FLSL: Feature Level Self-supervise Learning

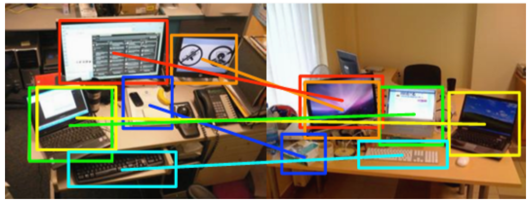
Qing Su[1], Anton Netchaev[2], Hai Li[3], and Shihao Ji[1]
[1]Georgia State University, [2]U.S. Army ERDC, [3]Duke University

Paper link: https://arxiv.org/abs/2306.06203

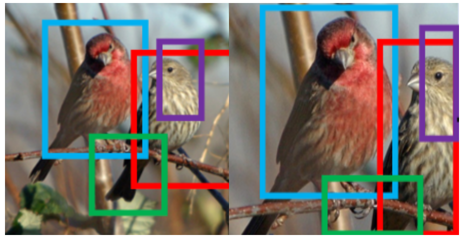
NEURAL INFORMATION PROCESSING SYSTEMS

Current Self-Supervised Learning (SSL) for Dense Prediction

RoI-based

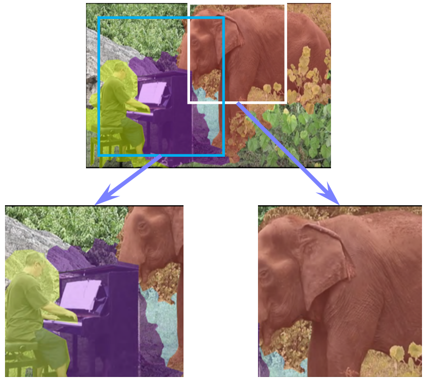


ORL [Xie, J. et la.]

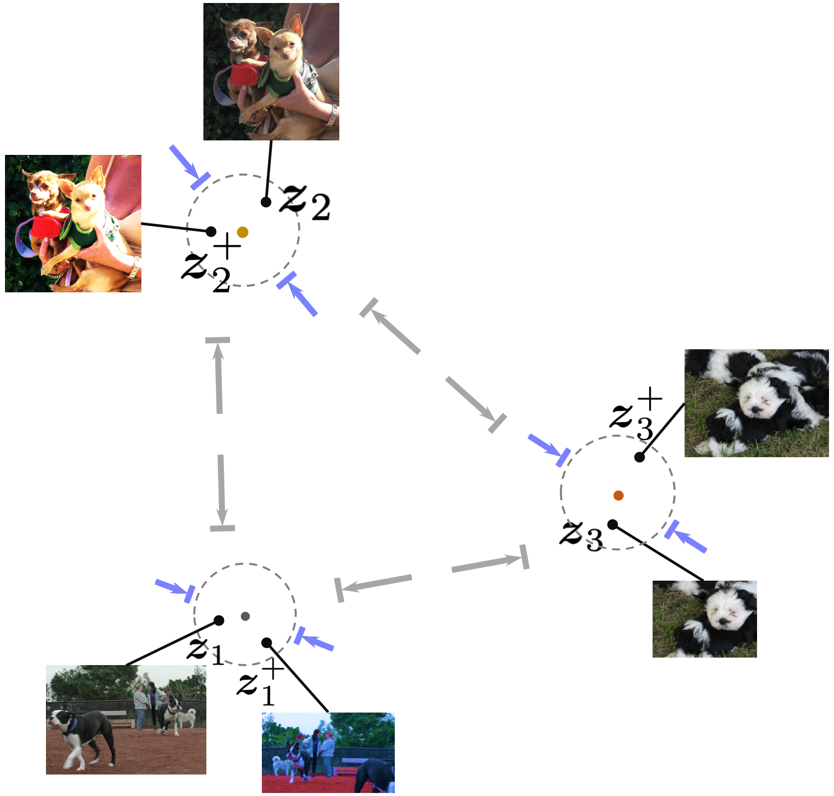


SoCo [Wei, F. et la.]

Cluster-based



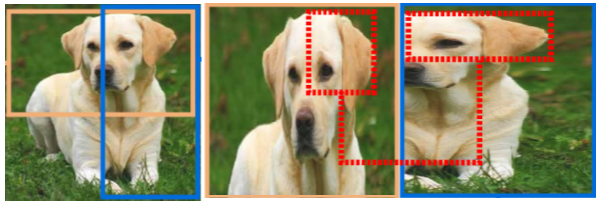
DetCon [Hénaff, O.J et la.]



Explicit / Implicit clustering

Joint-embedding

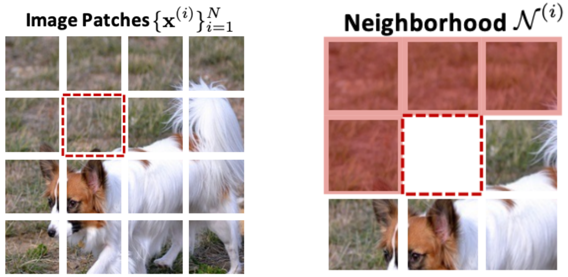
Overlap-based



ReSim [Xiao, T. et la.]

