SCALABLE PATCH-LEVEL SELF-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised learning (SSL) at scale produces powerful visual representations. However, most scalable SSL methods rely on ad hoc combinations of multiple objectives and stabilization mechanisms. Taking a step back, we ask if we can design a high-performing, yet principled SSL algorithm. Starting from the multi-view assumption—stipulating that task-relevant content is captured by the information common to different views—we construct an information-theoretic objective decomposing into interpretable terms. This derivation yields JEM, a student–teacher method that learns by aligning corresponding patch representations across views, explicitly regularized by information and structure preservation losses. JEM trains stably from 300M to 7B parameters, and, to our knowledge, is the first latent-space patch-level method demonstrated at 7B scale. Across all scales, JEM reaches strong performance on both global and dense probing tasks, on segmentation benchmarks consistently surpassing the DINOv2 algorithm—an influential foundation for today's strongest visual SSL methods. Notably, at 7B parameters, it also exceeds DINOv3's performance on panoptic segmentation, despite DINOv3 being trained on a 10× larger dataset with several refinement stages. These results demonstrate that we can indeed build an SSL algorithm which is *principled* and *stable*, without sacrificing *downstream performance*.

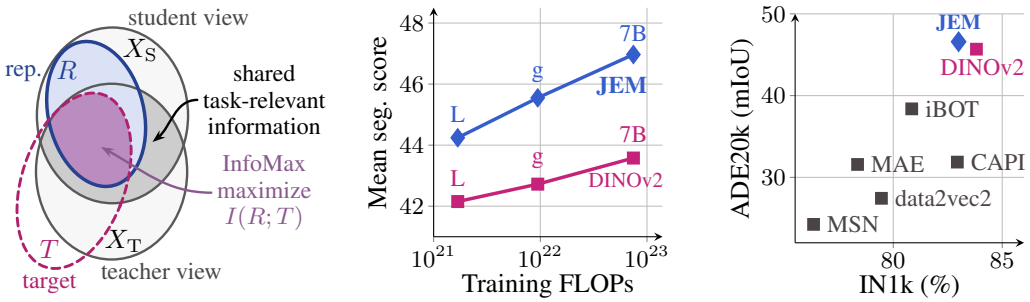
1 INTRODUCTION

Self-supervised learning has become a standard route to general-purpose visual representations, and the strongest encoders now match or exceed supervised pretraining across a wide range of tasks (Oquab et al., 2024; Siméoni et al., 2026). The best methods combine both image- and patch-level objectives, and have been successfully scaled (Oquab et al., 2024; Siméoni et al., 2026; Fan et al., 2025; Fu et al., 2026). However, this success comes at a price: these methods have grown complex, their objectives do not always compose naturally, and additional stabilization tricks are required to keep them from degrading each other (Oquab et al., 2024; Siméoni et al., 2026). Impressive as these results are, we take them as an invitation to take a step back and reconsider what a principled self-supervised objective could look like.

Many successful methods are built from the idea that different views (altered through augmentations, cropping, masking) of the same image should be described similarly (Chen et al., 2020; Caron et al., 2021; Zbontar et al., 2021). We believe that the underlying *multi-view assumption*—which suggests that if each view is sufficient for the downstream task, then a representation should retain the information shared between views (Fig. 1a)—can act as a core learning principle. From the multi-view assumption, we derive a multi-view InfoMax objective for student-teacher SSL with two terms: one maximizing cross-view alignment, one preventing information collapse.

We then discuss how to instantiate this objective into JEM, a novel joint-embedding multi-view method trained with a simple patch-level cross-view alignment loss, and two regularization terms, one preventing global collapse, one preserving the spatial structure of the features. We show that JEM trains stably and scales well (Fig. 1b), without additional machinery, to models of 7B parameters. JEM performs strongly on dense benchmarks, consistently surpassing a DINOv2 baseline on semantic, instance, and panoptic segmentation, while remaining competitive with the strongest models on image-level classification (Fig. 1c). At 7B scale, JEM even exceeds DINOv3's performance on panoptic segmentation, despite being trained for a shorter time on a smaller dataset without additional refinement steps.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Multi-view information diagram (more in Sec. 3.1), leading to JEM.

(b) Mean segmentation (over six datasets) scales with model size.

(c) Dense vs. global performance across SSL methods (VIT-L, IN1k).

Figure 1: **We introduce JEM**, a student–teacher SSL objective grounded in information theory. Following the multi-view principle (a), it maximizes the information a student representation retains about what views share. Average scores over semantic, panoptic and instance segmentation at different scales (b) show that JEM scales effectively and consistently outperforms DINOv2. (c) JEM surpasses methods using patch- and/or image-level objectives on dense tasks while remaining competitive on classification. JEM is *principled*, *stable* and does not sacrifice *downstream performance*.

To summarize, we (1) derive a principled information-theoretic SSL objective leading to JEM, a joint-embedding multi-view method trained with a patch-level objective, that (2) trains stably across scales up to 7B parameters and (3) outperforms the DINOv2 algorithm in controlled comparisons on dense prediction tasks while remaining competitive on classification, even exceeding the DINOv3 7B foundation model on panoptic segmentation.

2 Related Work

Self-Supervised Visual Representation Learning Modern SSL falls into four broad families. Contrastive methods (He et al., 2020; Chen et al., 2020) pull together representations of different views of the same image while pushing apart representations of other images. Redundancy-reduction methods (Zbontar et al., 2021; Bardes et al., 2022a) enforce view invariance while explicitly preventing collapse across dimensions. Self-clustering approaches such as SwAV Caron et al. (2020) and DINO (Caron et al., 2021) learn cluster assignments and align those across views. Masked prediction instead removes part of the input and predicts the missing information: MAE (He et al., 2022) predicts pixels, while I-JEPA (Assran et al., 2023) predicts latent representations of spatially localized target blocks, and iBOT (Zhou et al., 2022) and CAPI (Darcet et al., 2025) predict cluster assignments for masked patches.

Recent methods increasingly differ in whether they rely on a single objective family or combine complementary objectives. I-JEPA and CAPI are examples of single-family latent prediction approaches: I-JEPA predicts latent representations of masked target blocks, while CAPI predicts latent cluster assignments for masked patches. DINOv2 (Oquab et al., 2024), in contrast, combines global DINO self-distillation with patch-level iBOT prediction and KoLeo (Sablayrolles et al., 2019) regularization; DINOv3 (Siméoni et al., 2026) even adds a further term, Gram anchoring, to counter an effect observed during long training at large scale: global performance keeps improving while dense feature maps progressively lose spatial locality. Are such combinations necessary for scaling SSL? Our work explores a principled information-theoretic objective derived from maximizing mutual information between representations of different views. Unlike methods with separate global and local objectives, our model has no class token and applies the objective exclusively to patch representations, while still matching the performance of multi-objective methods such as DINOv2.

Information theory in SSL Information-theoretic SSL spans several closely related perspectives (Shwartz-Ziv & LeCun, 2024). The original InfoMax principle (Linsker, 1988) prescribes maximizing the information the representation retains about its input, implemented for representation learning by Deep InfoMax (Hjelm et al., 2019). Other works instead maximize the information between different *views* of the data (van den Oord et al., 2018; Bachman et al., 2019; Tian et al.,

2020a; Henaff, 2020), yielding the *multi-view InfoMax* principle; we also use it as our starting point. However, instead of estimating mutual information with contrastive objectives or separate critics, we use a variational approximation which naturally fits the student-teacher paradigm common in modern SSL. Complementarily, the information bottleneck principle suggests retaining predictive information while compressing superfluous information (Tishby et al., 1999; Alemi et al., 2017). Federici et al. (2020) combine it with multi-view InfoMax by preserving shared and suppressing view-specific information. In contrast, we derive an explicit anti-collapse term that *maintains* view-specific information; however, our use of discrete representation targets can be seen as an architectural bottleneck. The limits of mutual information as an objective for SSL have been discussed by Tschannen et al. (2020); Tian et al. (2020b); Tsai et al. (2021). For us, the multi-view InfoMax objective informs practical loss design, acknowledging that learning good representations also relies on inductive biases from architecture and view design. We review more related work in App. A.1.

3 Method

We aim to build an SSL method that is *principled* and *stable*, without any sacrifice in *downstream performance*. Starting from the multi-view assumption, we derive an information-theoretic objective (Sec. 3.1). From this objective, we derive an alignment loss (Sec. 3.2.1) and regularization terms that mitigate its collapse modes (Secs. 3.2.2 and 3.2.3). In Sec. 3.3, we then discuss the careful design of the views, which we found to be crucial for best performance. Throughout the section we provide detailed design choices that make JEM, a new joint-embedding SSL method yielding results competitive with more complex and highly tuned SSL objectives.

3.1 The Multi-View InfoMax Objective

The *multi-view assumption* stipulates that any view of the data alone contains sufficient information for the downstream tasks (Sridharan & Kakade, 2008). It follows that all task-relevant information must be contained in the *shared information* across views. Whether the multi-view assumption holds in practice is dependent on the choice of views and the downstream tasks, and its limitations are known (Tsai et al., 2021; Shwartz-Ziv & LeCun, 2024). Still, we use it as an idealized principle that motivates a criterion for learning representations.

Setup: Given a sample X, we apply masking, cropping, and photometric distortions to generate views. For generality, we assume the asymmetric student-teacher setup: given paired views X_S, X_T, the teacher outputs a target $T \sim p(\cdot \mid X_T)$, and the student f outputs a representation R for X_S. To predict the target T from X_S, a projection head maps the representation R to a predictive distribution $q(\cdot \mid X_S)$ over the target space. In practice, T is a *collection of discrete variables* as discussed later.

InfoMax objective The information we would like to capture in the representation R is the student-teacher shared mutual information (MI) $I(X_S; X_T)$. The teacher exposes information about its view X_T through the target T. By the data processing inequality, we have $I(X_S; X_T) \geq I(X_S; T) \geq I(R; T)$. This yields the *multi-view InfoMax* objective of maximizing $I(R; T)$, i.e., the shared information of the representation R with the target T (see Fig. 1a). We can maximize this objective using a variational lower bound (Barber & Agakov, 2003), where $H(T)$ is the entropy of the target and D_{KL} is the Kullback-Leibler divergence (all derivations in App. A.3.3):

$$I(R; T) \geq H(T) + \mathbb{E}[\log q(T \mid X_S)] \tag{1}$$

$$= \underbrace{I(X_T; T)}_{\text{anti-collapse}} - \underbrace{\mathbb{E}_{X_T, X_S} D_{\mathrm{KL}}\left(p(T \mid X_T) \,||\, q(T \mid X_S)\right)}_{\text{alignment}}. \tag{2}$$

The lower bound showcases two elements intuitively important for any SSL algorithm: an *anti-collapse term*, measuring the amount of information the targets T retain about the input X_T, and an *alignment term*, measuring how well the student approximates the teacher. In practice, we achieved the best performance with the teacher an EMA of the student, and therefore only optimize Eq. (2) with respect to the student parameters. Thus, the anti-collapse term $I(X_T; T)$ on the teacher side is not directly maximized, which we resolve in Secs. 3.2.2 and 3.2.3.

