Unsqueeze [CLS] Bottleneck to Learn Rich Representation

Qing Su, Shihao Ji
Georgia State University

1

Are the Current SSL Methods Good Enough?

- **Issues of multi-view methods**

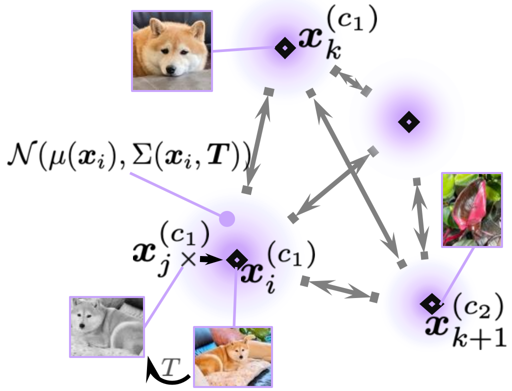
  — **Implicit Clustering (IC):** SimCLR, MoCo, BYOL, Barlow Twins, etc.
  Overfitting issue / under-compression ---> nuisance info.

  — **Explicit Clustering (EC):** SwAV, DINO, etc.
  Underfitting issue / over-compression ---> information loss

Impact of Compression of DINO

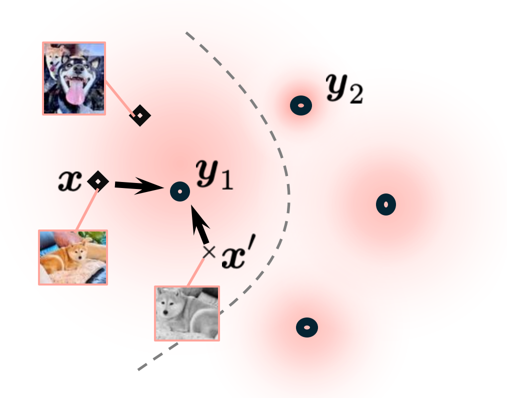
#epoch	300	800
lin. Top1	76.1	77.0 (+0.9)
AP	41.6	40.8 (−0.8)

➢ **Multi-level methods:** DenseCL, DetCo, EsViT, iBOT, DINOv2
Misaligned semantic constraints or tasks ---> suboptimal



Superscript in (·) denotes class label

Implicit Clustering



Explicit clustering

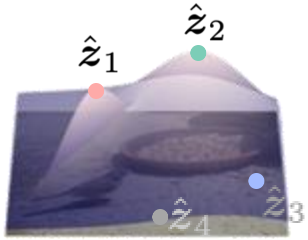
How to Learn Meaningful Representations While Preserving More Information?

- **Source of over-compression in Explicit-Clustering (EC) -based methods:**
  - Radical hyperparameter settings,
    e.g., small number of centroids, small temperature.
  - *Chasing a sharper target distribution (in distillation-based methods such as DINO).
- **Source of semantic misalignment in Multi-level SSL methods:**
  - Fixed-size semantic constraints that misaligned with local semantic distribution,
    e.g., matching representations of blocks with fixed size (as in DenseCL, DetCo, EsViT).
  - Shared projector for objectives of different semantic types.
    e.g., tying projector for image-level semantics and local patch recovering (via MIM) (as in iBOT)

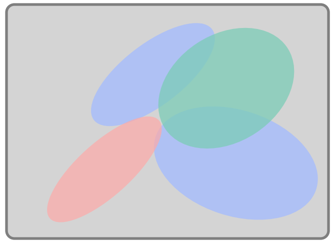
How to Learn Meaningful Representations While Preserving More Information?

Our Remedy: Natural Image Modeling

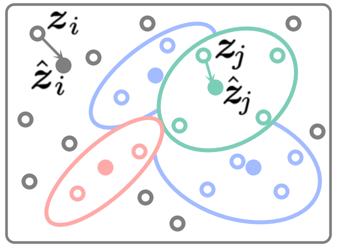
- A natural image should be mixture of semantic concepts:

Smoothed feature distribution of a natural image



Mixture model of semantics



Stratified Random Sampling preserving image structure

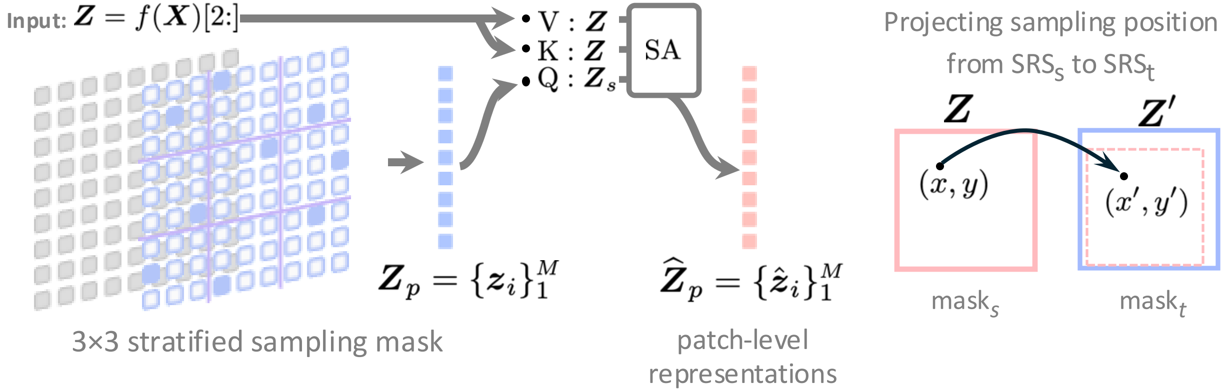
Multi-model target distribution:

$$p(y|\boldsymbol{Z}) \approx {}^{1}/_{M} \sum_{i=1}^{M} p(y|\hat{\boldsymbol{z}}_i), \quad \text{where} \quad \hat{\boldsymbol{z}}_i = \mathrm{SA}(\boldsymbol{z}_i, \boldsymbol{Z}, \boldsymbol{Z}), \quad i \sim \mathcal{U}_i.$$

- Learn an extra class token (`cls+`) to extract the "multi-modal" information.

OUR REMEDY: Stratified Random Sampling (SRS) module

- **Leverage self-attention to extract representations that align with local context.**

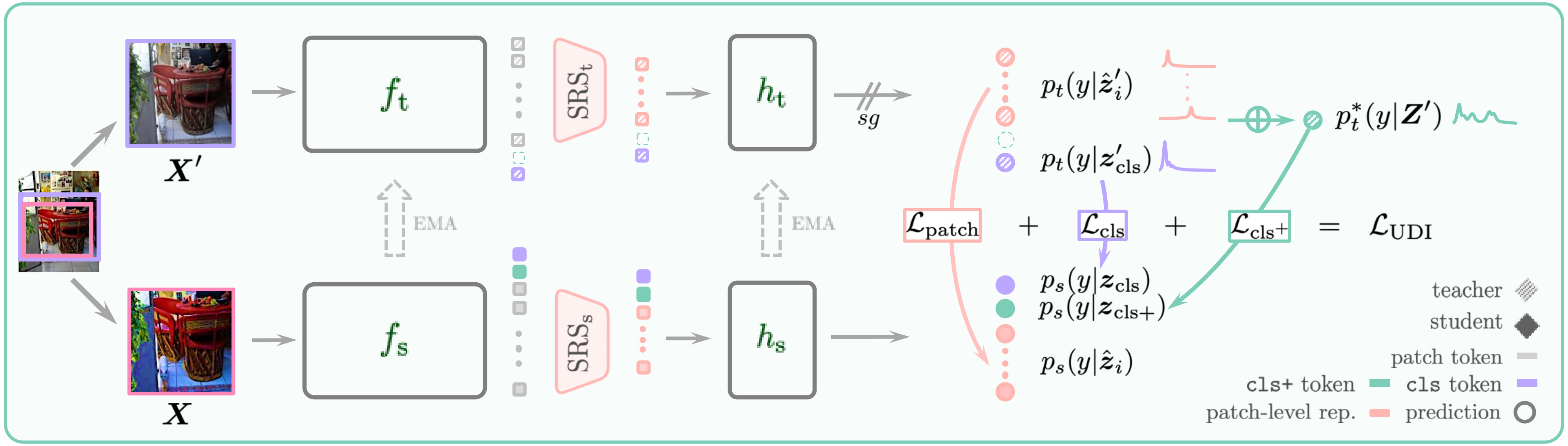


  - Attention map serves as soft-masked pooling, applying semantically coherent constraint that reflects local context.