Patch-based

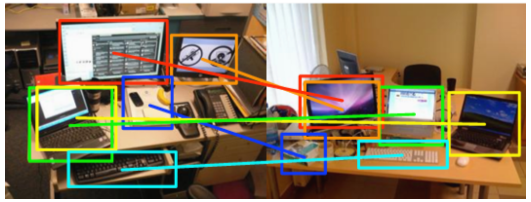
DetCo [Xie, E. et la.]



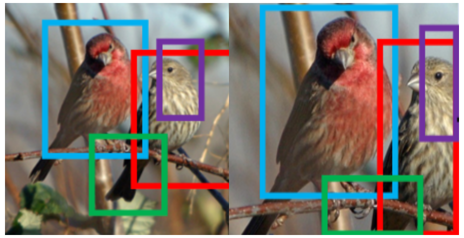
DINO + SelfPatch [Yun, S. et la.]

Current SSL for Dense Prediction

RoI-based

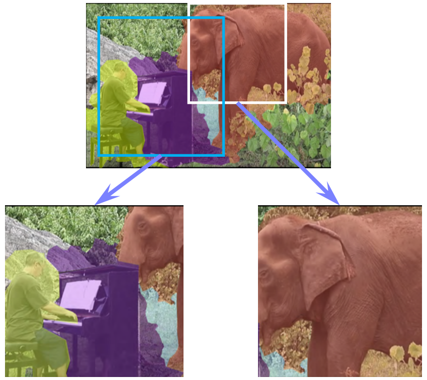


ORL [Xie, J. et la.]



SoCo [Wei, F. et la.]

Cluster-based



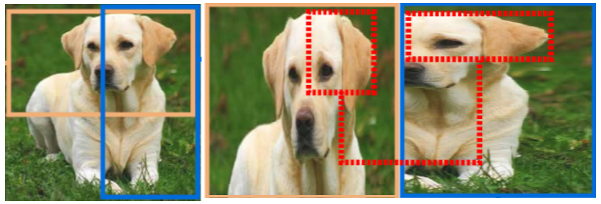
DetCon [Hénaff, O.J et la.]

- Relying on critical **non-trainable component** to find the positive pairs.
  e.g., *Selective search* (region proposal), *Felzenszwalb-Huttenlocher* algorithm (cluster proposal), corresponding regions mapping, etc.

- **Semantic misalignment** with dense prediction tasks.
  e.g., considering similar representations for correspondent regions only, considering distinct representation of every patch, etc.

- **Discrepancy** with actual semantics of an image.
  e.g., untended separation of representations from foreground and background (interest and non-interest), rectangular RoIs and misaligned cluster masks posing mismatched positive pairs.

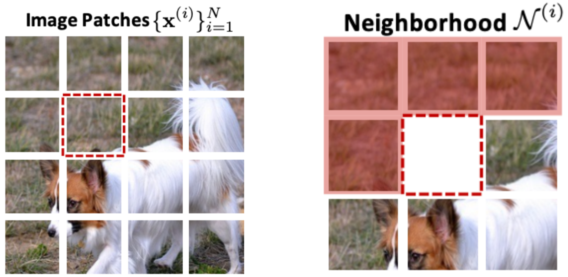
Overlap-based



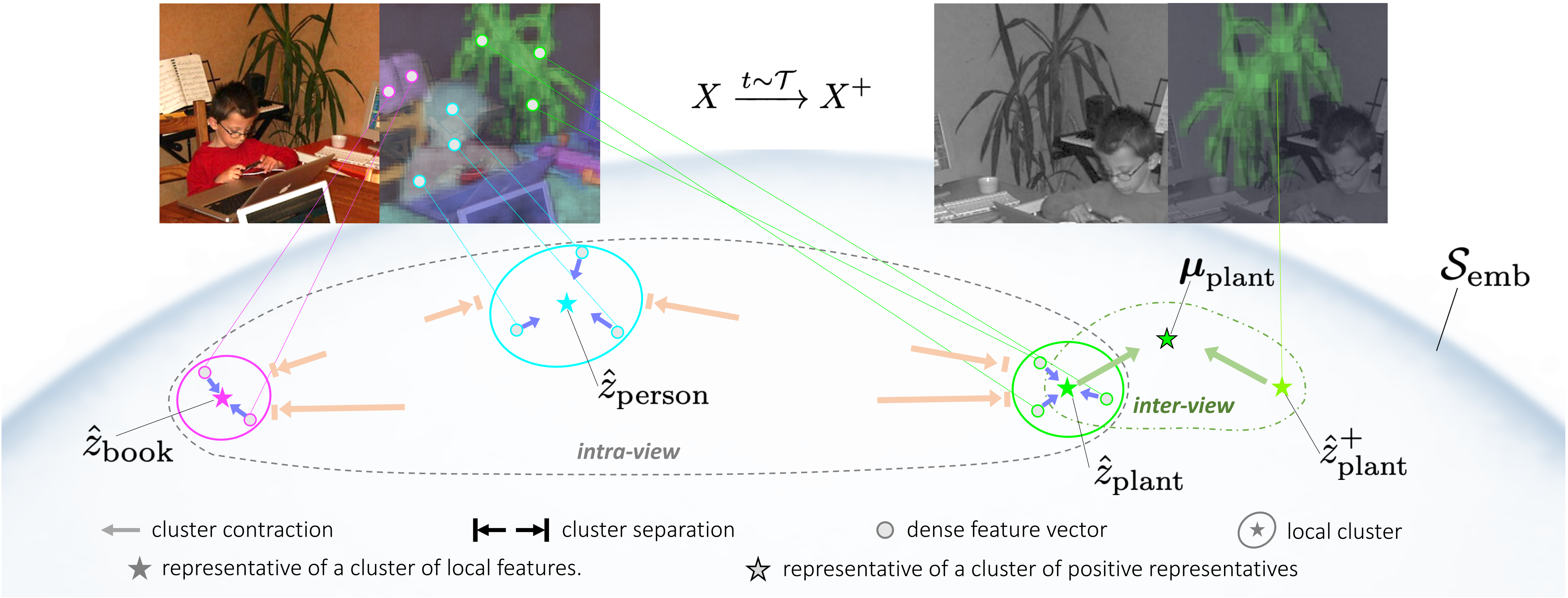
ReSim [Xiao, T. et la.]