Granularity of the objective The KL term in Eq. (2) is flexible regarding the granularity of the targets; for instance, we could align the two views through a single image-level target. Instead, to

leverage more information, we model the targets as a collection of L elements $T = (T^{(1)}, \dots, T^{(L)})$, where ℓ indexes the targets by patch and feature layer. In this way, the training signal is both spatial and stems from features from different depths of the model. In the following, we assume that student and teacher distributions factorize over the elements. The alignment term in Eq. (2) then decomposes into a sum of element-wise alignment terms.

Preventing collapse From Eq. (2), we can see that collapse occurs when $I(X_T; T) = H(T) - H(T \mid X_T)$ goes towards 0. Preventing collapse thus involves maintaining targets with high *marginal entropy* $H(T)$ and low *conditional entropy* $H(T \mid X_T)$. Intuitively, the targets should be spread around the data distribution, while remaining predictable for any particular input. For collections of elements, $I(X_T; T)$ decomposes as

$$I(X_T; T) = \sum_{\ell=1}^{L} \underbrace{I(X_T; T^{(\ell)})}_{\text{per-element info}} - \underbrace{\mathrm{TC}(T)}_{\text{redundancy}} + \underbrace{\mathrm{TC}(T \mid X_T)}_{0}, \tag{3}$$

where the *total correlation* $\mathrm{TC}(T)$ measures the redundancy among elements and $TC(T \mid X_T) = 0$ by assuming a factorized $p(T \mid X_T)$. Hence, preventing collapse can be achieved by maintaining high per-element information while limiting the uncontrolled growth of redundancy (Secs. 3.2.2 and 3.2.3, respectively).

3.2 JEM: A New Joint-Embedding Multi-View Objective

In this section, we discuss our SSL method JEM. Motivated by maximizing the multi-view InfoMax objective, JEM uses a principled loss function consisting of an alignment and regularization terms derived from Eqs. (2) and (3). We provide an overview of our method in Fig. 2.

3.2.1 Patch-Wise Alignment Loss

For a given image, we produce a pair of overlapping student-teacher views x_S, x_T. Both views are placed on the same patch grid so overlapping patches correspond one-to-one. We denote $\mathcal{O}_S$ as the set of indices of student patch-layer pairs also appearing in the teacher view, and $m \colon \mathcal{O}_S \mapsto \mathcal{O}_T$ the function mapping the student indices to the corresponding teacher patch-layer pairs with indices $\mathcal{O}_T$. This choice of mapping enforces an invariance to the index of the elements. The alignment loss $\mathcal{L}_{\text{align}}$ is the element-wise decomposition of the KL in Eq. (2) over $\mathcal{O}_S$:

$$\mathcal{L}_{\text{align}}(x_S, x_T) = \frac{1}{|\mathcal{O}_S|} \sum_{\ell \in \mathcal{O}_S} D_{\text{KL}} \left(p(T^{(m(\ell))} \mid x_T) \,||\, q(T^{(\ell)} \mid x_S) \right). \tag{4}$$

3.2.2 Preventing Global Collapse

To maximize the lower-bound in Eq. (2), our goal is to keep $I(X_T; T)$ large. As we use discrete targets, directly estimating $I(X_T; T)$ is hard for multi-variate T. Instead, we maximize element-wise information terms $I(X_T; T^{(\ell)})$, with ℓ the element index, using the decomposition in Eq. (3). We implement this with an explicit regularization loss on the student. Through EMA updates, the student acts as a proxy for the teacher. In particular, we maximize $I(X_S; S^{(\ell)}) = H(S^{(\ell)}) - H(S^{(\ell)} \mid X_S)$, where $S^{(\ell)}$ follows the K-class categorical $q(\cdot \mid Z_S^{(\ell)})$ parametrized by the student head output $Z_S^{(\ell)}$ for patch ℓ. Given head outputs $\mathbf{z}^{(\ell)}$ for a batch of N inputs $\mathbf{x}$, we can compute an estimate of the element-wise mutual information as:

$$\hat{I}(\mathbf{x}; \mathbf{z}^{(\ell)}) = \underbrace{-\sum_{j=1}^{K} \bar{q}_j \log \bar{q}_j}_{\text{marg. entropy}} \underbrace{-\left(-\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} q_j(z_i^{(\ell)}) \log q_j(z_i^{(\ell)})\right)}_{\text{cond. entropy}}, \tag{5}$$

where $\bar{q}_j = \frac{1}{N} \sum_{i=1}^{N} q_j(z_i^{(\ell)})$ denotes the mean batch probability for class j and patch ℓ. The global collapse regularization loss is then given as:

$$\mathcal{R}_{\text{global}}(\mathbf{x}) = -\frac{1}{L} \sum_{\ell=1}^{L} \hat{I}(\mathbf{x}; \mathbf{z}^{(\ell)}). \tag{6}$$

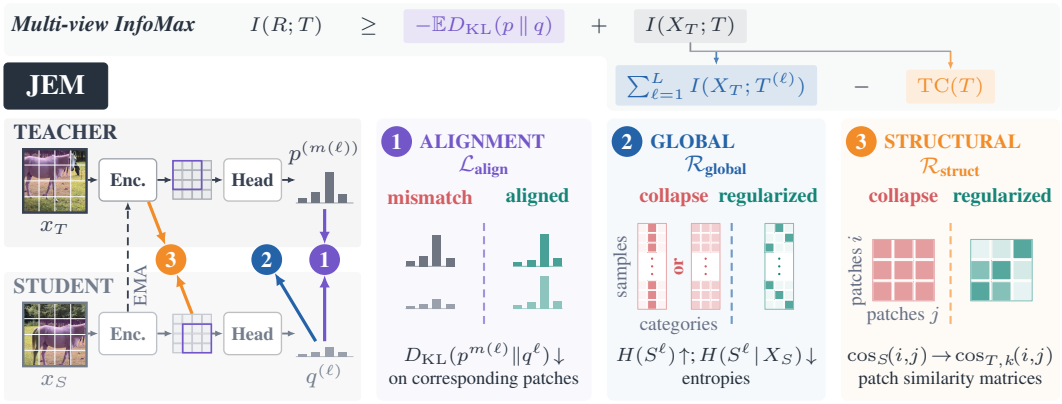


Figure 2: **From multi-view InfoMax to JEM.** Maximizing the variational lower bound on the multi-view InfoMax objective in Eq. (2) yields two requirements: cross-view alignment between teacher and student distributions p and q, and informative, non-collapsed targets $T \sim q(\cdot \mid X_T)$. JEM implements this with three loss terms: the KL term $\mathcal{L}_{\text{align}}$ aligns student patch ℓ with matching teacher target $m(\ell)$; $\mathcal{R}_{\text{global}}$ maintains per-patch mutual information through marginal and conditional entropy; and $\mathcal{R}_{\text{struct}}$ seeks to control the total correlation of the targets $\mathrm{TC}(T)$ by preserving the geometry of teacher features of layer k in the final student representation. Please zoom for details.

However, this regularization term only partially covers the anti-collapse term $I(X_T;T)$ in Eq. (2) due to the element-wise computation of $I(X_S;S^{(\ell)})$. We therefore define an additional regularization term which covers the *structural* collapse mode.

3.2.3 Preventing Structural Collapse

In the last section, we derived a regularizer to maintain the first term of Eq. (3), $\sum_{\ell=1}^{L} I(X_T;T^{(\ell)})$, preventing global collapse of the targets. In this section, we address the second term, the redundancy $\mathrm{TC}(T)$. Intuitively, the redundancy becomes large when all $T^{(\ell)}$ encode similar information, which we term a *structural collapse*. For example, if each patch target encodes the same input-dependent category, the jointly-encoded information $I(X_T;T)$ is the same as that of a single target $I(X_T;T^{(\ell)})$. Maximizing the sum of element-wise information $I(X_T;T^{(\ell)})$ does not stop this collapse mode; thus, we design a mechanism to prevent the redundancy from growing too large.

Estimating the total correlation $\mathrm{TC}(T)$ of a large set of discrete variables is intractable. Instead, we encourage the representation R to retain the dependency structure contained in the input, using this as a proxy for *controlling* $TC(T)$. As the true dependencies are inaccessible, we use the similarities of representations from different layers of the teacher as a proxy; this rests on the assumption that different layers encode different levels of structure. The representation must then match these similarities after a learned non-linear transformation. Specifically, given student representation r, and teacher representation h_k and transformation g_k at layer k, the loss for a pair of elements at indices i, j can be written as

$$\mathcal{L}_{\text{struct}}^{(i,j)}(r, h_k) = \left(\cos(g_k(r^{(i)}), g_k(r^{(j)})) - \cos(h_k^{(m(i))}, h_k^{(m(j))})\right)^2 \tag{7}$$

where cos denotes the cosine similarity and (i, j) are restricted to the subset of elements coming from the same layer of $\mathcal{O}_S \times \mathcal{O}_S$. Intuitively, by being able to reconstruct the early layer similarities, the representation avoids the collapse to a single, redundant vector, and is still free to encode higher-level semantic structure such as similar object patches. Given a set of target teacher layers $\mathcal{K}$, the total loss is then averaged over target layers and patch pairs:

$$\mathcal{R}_{\text{struct}}(x_S, x_T) = \frac{1}{|\mathcal{K}||\mathcal{O}_S|(|\mathcal{O}_S| - 1)} \sum_{k \in \mathcal{K}} \sum_{i \in \mathcal{O}_S} \sum_{\substack{j \in \mathcal{O}_S \\ i \neq j}} \mathcal{L}_{\text{struct}}^{(i,j)}(r, h_k). \tag{8}$$

3.3 View Construction

While often relegated to implementation details, the algorithm used to construct views of the input is a critical design choice in SSL (Chen et al., 2020; Tian et al., 2020b; He et al., 2022). For images, one obvious way to create views is cropping (Bachman et al., 2019; Chen et al., 2020), commonly combined with photometric transformations (Chen et al., 2020; Grill et al., 2020). The resulting crops overlap to varying degrees, leading to context-invariant representations. Additionally, Caron et al. (2020; 2021) exploit smaller local crops, which induce local-to-global correspondences. Masking-based methods, on the other hand, typically use fully overlapping student and teacher crops (He et al., 2022; Assran et al., 2022), where masking controls which patches are visible to the encoder.

From the perspective of the multi-view assumption, *cropping* and *masking* are complementary mechanisms for controlling the information shared between views. In this work, we treat them in a unified manner. We build several student views which partially overlap with the teacher view and are, in addition, masked, such that the views differ both in spatial support and patch visibility. The alignment loss of Eq. (4) is applied to *all overlapping patches*, whether masked or visible. Note that traditional masked prediction is a special case of this construction, obtained with a single student view and the student/teacher crops coinciding. Interestingly, we found that the method performs well without the use of masks, showing that patch-level SSL is possible using only cropping (see Sec. 4.2).

Concretely, we implement this design by sampling shifted but overlapping student and teacher crops aligned on a common patch grid, as shown in Fig. 3. We integrate *local crops* (Caron et al., 2020) by sampling smaller student than teacher crops, obtaining local-to-global prediction pairs. For a fraction of student views, we sample and apply a *patch mask*. Student and teacher views are also independently transformed with a standard set of photometric transformations (Grill et al., 2020). Further implementation details are provided in App. A.3.1.