SRS ATTENTION MAP WITH QUERY AS LOCAL PATCH

  - Stratified random sampling for efficiency while preserving the image layout

OUR REMEDY: UDI pipeline



$$\mathcal{L}_{image} = \mathcal{L}_{cls} = \mathbb{E}_{z_{cls}, z'_{cls} \in \mathcal{Z}_{cls}} \left[\mathrm{H}(p_t(y|z'_{cls}), p_s(y|z_{cls})) \right]$$

$$\mathcal{L}_{patch} = \mathbb{E}_{z_p, z'_p \in \mathcal{Z}_p} \left[\mathrm{H}(p_t(y|z'_p), p_s(y|z_p)) \right]$$

$$\mathcal{L}_{cls+} = \mathbb{E}_{Z' \in \mathcal{Z}, z_{cls+} \in \mathcal{Z}_{cls+}} \left[\mathrm{H}\left(p^*_t(y|Z'), p_s(y|z_{cls+})\right)\right].$$

$$p(y|Z) \approx \frac{1}{M}\sum_{i=1}^{M} p(y|\hat{z}_i), \quad \text{where} \quad \hat{z}_i = \mathrm{SA}(z_i, Z, Z), \quad i \sim \mathcal{U}_i.$$

$$p^*(y|Z) = \alpha\, p(y|Z) + (1-\alpha)\, p(y|z_{cls}).$$

Multi-level objective with an additional term for the extra class token

$$\begin{aligned} \mathcal{L}_{UDI} &= \mathcal{L}_{image} + \mathcal{L}_{patch} \\ &= \mathcal{L}_{cls} + \mathcal{L}_{cls+} + \mathcal{L}_{patch}. \end{aligned}$$

MAIN RESULTS

- Performance on image-level and dense prediction downstream tasks

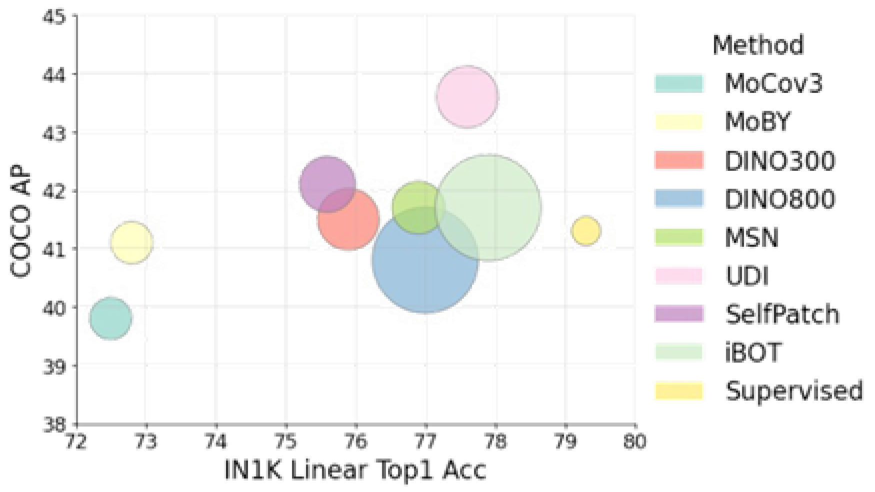
UDI achieves balanced improvement on both image-level (linear probing) and dense prediction task (MS-COCO) with 1.5% and 1.6 AP, respectively, with relatively small training budget.

TAB. 1.

Objective	Arch.	#Views	Epoch†	Linear	AP^{bb}
MoCov2	RN50	2	400	67.5	38.9
DenseCL [67]	RN50	2	400	64.6 (−2.9)	40.3 (+1.4)
ReSim [70]	RN50	2	400	66.1 (−1.4)	40.3 (+1.4)
DetCo [71]	RN50	2	400	68.6 (+1.1)	40.1 (+1.2)
SimCLR [12]	RN50	2	200	65.4	40.5
PixPro [12]	RN50	2	200	66.3 (+0.9)	40.9 (+0.5)
DINO [9]	ViT-S/16	10	1050	76.0	41.5
Selfpatch [76]	ViT-S/16	10	1050	75.6 (-0.4)	42.1 (+0.6)
DINO [9]	Swin-T/14	10	1050	77.1	46.0
EsViT [38]	Swin-T/14	10	1050	77.6 (+0.5)	46.2 (+0.2)
DINO [9]	ViT-S/16	12	1200	76.1	41.6
iBOT [79]	ViT-S/16	12	1200	77.4 (+1.3)	41.7 (+0.1)
DINO [9]	ViT-S/16	12	1200	76.1	41.6
UDI	ViT-S/16	12	1200	77.6 (+1.5)	43.2 (+1.6)

IMG. 1.

*bubble size ∝ #epochs in SSL



Main Results

- Using the regular class token (UDI) V.S. using the extra class token (UDI+)

- Regular class token in UDI extracts more linearly separable features, resulting in better performance in low-shot learning.

- The extra class token in UDI (denoated by UDI+) captures more information from images, resulting in superior performance in transfer learning than the regular class token.