Patch-based

DetCo [Xie, E. et la.]



DINO + SelfPatch [Yun, S. et la.]



The Bi-level Clustering of FLSL

1st level (Intra-view) representations corresponding to a semantic concept (as a cluster), $\boldsymbol{z} \in \tilde{\boldsymbol{z}}^c$, are close to its cluster representative (mode) $\hat{\boldsymbol{z}}^c$ and far away from the representations of other clusters;

2nd level (Inter-view) the cluster representatives (modes) $\hat{\boldsymbol{z}}$s corresponding to the same semantic concept in $\boldsymbol{X}$s over the dataset are pushed closer to each other.

Self-attention from *mean-shift* (MS) Clustering Perspective

KDE

$$p(\boldsymbol{z}) = \sum_{i=1}^{N} p(\boldsymbol{z}_i)p(\boldsymbol{z}|\boldsymbol{z}_i) = \sum_{i=1}^{N} \pi_i \frac{1}{T_i} K(d(\boldsymbol{z}, \boldsymbol{z}_i; \boldsymbol{\Sigma}_i))$$

Mode

Stationary points $\partial p(\boldsymbol{z}) / \partial \boldsymbol{z} = 0$

$$\hat{\boldsymbol{z}} = \boldsymbol{f}(\boldsymbol{z}) = \sum_{i=1}^{N} p(\boldsymbol{z}_i|\boldsymbol{z})\boldsymbol{z}_i$$

$$p(\boldsymbol{z}_i|\boldsymbol{z}) = \frac{\pi_i \frac{1}{T_i} K'(d(\boldsymbol{z}, \boldsymbol{z}_i; \boldsymbol{\Sigma}_i))\boldsymbol{\Sigma}_i^{-1}}{\sum_{j=1}^{N} \pi_j \frac{1}{T_j} K'(d(\boldsymbol{z}, \boldsymbol{z}_j; \boldsymbol{\Sigma}_j))\boldsymbol{\Sigma}_j^{-1}}$$

- ℓ_2-normalized vectors
- homoscedastic Gaussian kernel
- ...

$$\hat{\boldsymbol{z}} = \mathrm{meanshift}(\boldsymbol{z}, \tau) = \sum_{i=1}^{N} \frac{\exp(\tau \boldsymbol{z}^\top \boldsymbol{z}_i)}{\sum_{j=1}^{N} \exp(\tau \boldsymbol{z}^\top \boldsymbol{z}_j)} \boldsymbol{z}_i$$

Generalized MS with linear operator

$$\hat{\boldsymbol{Z}} = \mathrm{SA}(\boldsymbol{Z}) = \mathbf{W}_V\ \boldsymbol{Z}\ \mathrm{softmax}\left(1/\sqrt{D_{qk}}\, \boldsymbol{Z}^\top (\mathbf{W}_Q^\top \mathbf{W}_K)\ \boldsymbol{Z}\right)$$

Attention mechanism

$$\mathrm{attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{V}\mathrm{softmax}\left(\boldsymbol{Q}^\top \boldsymbol{K} / \sqrt{D_{qk}}\right)$$

NEURAL INFORMATION PROCESSING SYSTEMS

1st level: intra-view clustering representations corresponding to a semantic concept (as a cluster), $\boldsymbol{z} \in \tilde{\boldsymbol{z}}^c$, are close to its cluster representative (mode) $\hat{\boldsymbol{z}}^c$ and far away from the representations of other clusters.