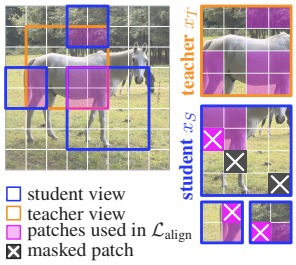


Figure 3: **View construction.**

4 Experiments

This section presents the experimental results validating JEM. We first provide our implementation details (Sec. 4.1) and ablate the individual components (Sec. 4.2). We evaluate the quality of the resulting features against models from the literature (Sec. 4.3) and study scaling behaviour (Sec. 4.4).

4.1 Implementation details

Architecture Our encoder is a Vision Transformer (ViT) (Dosovitskiy et al., 2021), at three model scales, ViT-L, ViT-g and ViT-7B. All models use 16×16 patches, omit the class token, and prepend four learned register tokens (Darcet et al., 2024). We follow the open source implementation of Siméoni et al. (2026), and inherit most hyperparameters from the published DINOv3 configurations.

Training Details For each image, we sample two view sets, each containing a global student view, four local student views and one teacher view. Global and teacher views are 256×256 pixels and local views 112×112. In 80% of the global and local views we mask 65% and 50% of the patches respectively; the rest are left unmasked. We use $K = 4096$ classes for the categorical distributions. We train ViT-g for 500k iterations with AdamW, a global batch size of 3072, and constant weight decay 0.04. The learning rate is warmed up linearly to 10^{-4} over 50k iterations. The teacher is an exponential moving average of the student with fixed momentum 0.994. The global collapse term $\mathcal{R}_{\text{global}}$ has a coefficient of 0.1 with a linear decay from 1 over the first 12.5k steps. The structural collapse term $\mathcal{R}_{\text{struct}}$ has a coefficient increased linearly from 0.05 to 1 over the course of training, with a linear warm-up from 0 over the first $12.5k$ steps. More hyperparameter details in App. A.3.2.

We first train on ImageNet-1k (Russakovsky et al., 2015) and ImageNet-22k (Deng et al., 2009) datasets. For larger scale, we follow Fu et al. (2026) and reproduce the curated LVD-142M dataset (Oquab et al., 2024), denoted "A140M" (App. A.3.5 for details).

Table 1: **Ablation study.** We isolate the contributions of view construction (a), the composition of the patch-level KL alignment loss (b), and the terms in the complete objective (c). We train ViT-L models on ImageNet-22k for 250k steps and report attention-probe accuracy on IN1k and segmentation mIoU on ADE20k. Tables compare the full model with variants removing (w/o) individual components.

(a) View construction.

Configuration	IN1k	ADE
Full JEM *model*	**82.8**	**48.1**
w/o independ. aug.	74.8	31.4
w/o local crops	81.2	46.2
w/o masking	82.3	45.3
w/o stud.-teach. shift	82.8	46.6

(b) KL alignment.

Configuration	IN1k	ADE
Full JEM *model*	82.8	48.1
w/o multi-layer loss	81.7	48.3
w/o visible-patch loss	70.8	28.1
w/o masked-patch loss	81.3	46.0
w/o EMA	83.0	47.0

(c) Objective.

Configuration	IN1k	ADE
Full JEM *model*	**82.8**	**48.1**
w/o $\mathcal{L}_{\text{align}}$	63.1	22.4
w/o $\mathcal{R}_{\text{global}}$	10.3	0.5
w/o $\mathcal{R}_{\text{struct}}$	82.3	48.0

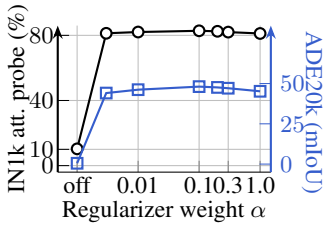


Figure 4: **Impact of the global regularization weight** on IN-1k and ADE20k performance. Models use ViT-L, IN-22k, and 250k steps.

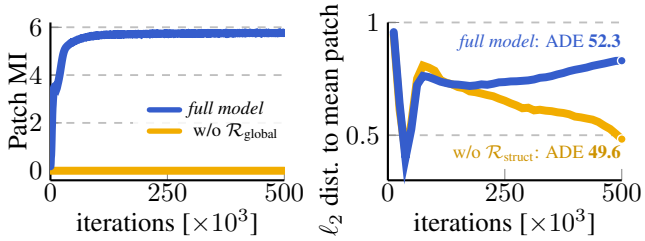


Figure 5: **Collapse prevention.** (Left) Teacher patch mutual information (MI) collapses when disabling $\mathcal{R}_{\text{global}}$. (Right) Mean ℓ_2-distance between patches and the image mean patch representation; low means loss of spatial structure. ViT-g on A140M.

Evaluation Protocol We evaluate all methods with a frozen backbone and simple probing setups. For classification, we report attention probe accuracy on ImageNet-1k (Russakovsky et al., 2015). For semantic segmentation, we report linear probe mIoU on ADE20k (Zhou et al., 2019), Pascal VOC (Everingham et al., 2010), and Cityscapes (Cordts et al., 2016); for monocular depth, we report linear probe RMSE on NYUv2 (Silberman et al., 2012) and KITTI (Geiger et al., 2012). We also consider harder benchmarks testing combined class- and instance understanding, and report mask AP for instance segmentation on COCO (Lin et al., 2014), and PQ for panoptic segmentation on COCONut (Deng et al., 2024) and ADE20k. See App. A.4.1 for details.

4.2 Ablations

In this section, we study the impact of the different components of JEM. In all experiments, if not stated otherwise, we train ViT-L-sized models on the ImageNet-22k dataset for 250k iterations. We evaluate on ImageNet-1k (IN1k) classification and ADE20k (ADE) semantic segmentation.

View construction (Tab. 1a) Removing per-crop-independent augmentations impacts our approach the most and greatly degrades the performance on both global and dense tasks. The other choices—using masking, shifts between teacher and student crops, and the use of local crops—are less influential on ViT-L scale, but still needed for the best results, mostly on dense tasks. We have also observed this effect to be more pronounced at larger model scale. Overall, we highlight the robustness to changes in the view construction: JEM can deliver strong performance without using masked patches or local crops, vital components of other methods (Caron et al., 2021; Oquab et al., 2024).

Alignment loss (Tab. 1b) Multi-layer prediction, masked-patch prediction, and an EMA teacher are not individually essential for strong performance: removing any one of them changes IN1k accuracy by at most 1.5 points and ADE20k mIoU by at most 2.1 points. This robustness is consistent with the InfoMax formulation: core requirements are only cross-view alignment and informative, non-collapsed targets. However, it is important to use visible patches in the alignment loss; removing their contribution costs 20 points on ADE20k. We hypothesize this is because it is important to train under the teacher's input condition to ground the targets.

Table 2: **Comparison to prior SSL methods**. Classification reports IN-1k attention-probe accuracy. Semantic, instance, and panoptic segmentation results are mIoU, mAP, and PQ, respectively. Depth reports RMSE (lower better). Tokens is the approximate number of tokens seen during training in trillions. '*' denotes the DINOv2 algorithm trained with the JEM setup for 125k steps on IN-1k and 500k steps on IN-22k or A140M. † denotes backbone training or fine-tuning above 256 px.

				Class.	Sem. Seg.			Inst. Seg.	Pan. Seg.		Depth	
Model	Size	Dataset	Tokens (T)	IN1k	ADE	VOC	City.	COCO	ADE	COCONut	NYU	KITTI
MAE	L	IN-1k	0.40	78.4	31.6	66.1	53.1	6.0	6.5	9.7	0.489	2.903
MSN	L	IN-1k	0.43	76.3	24.3	55.0	46.4	3.2	4.4	6.7	0.656	3.696
data2vec 2.0	L	IN-1k	0.60	79.5	27.4	52.7	48.6	2.1	4.8	7.1	0.512	2.901
CAPI	L	IN-1k	2.10	82.9	31.9	66.0	50.6	10.7	12.7	14.7	**0.361**	**2.604**
DINOv2*	L	IN-1k	0.23	**83.8**	45.7	83.3	66.5	12.1	18.0	22.1	0.408	2.718
JEM	L	IN-1k	0.23	83.0	**46.6**	**84.7**	**67.3**	**12.5**	**19.7**	**24.4**	0.404	2.723
I-JEPA	g	IN-22k	0.12	78.8	34.2	73.6	52.4	6.3	7.4	10.0	0.471	2.994
Franca	g	IN-22k	1.74	86.0	45.9	81.9	67.0	11.9	13.9	16.4	0.355	2.505
DINOv2*	g	IN-22k	1.39	**86.5**	50.0	83.6	68.5	13.0	18.1	21.1	**0.343**	2.494
JEM	g	IN-22k	1.39	85.5	**51.2**	**85.9**	**69.6**	**14.7**	**22.4**	**25.7**	0.363	**2.472**
>1B, trained on large datasets												
DINOv2*	g	A140M	1.39	86.9	50.5	83.6	69.4	12.7	18.7	21.4	0.339	2.455
DINOv2 (reg.)†	g	LVD-142M	1.83	87.3	49.5	83.1	69.1	12.3	16.7	18.6	0.357	2.569
LingBot-Vision†	g	Unk.-163M	1.86	86.5	53.6	85.9	74.1	15.2	22.2	25.0	0.287	2.127
JEM	g	A140M	1.39	85.6	52.3	86.1	71.2	14.7	22.9	26.1	0.329	2.324
V-JEPA 2.1†	G	VM-163M	0.24	83.2	47.8	84.3	68.3	12.5	15.5	19.6	**0.277**	2.167
WebSSL	7B	MC-8B	7.23	87.5	42.2	75.7	61.5	9.3	9.9	13.7	0.460	2.945
DINOv2*	7B	A140M	1.85	87.7	51.9	84.5	71.1	14.0	17.9	22.0	0.289	2.218
DINOv3†	7B	LVD-1689M	4.06	**89.0**	**55.9**	86.7	**75.2**	**16.0**	21.6	27.0	0.278	2.242
JEM	7B	A140M	1.85	86.9	54.4	**87.1**	73.6	15.6	**23.9**	**27.3**	0.297	**2.125**

Objective (Tab. 1c) We ablate the three terms of our objective. First, removing the *alignment loss* $\mathcal{L}_{\text{align}}$ (Eq. (4)) sharply reduces IN1k accuracy from 82.8 to 63.1 and ADE20k mIoU from 48.1 to 22.4, showing that the alignment loss constitutes a vital training signal. Second, removing *global collapse regularization* $\mathcal{R}_{\text{global}}$ (Eq. (6)) is catastrophic, and leads the model to collapse, attaining 10.3 IN1k accuracy. In Fig. 4, we also examine how sensitive downstream performance is to the weighting coefficient of this regularization. We observe that the concrete weight matters only in that it must be non-zero: a weight of 0 yields degenerate representations, while any value across a wide range performs comparably, with best performance obtained at 0.1. Third, we find the effect of *structural regularization* $\mathcal{R}_{\text{struct}}$ (Eq. (8)) to be marginal at ViT-L scale (-0.5 on IN1k), but crucial when scaling the model size further. In the next paragraph, we investigate its effect with such (larger) model sizes, and find that it is important to prevent the progressive loss of spatial diversity.