Tab. 2. Low-shot Learning

Method	Arch.	logistic regression 1%	logistic regression 10%	fine-tuning 1%	fine-tuning 10%
SimCLRv2 [14]	RN50	—	—	57.9	68.1
BYOL [28]	RN50	—	—	53.2	68.8
SwAV [8]	RN50	—	—	53.9	70.2
SCLRv2+SD	RN50	—	—	60.0	70.5
DINO [9]	ViT-S/16	64.5	72.2	60.3	74.3
iBOT [79]	ViT-S/16	65.9	73.4	61.9	75.1
MSN [1]	ViT-S/16	**67.2**	—	—	—
UDI	ViT-S/16	66.7	**74.1**	**65.8**	**76.4**
UDI_{+}	ViT-S/16	66.1	73.8	65.2	76.2

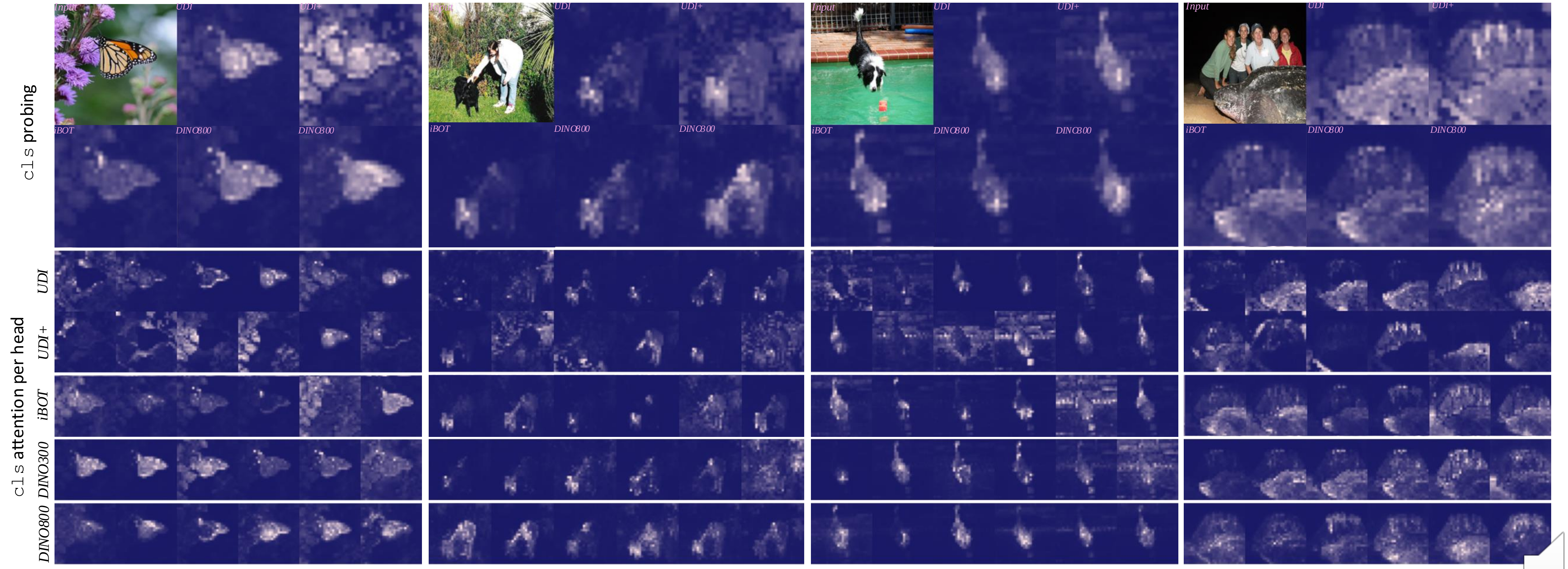
Tab. 3. Transfer Learning

Method	ViT-S/16 Cif_{10}	Cif_{100}	INat_{18}	INat_{19}	Flowers	Car
Supervised [9]	99.0	89.5	70.7	76.6	98.2	92.1
BEiT [3]	98.6	87.4	68.5	76.5	96.4	92.1
DINO [9]	99.0	90.5	72.0	78.2	98.5	93.1
DINO+reg.	98.8	90.5	72.1	78.2	98.5	93.2
iBOT [79]	**99.1**	90.7	73.7	78.5	98.6	94.0
UDI	**99.1**	90.8	74.1	78.9	98.6	**94.1**
UDI_{+}	**99.1**	**91.3**	**74.8**	**79.7**	**98.9**	94.0

Visualizations

- **Visualization of attention map using class token as a whole and per head.**

  UDI promotes attention maps that are more diverse and contextually aligned, in contrast to more focused attention of other SSLs.



Thank You!