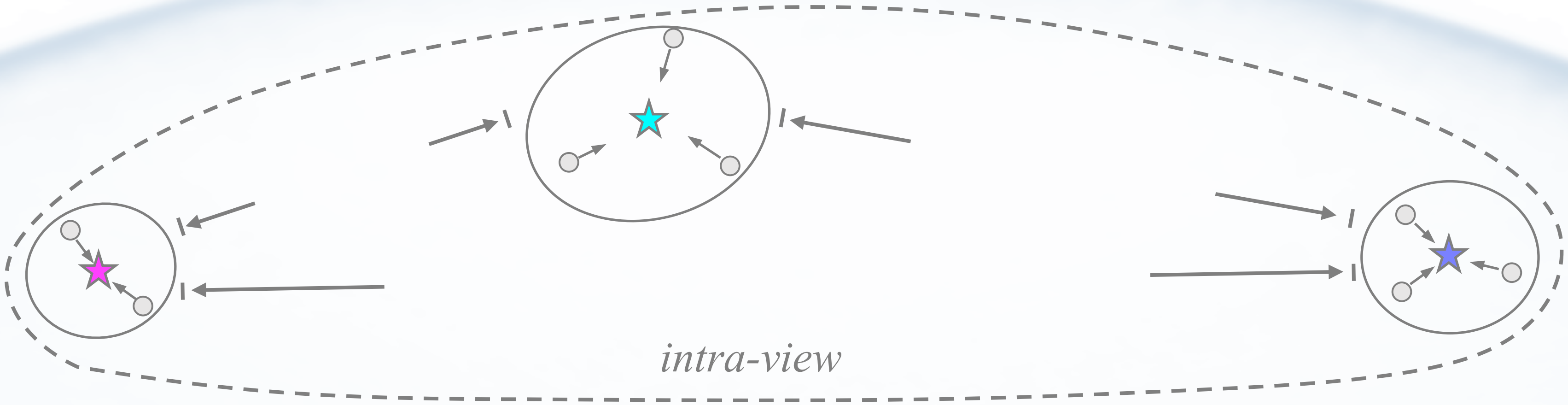
soft mask

$$\left[\mathrm{softmax}\left(\tau \boldsymbol{z}_i^\top \boldsymbol{Z}\right)\right]_j = p(\boldsymbol{z}_j|\boldsymbol{z}_i) \geq 1/\left(\left(\sum_{k\in c_i} e^{(\boldsymbol{z}_i^\top \boldsymbol{z}_k - \boldsymbol{z}_i^\top \boldsymbol{z}_j)\tau}\right) + (N-|c_i|)e^{-\Delta_{ij}\tau}\right), \forall j \in c_i$$

$$\Delta_{ij} = \boldsymbol{z}_i^\top \boldsymbol{z}_j - \max_{m\in[N]\setminus c_i} \boldsymbol{z}_i^\top \boldsymbol{z}_m,\ j \in c_i$$

cluster separation ↑
$\Delta_i = \min \Delta_{ij}\ , j \in c_i$

$$\min_{f_\theta} \sum_{i=1}^{N} \|\boldsymbol{z}_i - \hat{\boldsymbol{z}}_i\|_2^2 \qquad \hat{\boldsymbol{z}}_i = \boldsymbol{Z}\mathrm{softmax}(\tau \boldsymbol{z}_i^\top \boldsymbol{Z})$$



NEURAL INFORMATION PROCESSING SYSTEMS

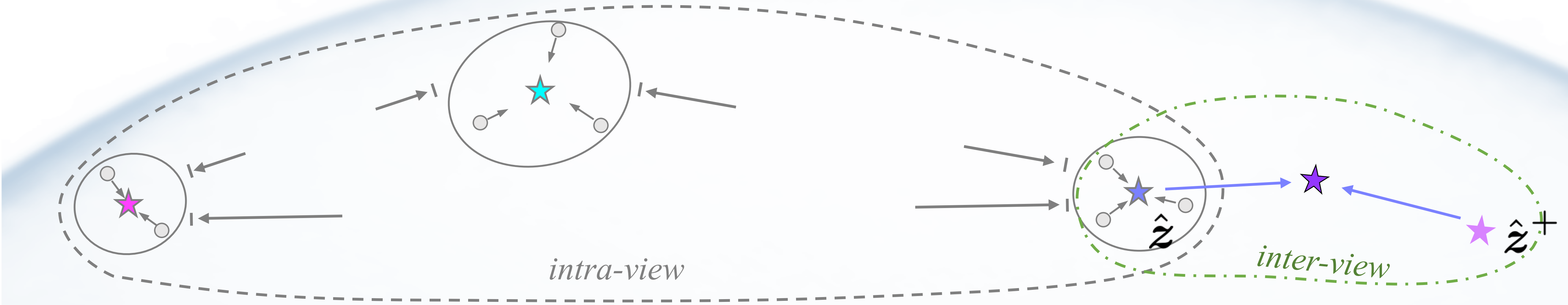
2nd-level: Inter-view Clustering the cluster representatives (modes) $\hat{\boldsymbol{z}}^c$s corresponding to the same semantic concept in $\boldsymbol{X}$s over the dataset are pushed closer to each other.