Preventing collapse (Fig. 5) Now we consider larger models (ViT-g) which suffer more and faster from collapse. First, in Fig. 5 (left), we show the evolution of the per-patch MI $I(X_T; T^{(\ell)})$ over training when disabling global regularization $\mathcal{R}_{\text{global}}$; in this case, the target distributions collapse to a uniform or constant distribution over categories. Second, in Fig. 5 (right), we show how the average patch ℓ_2-distance to the per-image mean patch develops over training with or without the structural regularizer. Low scores indicate that patch representations become more spatially uniform. Without regularization, increasing degradation starts after 250k iterations, leading to performance degradation. In contrast, the regularized model keeps patch representations distinct from their spatial average, thereby maintaining stronger performance on dense tasks (+2.7 ADE20k mIou).

4.3 Comparison to Other Algorithms

In Tab. 2, we compare JEM to other SSL methods on the same data, training ViT-L on Imagenet-1k and ViT-g on Imagenet-22k. We also report results with models trained on larger, not directly comparable datasets and discuss them in Sec. 4.4. For a controlled comparison, we re-train DINOv2 models (Oquab et al., 2024) (noted "DINOv2*"), using the same setup as JEM (similar to training

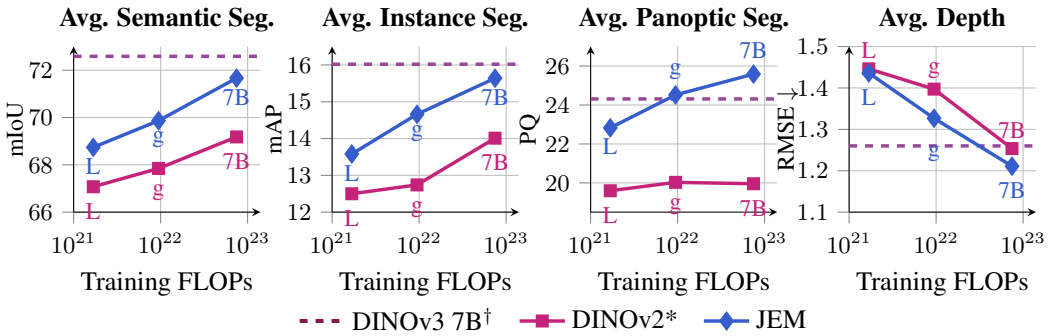


Figure 6: **Scaling model size.** Performance as a function of approximate training compute, averaged over datasets for each benchmark type. DINOv2* and JEM are trained by us for 500k steps on A140M. Dashed purple lines show DINOv3 ViT-7B/16 results also averaged.

DINOv3 without gram and high-resolution fine-tuning steps). This baseline is relevant as DINOv2, combining global and local losses, serves as a foundation for many of today's best SSL methods.

For both model sizes, JEM outperforms all other methods on six out of nine tasks, with consistent gains across the full range of dense prediction: from coarse semantic segmentation to fine-grained instance and panoptic segmentation. On global classification, we observe a small gap to DINOv2; still a strong result as DINOv2 was heavily designed to obtain the best global performance. Interestingly, JEM scales effortlessly to ViT-g. Moreover, its advantage over DINOv2 increases at scale, e.g. with a 4 PQ pt gain on both challenging panoptic segmentation tasks; the gap is maintained at 7B (Sec. 4.4).

4.4 Scaling JEM

An important aspect of modern ML algorithms is their scalability with model size and data. In SSL, scaling is often challenging due to feedback dynamics of self-supervision; different model scales show different behavior (Siméoni et al., 2026), and approaches that work well on small scales do not necessarily train well (or at all) beyond 1B parameters. Here we show that JEM scales well.

We train 300M (ViT-L), 1B (ViT-g), and 7B (ViT-7B) parameter models with JEM and DINOv2 on our A140M dataset for 500k steps. Fig. 6 plots task-averaged performance as a function of training FLOPs (individual benchmarks in appendix Fig. 7). We also include DINOv3 ViT-7B results (dashed line), though we note it is trained for twice as long on a 10× larger dataset and with additional refinement steps (gram and high-resolution fine-tuning). JEM trains stably even at 7B scale, with performance improving significantly when model size grows. To our knowledge, this is the first exclusively patch-level latent-space method that trains well at this scale. Compared to the DINOv2 baseline, JEM exhibits comparable or better scaling behavior and outperforms it at all model scales.

Tab. 2 also lists models above 1B parameters trained on different large datasets which are thus not directly comparable. Notably, JEM outperforms the best SSL method DINOv3 on several datasets in semantic and panoptic segmentation, as well as on depth estimation, despite DINOv3's longer training, larger dataset and its refinement steps (gram anchoring and high resolution fine-tuning)—showing the strength of our training paradigm. We also plot qualitative PCA and patch cosine similarity results in Figs. 9 and 10 and highlight the quality of the patch features, obtained without any refinement steps.

5 Conclusion

In this work, we introduced JEM, a principled and stable patch-level joint-embedding multi-view method that learns strong representations. JEM is derived from a multi-view InfoMax objective motivated by the multi-view assumption. This objective leads to three loss terms: one aligning representations across views, one preventing global collapse and one preserving structural integrity. JEM trains stably from 300M to 7B parameters—to our knowledge, the first exclusively patch-level latent-space method demonstrated at this scale—and achieves impressive results on segmentation tasks while being competitive on classification and depth estimation. These results suggest that multi-objective SSL's complexity is not necessary for strong performance, and establish principled patch-level objectives as a promising foundation for scalable visual representation learning.

AI USE STATEMENT

In this work, we used generative AI tools to help implement methods, help to develop conceptual frameworks, to formulate mathematical claims and assist in the writing of proofs. We have not used generative AI tools to provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, support qualitative and thematic data analysis, interpret results and generating synthetic data sets and cleaning and reformatting datasets are not applicable to this work. Additionally, we used generative AI tools to aid or polish writing, for the retrieval and discovery of related work, for creating or modifying scientific figures or images, to launch experiments and evaluations, and to create and edit software code. We have reviewed all AI-assisted work; implementations and proofs were manually checked for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017. URL https://openreview.net/forum?id=HyxQzBceg.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *ECCV*, pp. 456–473, 2022. doi: 10.1007/978-3-031-19821-2_26. URL https://link.springer.com/chapter/10.1007/978-3-031-19821-2_26.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pp. 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *NeurIPS*, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/ddf354219aac374f1d40b7e760ee5bb7-Abstract.html.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, pp. 1298–1312, 2022. URL https://proceedings.mlr.press/v162/baevski22a.html.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *ICML*, pp. 1416–1429, 2023. URL https://proceedings.mlr.press/v202/baevski23a.html.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics, 2025. URL https://arxiv.org/abs/2511.08544.

David Barber and Felix V. Agakov. Information maximization in noisy channels: A variational approach. In *NeurIPS*, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/hash/a6ea8471c120fe8cc35a2954c9b9c595-Abstract.html.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022a. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICRegL: Self-supervised learning of local visual features. In *NeurIPS*, pp. 8799–8810, 2022b. doi: 10.52202/068431-0640. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/39cee562b91611c16ac0b100f0bc1ea1-Abstract-Conference.html.

John S. Bridle, Anthony J. R. Heading, and David J. C. MacKay. Unsupervised classifiers, mutual information and 'phantom targets'. In *NeurIPS*, pp. 1096–1101, 1991. URL https://proceedings.neurips.cc/paper_files/paper/1991/hash/a8abb4bb284b5b27aa7cb790dc20f80b-Abstract.html.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9650–9660, 2021. doi: 10.1109/ICCV48922.2021.00951. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Mathilde Caron, Neil Houlsby, and Cordelia Schmid. Location-aware self-supervised transformers for semantic segmentation. In *WACV*, pp. 116–126, 2024. doi: 10.1109/WACV57701.2024.00019. URL https://openaccess.thecvf.com/content/WACV2024/html/Caron_Location-Aware_Self-Supervised_Transformers_for_Semantic_Segmentation_WACV_2024_paper.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, 34:17864–17875, 2021.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, pp. 3213–3223, 2016. doi: 10.1109/CVPR.2016.350. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Cordts_The_Cityscapes_Dataset_CVPR_2016_paper.html.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Timothée Darcet, Federico Baldassarre, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Cluster and predict latent patches for improved masked image modeling. *TMLR*, 2025. URL https://openreview.net/forum?id=Ycmz7qJxUQ.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Xueqing Deng, Qihang Yu, Peng Wang, Xiaohui Shen, and Liang-Chieh Chen. COCONut: Modernizing COCO segmentation. In *CVPR*, pp. 21863–21873, 2024. doi: 10.1109/CVPR52733.2024.02065. URL https://openaccess.thecvf.com/content/CVPR2024/html/Deng_COCONut_Modernizing_COCO_Segmentation_CVPR_2024_paper.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 27, 2014.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4. URL https://link.springer.com/article/10.1007/s11263-009-0275-4.

David Fan, Shengbang Tong, Jiachen Zhu, Koustuv Sinha, Zhuang Liu, Xinlei Chen, Michael Rabbat, Nicolas Ballas, Yann LeCun, Amir Bar, and Saining Xie. Scaling language-free visual representation learning. In *ICCV*, 2025. URL https://arxiv.org/abs/2504.01017.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *ICLR*, 2020. URL https://openreview.net/forum?id=B1xwcyHFDr.

Zelin Fu, Bin Tan, Changjiang Sun, Shaohui Liu, Kecheng Zheng, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Vision pretraining for dense spatial perception, 2026. URL https://arxiv.org/abs/2607.05247.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, pp. 3354–3361, 2012. doi: 10.1109/CVPR.2012.6248074. URL https://doi.org/10.1109/CVPR.2012.6248074.

Ryan Gomes, Andreas Krause, and Pietro Perona. Discriminative clustering by regularized information maximization. In *NeurIPS*, pp. 775–783, 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/hash/42998cf32d552343bc8e460416382dca-Abstract.html.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pp. 9729–9738, 2020. doi: 10.1109/CVPR42600.2020.00975. URL https://openaccess.thecvf.com/content_CVPR_2020/html/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022. doi: 10.1109/CVPR52688.2022.01553. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *ICML*, pp. 4182–4192. PMLR, 2020. URL https://proceedings.mlr.press/v119/henaff20a.html.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019. URL https://openreview.net/forum?id=Bklr3j0cKX.

Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *ICML*, pp. 1558–1567, 2017. URL https://proceedings.mlr.press/v70/hu17b.html.

Xu Ji, João F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, pp. 9864–9873, 2019. doi: 10.1109/ICCV.2019.00996. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Ji_Invariant_Information_Clustering_for_Unsupervised_Image_Classification_and_Segmentation_ICCV_2019_paper.html.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, pp. 9396–9405. IEEE, 2019.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *AISTATS*, pp. 562–570, 2015. URL https://proceedings.mlr.press/v38/lee15a.html.

Yihao Li, Saeed Salehi, Lyle Ungar, and Konrad P. Kording. Does object binding naturally emerge in large pretrained vision transformers? In *NeurIPS*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/050b8ff31bee2dfea65b731e71baccd5-Abstract-Conference.html.

Zhong-Yu Li, Shanghua Gao, and Ming-Ming Cheng. SERE: Exploring feature self-relation for self-supervised transformer. *IEEE TPAMI*, 45(12):15619–15631, 2023. doi: 10.1109/TPAMI.2023.3309979. URL https://ieeexplore.ieee.org/document/10234504.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

Ralph Linsker. Self-organization in a perceptual network. *IEEE Computer*, 21(3):105–117, 1988. doi: 10.1109/2.36. URL https://ieeexplore.ieee.org/document/36.

