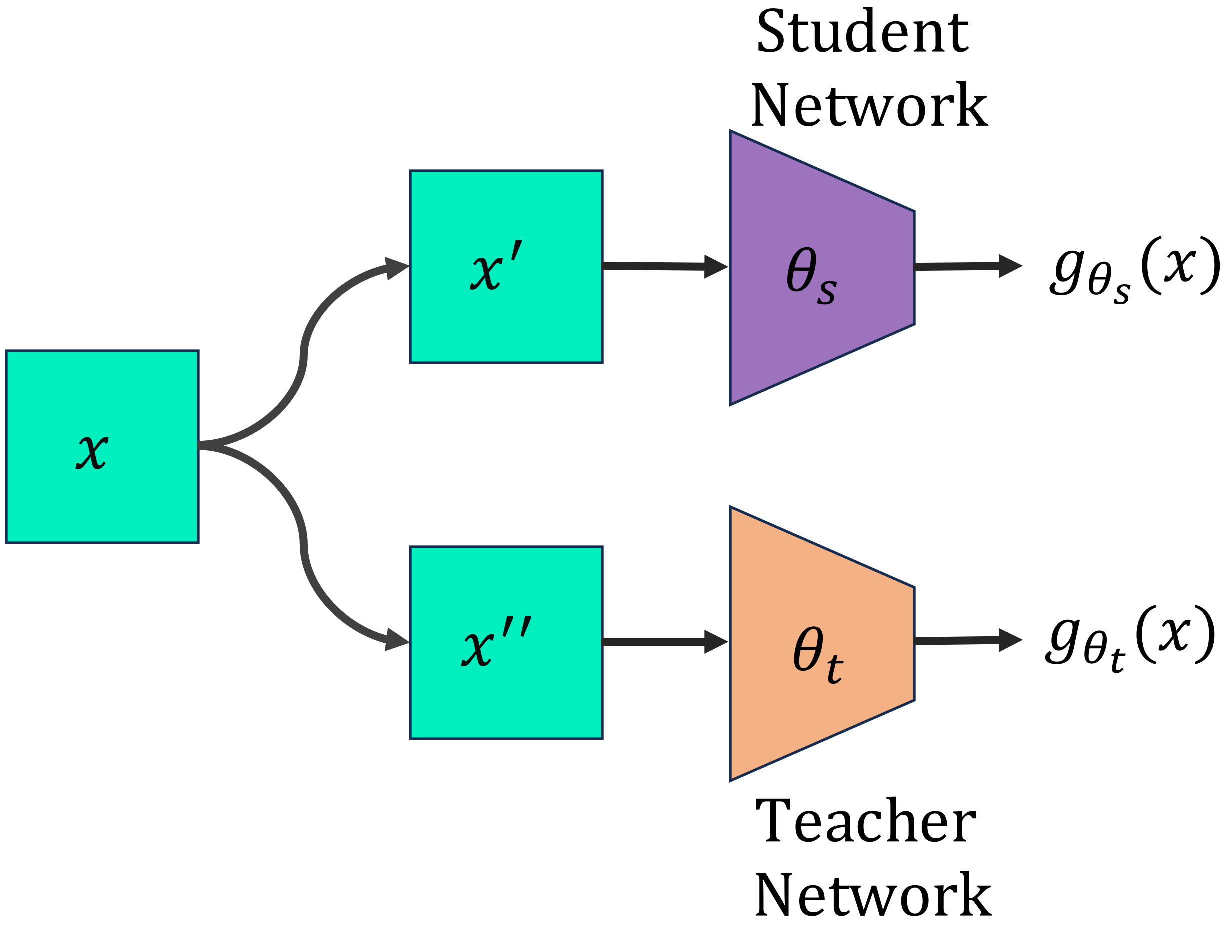
$$\mathcal{L}_{\theta,\zeta} = \left\| \overline{q_\theta}(z_\theta) - \overline{z'_\zeta} \right\|_2$$

Updating the target network:

Slowly moving average of θ

$$\zeta = \lambda\zeta + (1 - \lambda)\theta$$

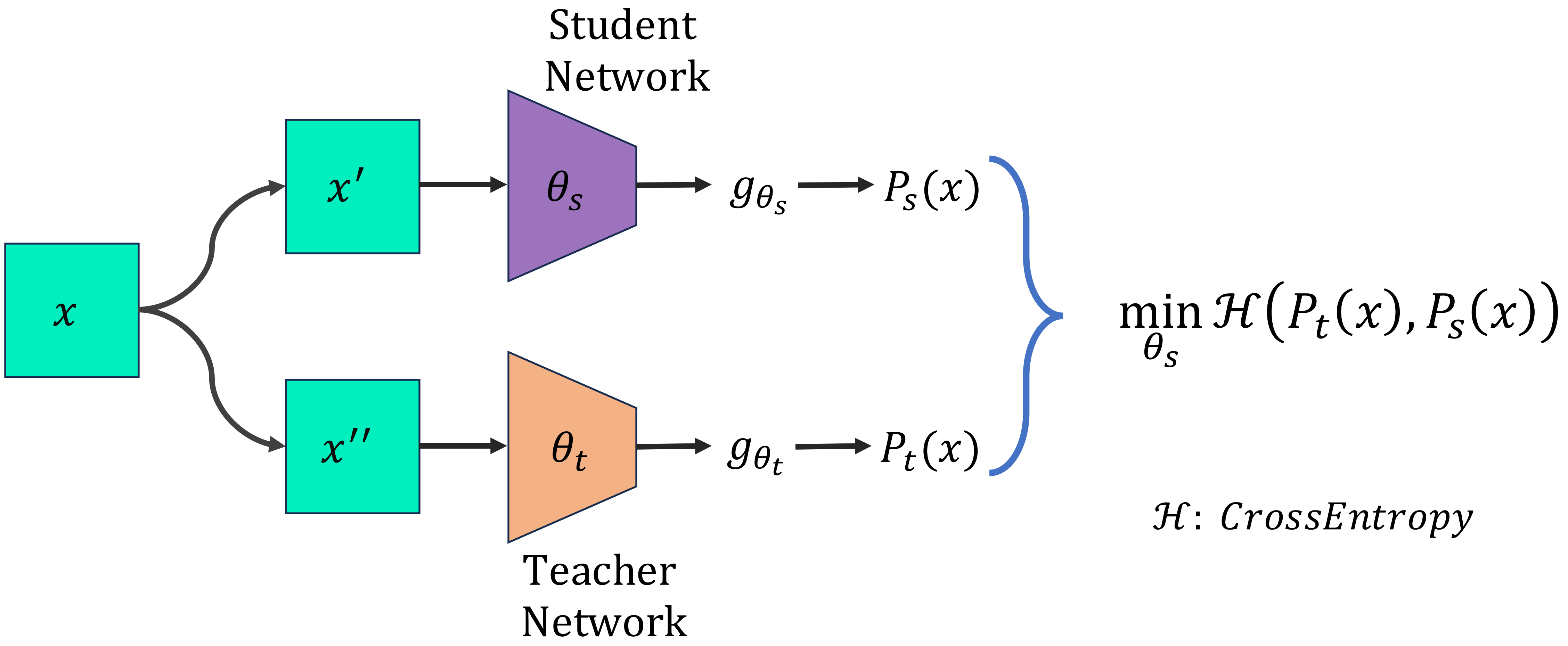
Intro to DINO Framework



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

$$P_t(x)^i = \frac{\exp\left(g_{\theta_t}(x)^i/\tau_t\right)}{\sum_{k=1}^{K}\exp\left(g_{\theta_t}(x)^k/\tau_t\right)}$$

DINO: Training the Student Network



DINO: SSL with Knowledge Distillation

- Generate a set of views V of the input image x containing:
  - Two global views x_1^g and x_2^g
  - Several local views with smaller resolutions

Global views:
Crops of 224×224, covering more than 50% of the original image

Local views:
Crops of 96×96, covering less than 50% of the original image

- The global views are passed to the teacher network
- The local views are passed to the student network

$$\min_{\theta_s} \sum_{x \in \{x_1^g, x_2^g\}} \sum_{\substack{x' \in V \\ -\{x_1^g, x_2^g\}}} \mathcal{H}\left(P_t(x), P_s(x)\right)$$

Teacher Network

- Built from the past iterations of the student network
- Freezing g_{θ_t} over the current training epoch
- Update rule: using momentum encoder

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

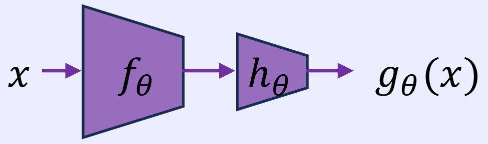
→ Mean teacher
→ Model averaging and ensembling effect
→ Better performance than the student

Network Architecture

- Backbone f_θ (ViT or ResNet)

- Projection head: h_θ
  - 3-layer MLP with hidden dimension 2048
  - l2-normalization
  - Weight-normalized fully connected layer with K dimensions