$$\min_{\mathcal{M}} \frac{1}{N'} \sum_{\hat{\boldsymbol{z}} \in \hat{\mathcal{Z}}} \sum_{k=1}^{K} \delta_{kk(\hat{\boldsymbol{z}})} \|\hat{\boldsymbol{z}} - \boldsymbol{\mu}_{k(\hat{\boldsymbol{z}})}\|_2^2 + D_{\mathrm{KL}}\left(\bar{\boldsymbol{p}} \| \boldsymbol{\pi}\right)$$

$$\min_{\mathcal{M}} \frac{1}{N'} \sum_{\hat{\boldsymbol{z}} \in \hat{\mathcal{Z}}} \left(\sum_{k=1}^{K} \delta_{kk(\hat{\boldsymbol{z}})} \|\hat{\boldsymbol{z}} - \boldsymbol{\mu}_{k(\hat{\boldsymbol{z}})}\|_2^2 + \underbrace{\left(1 - \delta_{k(\hat{\boldsymbol{z}}^+)k(\hat{\boldsymbol{z}})}\right) \|\hat{\boldsymbol{z}}^+ - \boldsymbol{\mu}_{k(\hat{\boldsymbol{z}})}\|_2^2}_{\text{separation margin for } \hat{\boldsymbol{z}}^+} \right) + D_{\mathrm{KL}}\left(\bar{\boldsymbol{p}} \| \boldsymbol{\pi}\right)$$

Positive pair retrieval: $\hat{\boldsymbol{z}}^+ = \boldsymbol{Z}^+ \operatorname{softmax}\left(\tau \boldsymbol{z}^\top \boldsymbol{Z}^+\right)$

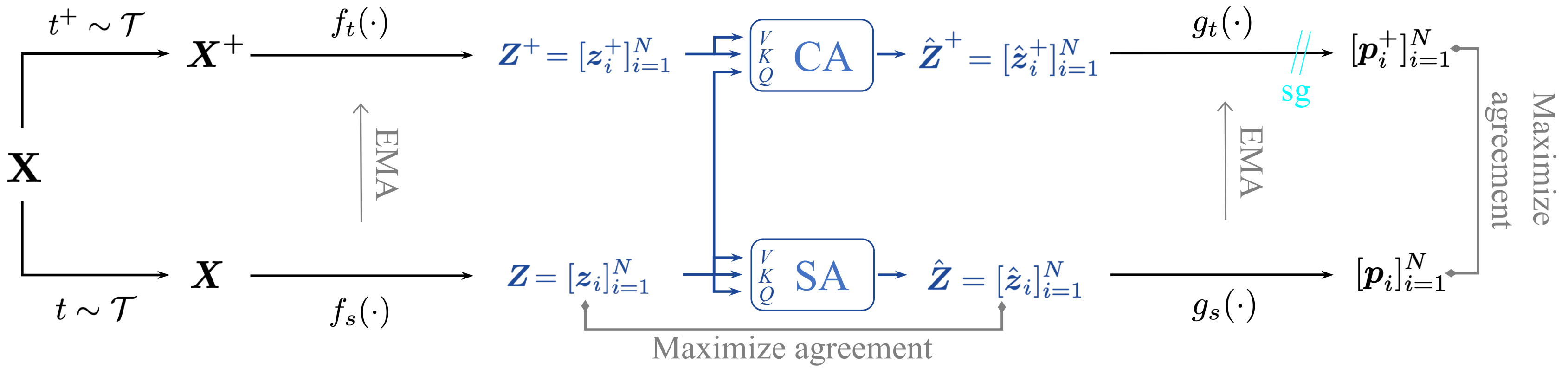
$$\boxed{\min_{\mathcal{M}} \frac{1}{N'} \sum_{\hat{\boldsymbol{z}} \in \hat{\mathcal{Z}}} \mathrm{H}(\boldsymbol{p}(\hat{\boldsymbol{z}}^+), \boldsymbol{p}(\hat{\boldsymbol{z}})) + D_{\mathrm{KL}}\left(\bar{\boldsymbol{p}} \| \boldsymbol{\pi}\right)}$$



FLSL Objective

$$\min \frac{1}{N'} \underbrace{\sum_{\boldsymbol{Z}\in\mathcal{Z}} \sum_{\boldsymbol{z}\in\boldsymbol{Z}} \upsilon \|\boldsymbol{z}-\hat{\boldsymbol{z}}\|_2^2}_{1^{st}\text{-level}} + \underbrace{\eta \sum_{\boldsymbol{z}\in\boldsymbol{Z}} \mathrm{H}(\boldsymbol{p}(\hat{\boldsymbol{z}}^+), \boldsymbol{p}(\hat{\boldsymbol{z}})) + \gamma D_{\mathrm{KL}}\left(\bar{\boldsymbol{p}} \| \boldsymbol{\pi}\right)}_{2^{nd}\text{-level}},$$

$$\text{with } \hat{\boldsymbol{z}} = \mathrm{SA}(\boldsymbol{z}, \boldsymbol{Z}, \boldsymbol{Z}),\ \hat{\boldsymbol{z}}^+ = \mathrm{CA}(\boldsymbol{z}, \boldsymbol{Z}^+, \boldsymbol{Z}^+),$$