Scott C. Lowe, Anthony Fuller, Sageev Oore, Evan Shelhamer, and Graham W. Taylor. Self-distillation of hidden layers for self-supervised representation learning, 2026. URL https://arxiv.org/abs/2603.15553.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-JEPA 2.1: Unlocking dense features in video self-supervised learning. In *ECCV*, 2026. URL https://arxiv.org/abs/2603.14482.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. URL https://openreview.net/forum?id=GLm1BA3C8p.

Pedro O. Pinheiro, Amjad Almahairi, Ryan Y. Benmalek, Florian Golemo, and Aaron Courville. Unsupervised learning of dense visual representations. In *NeurIPS*, pp. 4489–4500, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/3000311ca56a1cb93397bc676c0b7fff-Abstract.html.

Sucheng Ren, Fangyun Wei, Samuel Albanie, Zheng Zhang, and Han Hu. DeepMIM: Deep supervision for masked image modeling. In *WACV*, pp. 879–888, 2025. doi: 10.1109/WACV61041.2025.00095. URL https://openaccess.thecvf.com/content/WACV2025/html/Ren_DeepMIM_Deep_Supervision_for_Masked_Image_Modeling_WACV_2025_paper.html.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *ICLR*, 2015. URL https://arxiv.org/abs/1412.6550.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://link.springer.com/article/10.1007/s11263-015-0816-y.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. In *ICLR*, 2019.

Ravid Shwartz-Ziv and Yann LeCun. To compress or not to compress—self-supervised learning and information theory: A review. *Entropy*, 26(3):252, 2024. doi: 10.3390/e26030252. URL https://www.mdpi.com/1099-4300/26/3/252.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGB-D images. In *ECCV*, pp. 746–760, 2012. doi: 10.1007/978-3-642-33715-4_54. URL https://link.springer.com/chapter/10.1007/978-3-642-33715-4_54.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *TMLR*, 2026. URL https://openreview.net/forum?id=2NlGyqNjns.

Karthik Sridharan and Sham M. Kakade. An information theoretic framework for multi-view learning. In *Conference on Learning Theory*, pp. 403–414. Omnipress, 2008. URL https://dblp.org/rec/conf/colt/SridharanK08.

Thomas Stegmüller, Tim Lebailly, Behzad Bozorgtabar, Tinne Tuytelaars, and Jean-Philippe Thiran. CrOC: Cross-view online clustering for dense visual representation learning. In *CVPR*, pp. 7000–7009, 2023. doi: 10.1109/CVPR52729.2023.00676. URL https://openaccess.thecvf.com/content/CVPR2023/html/Stegmuller_CrOC_Cross-View_Online_Clustering_for_Dense_Visual_Representation_Learning_CVPR_2023_paper.html.

Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. In *CVPR*, pp. 2132–2141, 2023. doi: 10.1109/CVPR52729.2023.00212. URL https://openaccess.thecvf.com/content/CVPR2023/html/Tao_Siamese_Image_Modeling_for_Self-Supervised_Vision_Representation_Learning_CVPR_2023_paper.html.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, pp. 776–794, 2020a. doi: 10.1007/978-3-030-58621-8_45. URL https://link.springer.com/chapter/10.1007/978-3-030-58621-8_45.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In *NeurIPS*, 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/4c2e5eaae9152079b9e95845750bb9ab-Abstract.html.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Allerton Conf. on Communication, Control, and Computing*, pp. 368–377, 1999. URL https://arxiv.org/abs/physics/0004057.

Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *ICLR*, 2021. URL https://openreview.net/forum?id=-bdp_8Itjwp.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *ICLR*, 2020. URL https://openreview.net/forum?id=rkxoh24FPH.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. SCAN: Learning to classify images without labels. In *ECCV*, pp. 268–285, 2020. doi: 10.1007/978-3-030-58607-2_16. URL https://link.springer.com/chapter/10.1007/978-3-030-58607-2_16.

Shashanka Venkataramanan, Valentinos Pariza, Mohammadreza Salehi, Lukas Knobel, Elias Ramzi, Spyros Gidaris, Andrei Bursuc, and Yuki M. Asano. Franca: Nested matryoshka clustering for scalable visual representation learning. In *CVPR*, pp. 10533–10544,

2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Venkataramanan_Franca_Nested_Matryoshka_Clustering_for_Scalable_Visual_Representation_Learning_CVPR_2026_paper.html.

Feng Wang, Tao Kong, Rufeng Zhang, Huaping Liu, and Hang Li. Self-supervised learning by estimating twin class distributions. *IEEE Transactions on Image Processing*, 2023a. URL https://arxiv.org/abs/2110.07402.

Haoqing Wang, Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhi-Hong Deng, and Kai Han. Masked image modeling with local multi-scale reconstruction. In *CVPR*, pp. 2122–2131, 2023b. doi: 10.1109/CVPR52729.2023.00211. URL https://openaccess.thecvf.com/content/CVPR2023/html/Wang_Masked_Image_Modeling_With_Local_Multi-Scale_Reconstruction_CVPR_2023_paper.html.

Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. In *NeurIPS*, pp. 16423–16438, 2022. doi: 10.52202/068431-1195. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6818dcc65fdf3cbd4b05770fb957803e-Abstract-Conference.html.

Ziyang Wu, Jingyuan Zhang, Druv Pai, Xudong Wang, Chandan Singh, Jianwei Yang, Jianfeng Gao, and Yi Ma. Simplifying DINO via coding rate regularization. In *ICML*, pp. 68036–68059, 2025. URL https://proceedings.mlr.press/v267/wu25ar.html.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *CVPR*, pp. 16684–16693, 2021. doi: 10.1109/CVPR46437.2021.01641. URL https://openaccess.thecvf.com/content/CVPR2021/html/Xie_Propagate_Yourself_Exploring_Pixel-Level_Consistency_for_Unsupervised_Visual_Representation_Learning_CVPR_2021_paper.html.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, pp. 12310–12320, 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *IJCV*, 127(3):302–321, 2019. doi: 10.1007/s11263-018-1140-0. URL https://link.springer.com/article/10.1007/s11263-018-1140-0.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *ICLR*, 2022. URL https://openreview.net/forum?id=ydopy-e6Dg.

Adrian Ziegler and Yuki M. Asano. Self-supervised learning of object parts for semantic segmentation. In *CVPR*, pp. 14502–14511, 2022. doi: 10.1109/CVPR52688.2022.01410. URL https://openaccess.thecvf.com/content/CVPR2022/html/Ziegler_Self-Supervised_Learning_of_Object_Parts_for_Semantic_Segmentation_CVPR_2022_paper.html.

A APPENDIX

In this appendix, we provide additional analyses, implementation details, and experiments for JEM.

Specifically:

- **Additional discussion**. In App. A.1, we show how existing self-supervised methods can be interpreted from the perspective of the lower bound of the multi-view InfoMax objective, and discuss prior work more in depth.
- **Additional results**. In App. A.2, we report scaling trends on individual benchmarks.
- **Additional details**. In App. A.3, we provide the full derivations of our objective, further details on the structural collapse regularizer and on view construction, the complete set of training hyperparameters, and details about the training data.
- **Evaluation details**. In App. A.4, we describe the evaluation protocol of each benchmark, including the implementation and reproduction of competing methods, and list the evaluated model checkpoints.
- **Qualitative examples**. In App. A.5, we report additional qualitative feature visualizations of JEM and competing models.

A.1 RELATIONSHIP TO EXISTING SSL PARADIGMS

In this section, we discuss how existing self-supervised methods can be interpreted from the multi-view InfoMax perspective. Further, we relate JEM to prior works in which some of its key components also appear, namely: explicit entropy regularization of discrete targets, dense losses between partially overlapping views, multi-layer targets, and structural regularization, and discuss how our formulation differs.

How Different Methods Fit the Multi-View InfoMax Objective Self-clustering methods such as SwAV (Caron et al., 2020), DINO (Caron et al., 2021), or iBOT (Zhou et al., 2022) use discrete categorical targets T, minimizing a cross entropy loss between student and teacher views that matches the KL divergence in Eq. (2) up to a constant. For the anti-collapse term, centering or Sinkhorn-Knopp balancing of the targets plays the role of maintaining $H(T)$; temperature sharpening makes sure that $H(T \mid X_T)$ does not grow too large. In SimDINO (Wu et al., 2025), the targets are continuous vectors, with the minimized mean-squared-error (MSE) loss again matching the KL divergence; $H(T)$ is maintained by maximizing the log-determinant of the covariance matrix of the head outputs, and $H(T \mid X_T = x)$ is minimal for all x. Similarly, redundancy-reduction methods (Zbontar et al., 2021; Bardes et al., 2022a; Balestriero & LeCun, 2025) combine cross-view alignment with output-distribution regularization that can be seen as an indirect proxy for maintaining $H(T)$.

A principled view of the DINO family of methods. Owing to its strong performance, the DINO family of self-distillation methods (Caron et al., 2021; Zhou et al., 2022; Oquab et al., 2024; Siméoni et al., 2026) has become a common foundation for many recent methods in the literature (Wu et al., 2025; Fan et al., 2025; Venkataramanan et al., 2026). As discussed above, these methods can be interpreted as instantiations of the multi-view InfoMax objective, in which the cross-entropy between student and teacher implements the alignment term, while the anti-collapse term is only maintained implicitly, through mechanisms that were introduced empirically (Caron et al., 2020; 2021). In JEM, each component of the training objective instead follows from the lower bound in Eq. (2), which simplifies the recipe of DINOv2 in three ways. (i) *A single loss and single head for the final layer.* DINOv2 applies the DINO loss to the class token and the iBOT loss to the masked patches, with separate heads, and adds the KoLeo regularizer (Sablayrolles et al., 2019), a nearest-neighbor estimator of the differential entropy of the class-token features, which acts as a further proxy for $H(T)$. JEM instead applies the same alignment and anti-collapse terms to all elements of T, through one head at each supervised depth. (ii) *Explicit regularization instead of heuristics.* Sinkhorn-Knopp balancing/centering and temperature sharpening indirectly control the marginal entropy $H(T)$ and the conditional entropy $H(T \mid X_T)$, respectively. JEM replaces them with $\mathcal{R}_{\text{global}}$, which optimizes both quantities explicitly, so that student and teacher can share the same constant temperature, without a teacher temperature schedule. (iii) *No prototype freezing.* DINO and DINOv2 freeze the last layer of the head, which holds the prototypes, during the first iterations of training to stabilize it; JEM

requires no such mechanism. Moreover, the class token fits naturally into our information-theoretic derivation as *one more element* of T, to which the same alignment and anti-collapse terms apply. From this perspective, DINO corresponds to the special case in which the class token is the only element of T, and iBOT and DINOv2 to adding the masked patches as further elements, each with its own loss and head.