➔ $g = h \circ f = h\big(f(x)\big)$



Backbone

model	blocks	dim	heads	#tokens	#params	im/s
ResNet-50	–	2048	–	–	23M	1237
ViT-S/16	12	384	6	197	21M	1007
ViT-S/8	12	384	6	785	21M	180
ViT-B/16	12	768	12	197	85M	312
ViT-B/8	12	768	12	785	85M	63

- ➢ Student and teacher have the same exact architecture
- ➢ Complete BN-free when using ViT as backbone

Avoiding Collapse

Centering and sharpening

- ➢ Centering prevents one dimension to dominate while encouraging collapse to uniform distribution
- ➢ Sharpening has the opposite effect of centering

➔ Applying both <u>centering</u> and <u>sharpening</u> effectively prevents collapse

Experiments

Training

- Pre-train on ImageNet dataset without labels (DINO)
  - AdamW optimizer
  - Learning-rate warmup

Evaluation Protocol

- Linear evaluation
  - Train with random resize crops and horizontal flip, evaluate on center crop
- k-NN evaluation (with k=20)
- Finetuning

Results: linear and k-NN

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**

Comparing across SSL frameworks

- Using ResNet-50: DINO shows on par performance
  ➔ DINO works in standard settings
- Using ViT-S: DINO outperforms other SSL methods by at least 3.5%

https://arxiv.org/pdf/2104.14294

Properties of ViTs Trained with SSL

1. Nearest Neighbor Retrieval

- Image retrieval
  - Evaluated on Oxford and Paris image retrieval datasets
  - DINO outperforms models trained with labels (supervised)

			$\mathcal{R}$Ox		$\mathcal{R}$Par	
Pretrain	Arch.	Pretrain	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

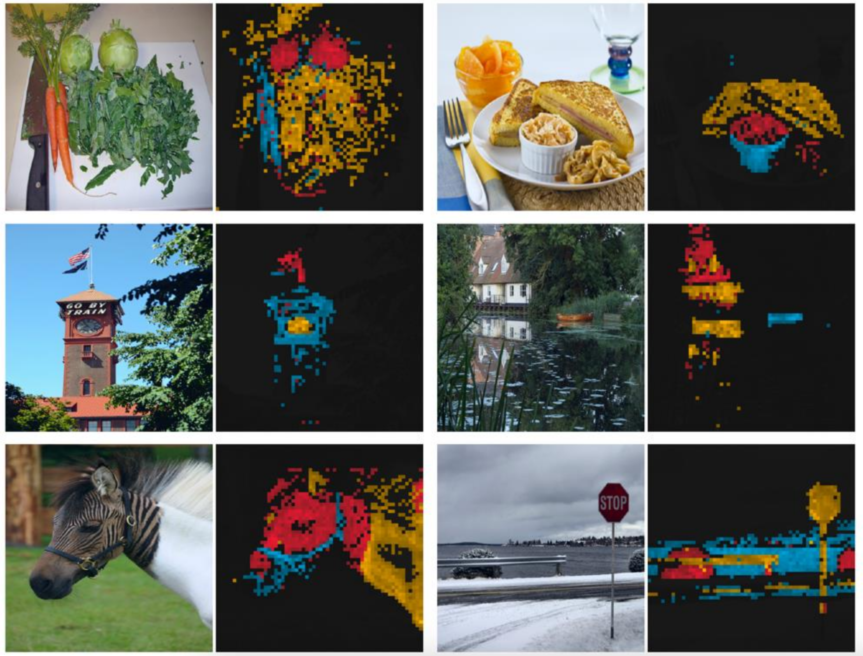
- Copy detection
  - INRIA Copydays dataset

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**

https://arxiv.org/pdf/2104.14294

Properties of ViTs Trained with SSL

2. Discovering Semantic Layouts



Method	Data	Arch.	$(\mathcal{J}\&\mathcal{F})_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Supervised					
ImageNet	INet	ViT-S/8	66.0	63.9	68.1
STM [48]	I/D/Y	RN50	81.8	79.2	84.3
Self-supervised					
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**

- Probing self-attention maps

- Video instance segmentation
  - DAVIS-2017 dataset
  - Without finetuning

https://arxiv.org/pdf/2104.14294

DINO: self-distillation with no labels

- Pre-train models (ViT or CNN) on unlabeled data
- Identified two key properties for pre-train ViTs with SSL
  - High quality features for k-NN classification, useful for image retrieval application
  - Discovering semantic layout, useful for weakly supervised semantic segmentation
- Developing BERT-like model with ViT and SSL pre-trainin

Next Video: CLIP

Thanks for watching