FLSL

Pretrain	Backbone	Epoch	#Params	AP^{bbox}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{70}
MoCo-v2	RN50	200	23M	38.9	59.2	42.4	35.5	56.2	37.8
DetCo	RN50	200	23M	40.1	61.0	43.9	36.4	58.0	38.9
DenseCL	RN50	200	23M	40.3	59.9	44.3	36.4	57.0	39.2
BYOL	RN50	1000	23M	40.4	61.6	44.1	37.2	58.8	39.8
SCRL	RN50	1000	23M	41.3	62.4	45.0	37.7	59.6	40.7
MOCO-v3	ViT-S/16	300	21M	39.8	62.6	43.1	37.1	59.6	39.2
MoBY	ViT-S/16	300	21M	41.1	63.7	44.8	37.6	60.3	39.8
DINO	ViT-S/16	300	21M	40.8	63.4	44.2	37.3	59.9	39.5
DINO+SelfPatch	ViT-S/16	200	21M	42.1	64.9	46.1	38.5	61.3	40.8
ADCLR	ViT-S/16	300	21M	44.3	65.4	47.6	39.7	62.1	41.5
FLSL	ViT-S/16	300	21M	44.9	66.1	48.1	40.8	64.7	44.2
FLSL	ViT-S/8	300	21M	46.5	69.0	51.3	42.1	65.3	45.0

Table 1: MASK R-CNN ON COCO

Pretrain	AP^{bbox}	AP^{bbox}_{s}	AP^{bbox}_{m}	AP^{bbox}_{l}	AP^{mk}
None	48.1	-	-	-	42.6
IN-1k Supv.	47.6	-	-	-	42.4
IN-21k Supv.	47.8	-	-	-	42.6
IN-1k DINO	48.9	32.9	52.2	62.4	43.7
IN-1k MAE	51.2	34.9	54.7	66.0	45.5
IN-1k FLSL	53.1	36.9	56.2	67.4	47.0

Table 2: VITDET-B/16 WITH MASK R-CNN ON COCO

Pretrain	Backbone	AP_{VOC}
IN-1k DINO	ViT-S/16	48.9
IN-1k DINO	ViT-B/16	49.1
IN-1k DINO	ViT-S/8	51.1
IN-1k FLSL	ViT-S/16	53.1
IN-1k FLSL	ViT-B/16	53.5
IN-1k FLSL	ViT-S/8	55.2