Explicit entropy regularization of discrete assignments. For discrete targets, the InfoMax anti-collapse term $I(X_T; T) = H(T) - H(T \mid X_T)$ corresponds to maximizing the entropy of the average prediction while minimizing the entropy of each individual prediction. Early works (Bridle et al., 1991; Gomes et al., 2010) employ explicit mutual information maximization for training unsupervised classifiers. More recently, unsupervised clustering methods have employed similar regularization techniques. SCAN (Van Gansbeke et al., 2020) maximizes the entropy of the mean prediction to avoid collapse of the cluster assignments, without an explicit term for the sharpness of individual predictions, which is only encouraged implicitly by its consistency loss between neighboring samples. IIC (Ji et al., 2019) instead directly maximizes the mutual information between the cluster assignments of two views, without explicit regularization for each input's assignment. IMSAT (Hu et al., 2017) includes both marginal and conditional entropy terms, combined with an alignment loss to enforce augmentation invariance. Among self-clustering methods, MSN (Assran et al., 2022) optimizes the marginal entropy, but also applies Sinkhorn-Knopp balancing to the targets and resorts to temperature sharpening to control the conditional entropy.

More similar to ours, TWIST (Wang et al., 2023a) employs an alignment loss between teacher and student assignment distributions, alongside the entropy regularization terms. It is the closest image-level analogue of our $\mathcal{R}_{\text{global}}$ combined with $\mathcal{L}_{\text{align}}$. In contrast to these works, our $\mathcal{R}_{\text{global}}$ is applied to the discrete assignments of every patch in a patch-level student-teacher framework. Moving from per-image to per-patch assignment also introduces two collapse modes that entropy terms pooled over all patches cannot control: *structural collapse* and *positional collapse*. The latter occurs when the model simply encodes the position of each patch in the grid. The marginal entropy over *all* patches can be increased simply by encoding the patch position. Instead, computing the marginal entropy separately for each patch position leads to estimating $I(X_T; T \mid \text{Pos})$, which prevents this failure mode. In our derivation, this formulation naturally follows the decomposition in Eq. (3). Among patch-level methods, CAPI (Darcet et al., 2025) identifies the same positional collapse and alleviates it implicitly, by running the Sinkhorn-Knopp balancing of the targets separately at each patch position. On the other hand, the *structural collapse* failure mode entails patches within an image falling to an input-dependent constant distribution, effectively encoding the same image-level content. This motivates our structural regularizer $\mathcal{R}_{\text{struct}}$ (Sec. 3.2.3). Both of these terms have no counterpart in the image-level methods mentioned above.

Dense losses between partially overlapping views. A large family of dense SSL methods is based on masked image modeling, which predicts the content of masked patches, either in pixel (He et al., 2022), or latent space (Baevski et al., 2022; Assran et al., 2023). In pure masking methods, input and target fully overlap, whereas in our formulation we treat masking and cropping as complementary strategies to control information sharing between views. Therefore, we utilize *partially overlapping* views for teacher and student, additionally masking the student views. Another category of methods applies joint-embedding losses between corresponding locations of two partially overlapping crops, but without masking. Among these, PixPro (Xie et al., 2021) and VADeR (Pinheiro et al., 2020) are pixel-level contrastive methods. Several patch-level clustering-based methods also rely on partially overlapping crops, aligning the patch assignments of the overlapping region with RoIAlign (Ziegler & Asano, 2022; Wen et al., 2022) or clustering the patches of both views jointly (Stegmüller et al., 2023). VICRegL (Bardes et al., 2022b) applies a similar idea to continuous embeddings, matching each feature to its spatial nearest neighbor in the other view. A few methods combine the two families, applying masked modeling across partially overlapping views, like ours. LOCA (Caron et al., 2024) trains a student on masked query crops to predict teacher cluster assignments for a larger reference crop, where correspondences are interpolated. SIM (Tao et al., 2023) uses a decoder to predict, from the visible patches of a masked view, the features of *all* patches of a different view. Differently from all these prior works, we sample student and teacher crops at integer patch offsets on a shared patch lattice, so that overlapping patches correspond exactly one-to-one; we found that interpolating targets between misaligned patch grids degrades the resulting dense features. Moreover, we treat masked and visible patches uniformly, applying the alignment loss to all overlapping patches.

Multiple layers as targets. Supervising intermediate layers dates back to deeply-supervised networks (Lee et al., 2015), and to intermediate layer distillation (Romero et al., 2015). In SSL, data2vec (Baevski et al., 2022; 2023) utilizes projections to regress the top-K layers of an EMA teacher from the final layer of the student. Among masking models, DeepMIM (Ren et al., 2025) and LocalMIM (Wang et al., 2023b) add masked prediction losses on intermediate encoder layers, using as targets either pixels, image gradients, or teacher features. More recently, V-JEPA 2.1 (Mur-Labadia et al., 2026) fuses together the features of several student layers, for a shared predictor to regress the EMA-teacher features at the different layers, for both masked and visible tokens. Similarly, Bootleg (Lowe et al., 2026) extends I-JEPA by regressing, from the final student layer through a single predictor, the concatenated EMA-teacher features at the masked locations from different layers. In JEM, targets are indexed by patch *and* layer: each supervised student layer k predicts the categorical distribution computed by the EMA teacher at the *same* depth k, so that the prediction at depth k only depends on the first k layers, and the alignment and anti-collapse terms are applied at every depth. In addition, $\mathcal{R}_{\text{struct}}$ uses intermediate teacher layers as targets for the *final* student layer, but only through their similarity structure (see next paragraph).

Structural regularization and Gram anchoring. Our structural regularizer $\mathcal{R}_{\text{struct}}$ (Sec. 3.2.3) addresses the redundancy term $\mathrm{TC}(T)$ of the anti-collapse objective, whose uncontrolled growth leads to *structural collapse*, in which all patches of an image encode the same image-level content. We implement this regularizer by matching the pairwise cosine similarities between patches of a set of early-to-middle teacher layers from a projection of the features from the final student layer. Using pairwise similarities as a target has also been employed by other methods. In SSL, SERE (Li et al., 2023) extends iBOT with a patch self-relation loss, training the similarities between the final-layer patch features of the student to match those of the EMA teacher on another view. More recently, DINOv3 (Siméoni et al., 2026) adopts gram anchoring, to address the different phenomenon of the local patch structure becoming increasingly noisy with long training. In DINOv3, after $1M$ iterations, an earlier checkpoint is used as a *gram teacher*, to anchor the similarity matrix of the final-layer patch features of the student. Differently, our targets come from *early-to-middle* teacher layers and are only matched after a learned projection, so the final representation is not *forced* to share the geometry of the earlier layers, but must retain the information required to recover it. Moreover, our $\mathcal{R}_{\text{struct}}$ is applied from the start of training, uniformly on visible and masked patches, as well as local crops.

A.2 Additional Results

Fig. 7 reports the individual benchmark results underlying the task averages in Fig. 6, together with ImageNet-1k classification accuracy.

A.3 Additional Details

A.3.1 View Construction

Given an image x, we sample student view x_S and teacher view x_T as follows. We first extract a *super crop* with pre-determined scale and random aspect ratio from the image x and resize it to a square set of pixels with the side length being a multiple of the patch size. The scale is chosen such that an overlap of at least 10% is achieved between the student and teacher crops when randomly picking the scale of the teacher crop from [0.32, 1.0] of the area of the original image. Then, at random patch-grid aligned positions, we extract square student and teacher crops with a size matching the input resolution of the model. A random horizontal flip is applied to the whole image before cropping, such that all views share the same orientation. Each crop is then independently transformed with standard photometric augmentations (Grill et al., 2020; Caron et al., 2021), i.e. color jittering, grayscale conversion, Gaussian blur, and solarization. Different from DINO (Caron et al., 2021), where the teacher processes the same augmented global crops as the student, teacher crops are augmented independently of their student crops.

For 80% of the student crops, we additionally sample a patch mask M that is applied to the student view, for which we use multi-block masking (Zhou et al., 2022; Oquab et al., 2024) with cyclic shifts (Darcet et al., 2025; Venkataramanan et al., 2026) for the global crops, and uniform random masking for the local crops. For global crops, we mask 65% of the patches. For local crops, we mask 50% of the patches.

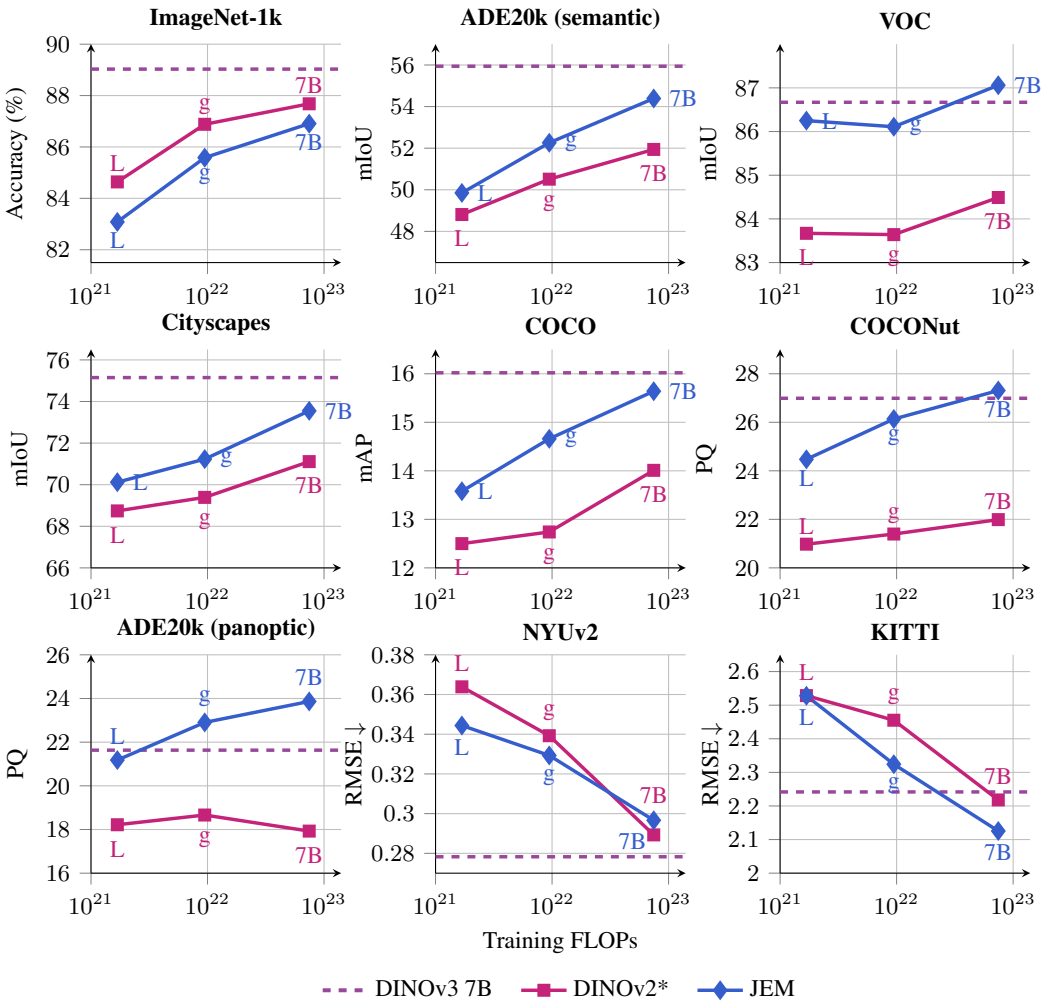


Figure 7: **Scaling across individual benchmarks.** Performance as a function of approximate training compute. DINOv2* and JEM are trained by us on A140M for 500k steps. Dashed lines show DINOv3 ViT-7B/16 performance. ADE20k is evaluated with both semantic segmentation (mIoU) and panoptic segmentation (PQ) probes.