Table 3: FASTER R-CNN FPN ON UAVDT

Dense prediction benchmark 1

MS-COCO Object Detection and Segmentation

Mask RCNN + ViT-S/16 and ViT-S/8, ViTDet + ViT-B/16

Dense prediction benchmark 2

UAVDT Vehicle Detection

Faster R-CNN FPN

+ ViT-S/16, ViT-S/8 and ViT-B/16

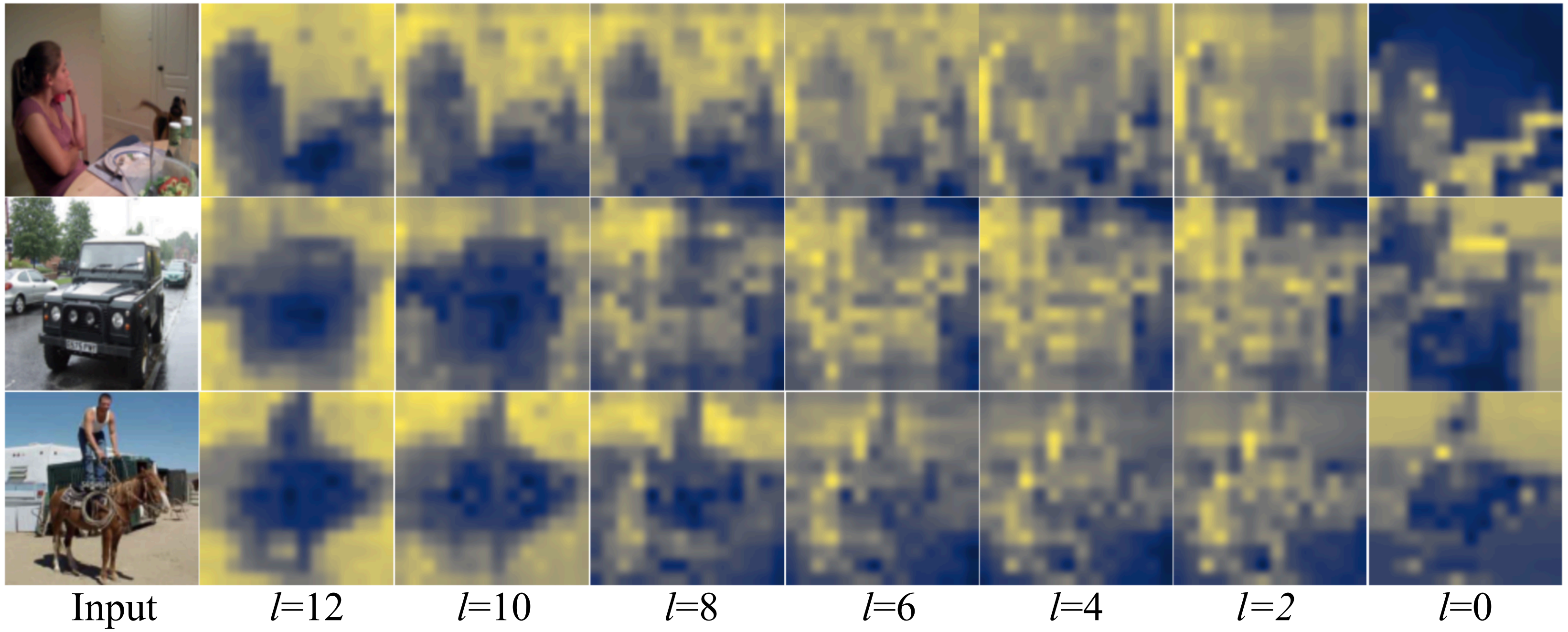
NEURAL INFORMATION PROCESSING SYSTEMS

Method	Arch	Backbone	#Iter.	mIoU	aAcc	mAcc
MoCo-v2	FPN	RN50	40k	35.8	77.6	45.1
SwAV	FPN	RN50	40k	35.4	77.5	44.9
ReSim	FPN	RN50	40k	36.6	78.4	46.4
DenseCL	FPN	RN50	40k	37.2	78.5	47.1
MoCo-v3	FPN	ViT-S/16	40k	35.3	78.9	47.1
MoBY	FPN	ViT-S/16	40k	39.5	79.9	47.1
DINO	FPN	ViT-S/16	40k	38.3	79.0	47.1
DINO+SelfPatch	FPN	ViT-S/16	40k	41.2	80.7	52.1
ADCLR	FPN	ViT-S/16	40k	42.4	81.1	54.2
FLSL	FPN	ViT-S/16	40k	**42.9**	**81.5**	**55.1**

Pretrain	Arch.	$(\mathcal{J}\&\mathcal{F})_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
IN-1k supv.	ViT-S/8	66.0	63.9	68.1
VLOG CT	RN50	48.7	46.4	50.0
YT-VOS MAST	RN18	65.5	63.3	67.6
IN-1k DINO	ViT-S/16	61.8	60.2	63.4
IN-1k DINO	ViT-B/16	62.3	60.7	63.9
IN-1k DINO	ViT-S/8	69.9	66.6	73.1
IN-1k FLSL	ViT-S/16	65.6	62.4	69.4
IN-1k FLSL	ViT-B/16	66.1	62.9	70.0
IN-1k FLSL	ViT-S/8	73.5	69.9	78.1

Dense prediction benchmark 3
ADE20K Semantic Segmentation
FPN + ViT-S/16

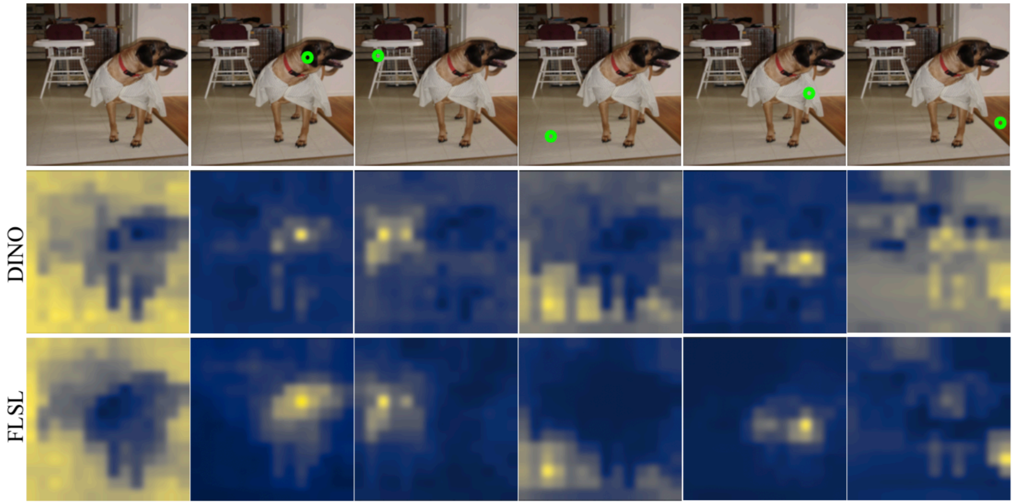
Dense prediction benchmark 4
Davis 2017 Video Instance Segmentation
ViT-S/16, ViT-S/8 and ViT-B/16

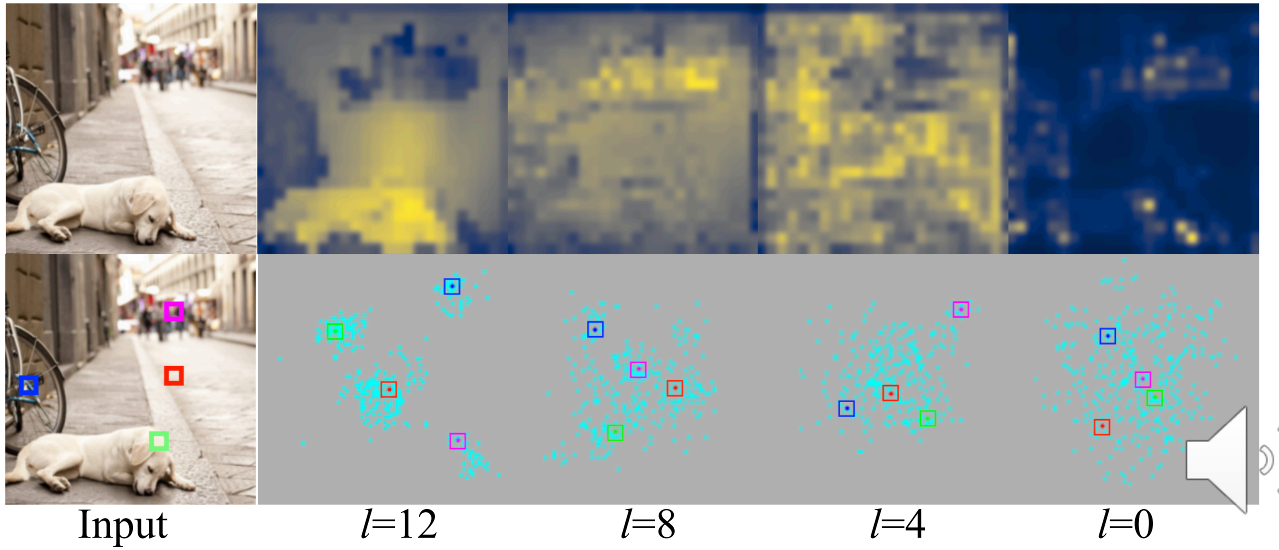


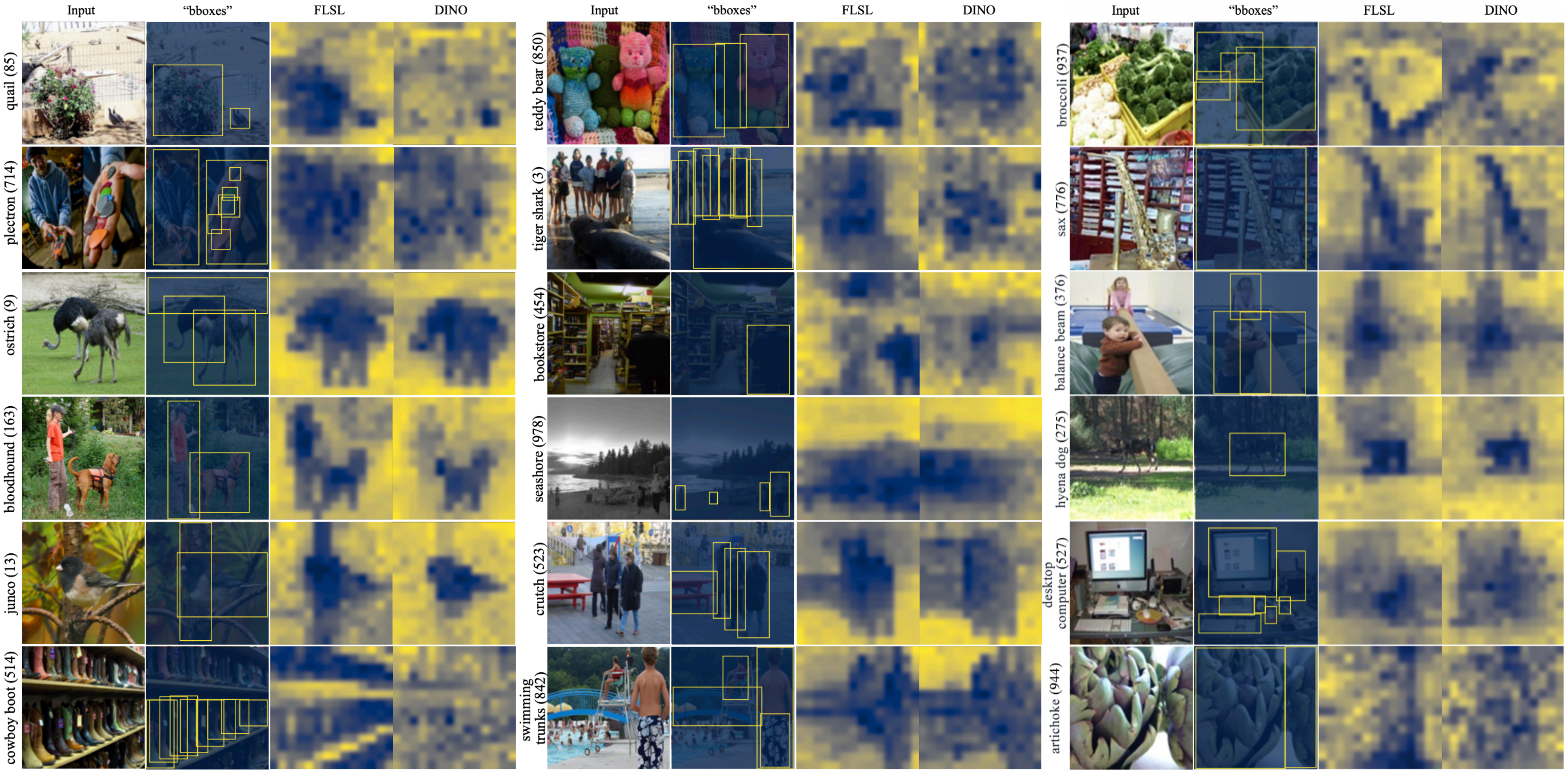
← Visualization of Aggregated Similarity Score (ASS) map from different ViT layers.

↙ Self-attention probing maps for features learned via FLSL and DINO

↓ Visualization of separability of features from different layers via t-SNE







ASS map visual comparison between FLSL and DINO