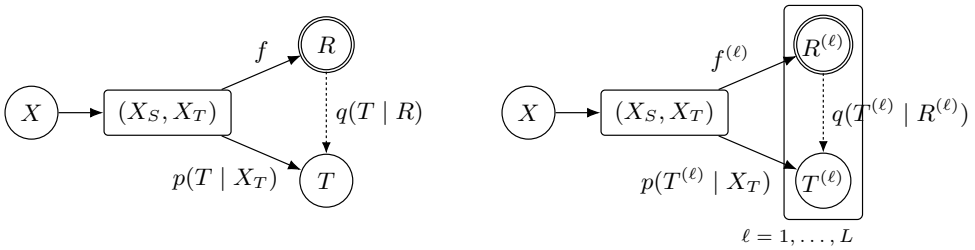
A.3.2 Training Details

We closely build both the JEM and the DINOv2 baseline implementation and hyperparameters on the open source implementation of DINOv3 Siméoni et al. (2026). Our self-trained DINOv2 baseline can thus be seen as a version of DINOv3 without gram anchoring and high-resolution refinement stages. We give the backbone architectures and training hyperparameters shared by JEM and the DINOv2 baselines in Tab. 3, and their remaining training hyperparameters in Tabs. 4 and 5, respectively.

The loss of JEM is a weighted combination of (i) alignment loss (Eq. (4)) on visible ($\mathcal{L}^{\text{vis}}_{\text{align}}$) and masked ($\mathcal{L}^{\text{msk}}_{\text{align}}$) patches, (ii) global collapse regularization loss $\mathcal{R}_{\text{global}}$ (Eq. (6)) and (iii) structural regularization loss $\mathcal{R}_{\text{struct}}$ (Eq. (8)):

$$\mathcal{L} = w_{\text{vis}} \cdot \mathcal{L}^{\text{vis}}_{\text{align}} + w_{\text{msk}} \cdot \mathcal{L}^{\text{msk}}_{\text{align}} + w_{\text{global}} \cdot \mathcal{R}_{\text{global}} + w_{\text{struct}} \cdot \mathcal{R}_{\text{struct}}. \tag{9}$$

In practice, we use different loss coefficients for targets from different layers (see Tab. 4), and apply the structural regularizer only to the final layer representation.



(a) Generic model.

(b) Element-wise model.

Figure 8: **Student-teacher graphical models.** (a) shows the generic model with the form of T not specified; (b) shows the element-wise model with additional factorization assumptions. The views (X_S, X_T) are sampled jointly given input X. Solid arrows describe the sampling model, with $R = f(X_S)$ and $T \sim p(\cdot \mid X_T)$; double circles mark deterministic nodes. Dashed arrows denote the variational predictors $q(T \mid R)$ and $q(T^{(\ell)} \mid R^{(\ell)})$. In (b), ℓ indexes matched patch/layer pairs, with teacher targets relabeled accordingly and $R^{(\ell)} = f^{(\ell)}(X_S)$. The plate represents the factorizations $p(T \mid X_T) = \prod_{\ell=1}^{L} p(T^{(\ell)} \mid X_T)$ and $q(T \mid R) = \prod_{\ell=1}^{L} q(T^{(\ell)} \mid R^{(\ell)})$.

A.3.3 Derivations

We assume X_S and X_T are jointly generated from X. For the student, $R = f(X_S)$, where f is a deterministic function, and thus $q(t \mid x_S) := q(t \mid f(x_S))$. The teacher generates T from X_T, resulting in the joint distribution $p(x_S, x_T, t) = p(x_S, x_T)\, p(t \mid x_T)$. See Fig. 8 for graphical models.

Lower bound on mutual information (Eq. (1)) This is the lower bound from Barber & Agakov (2003). Starting with the definition of mutual information, we have:

$$
\begin{aligned}
I(R;T) &= H(T) - H(T \mid R) && (10)\\
&= H(T) + \mathbb{E}_{R,T} \log p(T \mid R) && (11)\\
&= H(T) + \mathbb{E}_{R,T} \log q(T \mid R) + \mathbb{E}_{R,T} \log \frac{p(T \mid R)}{q(T \mid R)} && (12)\\
&= H(T) + \mathbb{E}_{R,T} \log q(T \mid R) + \underbrace{\mathbb{E}_R D_{\mathrm{KL}}\left(p(\cdot \mid R) \,\|\, q(\cdot \mid R)\right)}_{>0} && (13)\\
&\geq H(T) + \mathbb{E}_{X_S,T} \log q(T \mid X_S). && (14)
\end{aligned}
$$

KL form of lower bound (Eq. (2)) Expanding the expected KL divergence:

$$
\begin{aligned}
&\mathbb{E}_{X_S,X_T} D_{\mathrm{KL}}\left(p(\cdot \mid X_T) \,\|\, q(\cdot \mid X_S)\right) && (15)\\
&\quad = \mathbb{E}_{X_S,X_T} \mathbb{E}_{T \sim p(\cdot \mid X_T)} \left[\log p(T \mid X_T) - \log q(T \mid X_S)\right] && (16)\\
&\quad = \mathbb{E}_{X_T} \mathbb{E}_{T \sim p(\cdot \mid X_T)} \log p(T \mid X_T) - \mathbb{E}_{X_S} \mathbb{E}_{T \sim p(\cdot \mid X_T)} \log q(T \mid X_S) && (17)\\
&\quad = -H(T \mid X_T) - \mathbb{E}_{X_S,T} \log q(T \mid X_S). && (18)
\end{aligned}
$$

It follows that $\mathbb{E}_{X_S,T} \log q(T \mid X_S) = -D_{\mathrm{KL}}\left(p(\cdot \mid X_T) \,\|\, q(\cdot \mid X_S)\right) - H(T \mid X_T)$. Plugging this into Eq. (1) gives

$$
\begin{aligned}
I(R;T) &\geq H(T) + \mathbb{E}_{X_S,T} \log q(T \mid X_S) && (19)\\
&= I(X_T;T) - D_{\mathrm{KL}}\left(p(\cdot \mid X_T) \,\|\, q(\cdot \mid X_S)\right). && (20)
\end{aligned}
$$

Element-wise decomposition of mutual information (Eq. (3)) With element-wise targets $T = (T^{(1)}, \ldots, T^{(L)})$, we have that

$$
H(T) = \sum_{\ell=1}^{L} H(T^{(l)}) - \mathrm{TC}(T), \tag{21}
$$

where $\mathrm{TC}(T) := D_{\mathrm{KL}}\left((T^{(1)},\dots,T^{(L)}) \,\|\, \prod_{\ell=1}^{L} T^{(\ell)}\right)$ is the total correlation measuring dependence between the elements of T. The total correlation becomes zero if and only if the variables are mutually independent; it is an extension of mutual information for more than two variables. Then,

$$\begin{aligned} I(X_T;T) &= H(T) - H(T \mid X_T) \\ &= \sum_{\ell=1}^{L} H(T^{(\ell)}) - \mathrm{TC}(T) - \left(\sum_{\ell=1}^{L} H(T^{(\ell)} \mid X_T) - \mathrm{TC}(T \mid X_T)\right) \\ &= \sum_{\ell=1}^{L} H(T^{(\ell)}) - \sum_{\ell=1}^{L} H(T^{(l)} \mid X_T) - (\mathrm{TC}(T) - \mathrm{TC}(T \mid X_T)) \\ &= \sum_{\ell=1}^{L} I(X_T;T^{(\ell)}) - (\mathrm{TC}(T) - \mathrm{TC}(T \mid X_T))\,. \end{aligned} \tag{22}$$

Element-wise decomposition of KL divergence (Sec. 3.2.1) We want to show how the KL alignment term in Eq. (2) decomposes for element-wise targets. Let $T = (T^{(1)},\dots,T^{(L)})$, and assume that the teacher and student distributions factorize conditionally over elements:

$$p(t \mid x_T) = \prod_{\ell=1}^{L} p(t^{(\ell)} \mid x_T), \qquad q(t \mid x_S) = \prod_{\ell=1}^{L} q(t^{(\ell)} \mid x_S). \tag{23}$$

For any fixed pair of views (x_S, x_T), we then have

$$\begin{aligned} D_{\mathrm{KL}}\left(p(\cdot \mid x_T) \,\|\, q(\cdot \mid x_S)\right) &= \mathbb{E}_{T\sim p(\cdot\mid x_T)}\left[\log \frac{p(T \mid x_T)}{q(T \mid x_S)}\right] \\ &= \mathbb{E}_{T\sim p(\cdot\mid x_T)}\left[\sum_{\ell=1}^{L} \log \frac{p(T^{(\ell)} \mid x_T)}{q(T^{(\ell)} \mid x_S)}\right] \\ &= \sum_{\ell=1}^{L} \mathbb{E}_{T^{(\ell)}\sim p_\ell(\cdot\mid x_T)}\left[\log \frac{p(T^{(\ell)} \mid x_T)}{q(T^{(\ell)} \mid x_S)}\right] \\ &= \sum_{\ell=1}^{L} D_{\mathrm{KL}}\left(p(\cdot \mid x_T) \,\|\, q(\cdot \mid x_S)\right). \end{aligned} \tag{24}$$

The third equality marginalizes out all elements other than $t^{(\ell)}$, since the logarithm depends only on that element. Taking the expectation over paired views gives

$$\mathbb{E}_{X_T,X_S} D_{\mathrm{KL}}(p(\cdot \mid X_T) \,\|\, q(\cdot \mid X_S)) = \sum_{\ell=1}^{L} \mathbb{E}_{X_T,X_S} D_{\mathrm{KL}}\left(p(\cdot \mid X_T) \,\|\, q(\cdot \mid X_S)\right). \tag{25}$$

A.3.4 Structural Collapse Regularizer

We detail the structural collapse regularizer introduced in Sec. 3.2.3. For student and teacher views x_S, x_T, we define $\mathcal{O}_S$ and $\mathcal{O}_T$, the sets of patch indices of student/teacher that geometrically overlap with each other, and $m\colon \mathcal{O}_S \mapsto \mathcal{O}_T$, the function mapping the student indices to the spatially corresponding teacher patch. Given student representation r and teacher representation h_k at layer k, the loss for a pair of elements at indices i, j can be written as

$$\mathcal{L}_{\text{struct}}^{(i,j)}(r, h_k) = \left(\cos(g_k(r^{(i)}), g_k(r^{(j)})) - \cos(h^{(m(i))}, h^{(m(j))})\right)^2, \quad (i,j) \in \mathcal{O}_S \times \mathcal{O}_S, \tag{26}$$

where g_k is a learned non-linear projection for the kth teacher layer. Thus, to compute the loss, only geometrically corresponding pairs of patches are used. Given a set of target teacher layers $\mathcal{K}$, the total loss is then averaged over target layers and patch pairs:

$$\mathcal{R}_{\text{struct}}(x_S, x_T) = \frac{1}{|\mathcal{K}||\mathcal{O}_S|(|\mathcal{O}_S|-1)} \sum_{k\in\mathcal{K}} \sum_{i\in\mathcal{O}_S} \sum_{\substack{j\in\mathcal{O}_S \\ i\neq j}} \mathcal{L}_{\text{struct}}^{(i,j)}(r, h_k). \tag{27}$$

Hyperparameter	Model size		
	ViT-L	ViT-g	ViT-7B
Backbone parameters	303.1M	1.13B	6.7B
Transformer blocks	24	40	40
Embedding dimension	1024	1536	4096
Attention heads	16	24	32
Attention-head dimension	64	64	128
FFN type	MLP	SwiGLU	SwiGLU
FFN hidden dimension	4096	4096	8192
Patch size		16×16	
Storage tokens		4	
Positional encoding		RoPE	
LayerScale initialization		1×10^{-5}	
Drop-path rate	0.3	0.4	0.4

(a) Backbone architectures.

Hyperparameter	Model size		
	ViT-L	ViT-g	ViT-7B
Global crops		2×256^2	
Local crops		8×112^2	
Training iterations		500k	
Global batch size	2048	3072	4096
Optimizer		AdamW	
LR warm-up iterations		50k	
Layer-wise LR decay	0.9	0.995	0.98
LR schedule after warm-up	Cosine decay	Constant	Constant
AdamW β_1		0.9	
AdamW β_2	0.999	0.99	0.99
Initial weight decay		0.04	
Final weight decay	0.4	0.04	0.04
Teacher EMA momentum	$0.992 \rightarrow 1.0$	0.994	0.994
Parameter / reduction precision		BF16 / FP32	

(b) Shared training hyperparameters.

Table 3: **Shared architectures and training hyperparameters for JEM and DINOv2 baselines**. Entries spanning the three model columns are shared across all model sizes. Parameter counts include the student backbone only.

Hyperparameter	Model size		
	ViT-L	ViT-g	ViT-7B
Teacher crop scale		[0.32, 1.0]	
Number of categories		4096	
Student and teacher temperature		0.1	
Global mask probability / ratio		0.8 / 0.65	
Local mask probability / ratio		0.8 / 0.50	
Intermediate target blocks	{8,16}	{10,20,30}	{10,20,30}
Struct. regularizer target blocks	{0,8,16}	{0,10,20,30}	{0,10,20,30}
Struct. head output dimension		1024	
Proj./Struct hidden layers		3	
Proj./Struct heads hidden dimension	2048	2048	4096
Proj. head bottleneck dimension	256	384	384
Masked KL weight w_{masked}		0.8	
Visible KL weight w_{vis}		0.2	
Final layer weight		1	
Early layer weight per-layer	0.25	0.166	0.166
Global collapse regularization weight w_{global}		$1.0 \xrightarrow{\text{12.5k linear}} 0.1$	
Structural regularizer weight w_{struct} schedule		$0.0 \xrightarrow{\text{12.5k linear}} 0.05 \xrightarrow{\text{linear}} 1.0$	
Effective peak LR	1.697×10^{-3}	6.928×10^{-4}	1.200×10^{-3}
Effective final LR	1.697×10^{-6}	6.928×10^{-4}	1.200×10^{-3}

Table 4: **JEM training hyperparameters**. Entries spanning the three model columns are shared across all model sizes. Shared architectures and training hyperparameters are given in Tab. 3.

We use target layers [0, 8, 16] for ViT-L (24 total layers), and [0, 10, 20, 30] for ViT-g and ViT-7B (40 total layers), where layer 0 refers to the output of the patch embedding. The target features h_k are processed by an element-wise layernorm without affine parameters before computing the cosine similarity. We implement g_k with a MLP of the same size as the projection head, except that it maps to an output space of dimensionality 1024 and does not have the final linear projection to logits. The regularizer is applied both on global and local crops, and masked and unmasked patches.

Note that it is also possible to apply the regularizer only using student representations, and we successfully experimented with this option. This simplifies the implementation, as patches at the same index do directly geometrically correspond to each other. We opted for the slightly more complex teacher version as it allowed us to also apply the loss on the masked patches (which have no good similarity targets on the student), the EMA provides more stable targets, and the fact that the targets come from a different view may have an additional positive effect through enforcing a view invariance on the learned similarity structure.

Hyperparameter	Model size		
	ViT-L	ViT-g	ViT-7B
Global-crop scale		$[0.32,\ 1.00]$	
Local-crop scale		$[0.05,\ 0.32]$	
Global mask probability		0.5	
Mask-ratio range		$[0.1,\ 0.5]$	
DINO prototypes	65,536	131,072	262,144
iBOT prototypes	65,536	131,072	98,304
DINO/iBOT heads layers		3	
DINO head hidden dimension	2048	2048	8192
iBOT head hidden dimension	2048	2048	4096
DINO head bottleneck dimension	256	384	512
iBOT head bottleneck dimension	256	384	384
Student temperature		0.1	
Teacher temperature		$0.04 \rightarrow 0.07$ over 50k iterations	
DINO		1.0	
iBOT		0.5	
KoLeo		0.1	
Effective peak LR	1.697×10^{-3}	4.157×10^{-4}	4.800×10^{-4}
Effective final LR	5.657×10^{-6}	4.157×10^{-4}	4.800×10^{-4}
Last-layer freeze		Initial 5k iterations	

Table 5: **DINOv2 baseline training hyperparameters**. Entries spanning all model columns are shared across model sizes. Shared architectures and training hyperparameters are given in Tab. 3. Loss weights are configured coefficients.

A.3.5 DATA

Following the protocol established by Oquab et al. (2024), we construct A140M by retrieving from a large pool images that align with the train splits of established public datasets, including IN-1k, IN-22k, ADE20K, etc. Our source pool comprises approximately 750M images, obtained by deduplicating an internal collection of 1.3 billion images. This process involves both self-deduplication which remove duplicates to improve retrieval quality, and relative-deduplication which prevents data contamination from the "test" splits of the seed datasets, which are used as standard benchmarks in evaluation. Consistent with Oquab et al. (2024), we also implement NSFW filtering and facial blurring.

A.4 EVALUATION DETAILS

A.4.1 BENCHMARK PROTOCOLS

We evaluate models using standard lightweight probing setups. Following Siméoni et al. (2026), we compare models of different patch sizes at matched input sequence length to equalize the compute of the forward pass. All evaluations use final patch features of the backbone after the final layer norm.

Image Classification For classification, we apply the *single-layer attention probing* setup from Darcet et al. (2025). A learned query token is used to perform scaled multi-head attention (head dim=64) over key-value transformed features; the result is linearly projected to logits. We perform grid search on a validation set to select the best hyperparameters. Thereby, we scan the full cross product over learning rates from $(1 \cdot 10^{-5}, 2 \cdot 10^{-5}, 5 \cdot 10^{-5}, 1 \cdot 10^{-4}, 2 \cdot 10^{-4}, 5 \cdot 10^{-4}, 1 \cdot 10^{-3}, 2 \cdot 10^{-3}, 5 \cdot 10^{-3}, 1 \cdot 10^{-2})$ and weight decays from $(5 \cdot 10^{-4}, 1 \cdot 10^{-3}, 5 \cdot 10^{-2})$. The probe is trained with a batch size of 1024 with AdamW for 12500 steps, with linear learning rate warm-up of 1250 steps followed by cosine decay. The applied learning rate is linearly scaled with global batch as $\text{lr} = \frac{\text{global_bs}}{256} \cdot \text{base_lr}$.

For training on ImageNet, we split the original train set into training and validation sets, using 90% for training, and 10% for validation. We apply light random augmentations during training, specifically RandomResizedCrop and random horizontal flips. During evaluation, we resize the image to 256/292 pixels on the short side and take the central square crop of 224/256 pixels (87.5% zoom crop) for patch size 14/16. Models are evaluated on ImageNet val (Russakovsky et al., 2015). We report top-1 accuracy.

Semantic Segmentation Semantic segmentation is the task of assigning a class-label to every pixel of the input image. We use ADE20k (Zhou et al., 2019), PASCAL VOC (Everingham et al., 2010) and Cityscapes (Cordts et al., 2016) datasets to evaluate this task and use the mean intersection-over-union (mIoU) as the reported metric. We evaluate the task under two protocols with frozen backbones: kNN and linear probing. First, we use the kNN protocol for fast evaluations and assessment of the quality of dense patches. There, images and label maps are resized to 256×256 without cropping, using bicubic and nearest-neighbor interpolation respectively. The mIoU metric is computed on the full resized label maps. Second, for the linear probing, we follow Siméoni et al. (2026) and take patch tokens from the last layer after the backbone's final LayerNorm. On top of those, we train a linear head consisting of a BatchNorm layer followed by a 1×1 convolution that outputs N logits, where N is the number of classes in a given dataset. The logits are bilinearly upsampled to the label resolution and trained with pixel-wise cross-entropy. We scan the full cross product of learning rates $\{1 \cdot 10^{-4}, 3 \cdot 10^{-4}, 1 \cdot 10^{-3}\}$ and weight decays $\{1 \cdot 10^{-3}, 1 \cdot 10^{-2}\}$. All six probes are trained together on the same frozen features. Each probe is trained with AdamW optimizer (using $\beta_1 = 0.9$, $\beta_2 = 0.999$) and a batch size of 16 for 40,000 steps. The learning rate warms up linearly over 1,500 steps and then linearly decays to zero. During training, we resize each image so that its short side is $448/512 \cdot s$ pixels for models with patch size 14/16, with s drawn uniformly from [0.5, 2.0]. We then take a random $r \times r$ square crop adapted to 1024 tokens (i.e., $448^2/512^2$ crop for patch size 14/16) and apply random horizontal flips and photometric distortions (consisting of brightness, contrast, saturation and hue). During evaluation, we resize the image to 448/512 pixels on the short side and run $r \times r$ sliding-window inference with windows and a stride of 298/341 pixels. Logits are averaged where windows overlap and bilinearly upsampled to the original image resolution. We report the mIoU of the best probe over all hyperparameter configurations and evaluation steps.

Instance and Panoptic Segmentation Instance segmentation is the task of recognizing all pixels that belong to different object instances and their class-label. Panoptic segmentation (Kirillov et al., 2019) extends instance segmentation beyond objects: every pixel is assigned a class label and, for countable objects (*things*), an instance identity, while regions such as sky or road (*stuff*) are only labeled by class.

We evaluate instance segmentation on the COCO dataset (Lin et al., 2014), and report the standard COCO average precision metric. We evaluate panoptic segmentation on COCONut (Deng et al., 2024) and ADE20K-Panoptic (Zhou et al., 2019). For COCONut, we train on COCONut-S and evaluate on the relabeled COCO-val set with panoptic annotations; for ADE20K-Panoptic, we follow prior work (Cheng et al., 2021) and use the standard ADE20K training and validation splits. We report panoptic quality (PQ) (Kirillov et al., 2019), computed from IoU-based matching between predicted and ground-truth segments.

We use the same lightweight probe for both tasks, which operates on the last-layer patch features of the frozen backbone, taken after the final layer normalization. The probe is split into two independent modules: a patch-wise linear classifier that predicts the semantic category of each patch, as in semantic segmentation, and a patch-wise 2-layer MLP that is used to split the patches into different segments (instances). To get the masks, we compute the pairwise dot-product between the MLP outputs for different patches, whose sigmoid gives a score for the binary relation "patches belong to the same segment"; we supervise the MLP with the binary cross entropy loss against patch pairs derived from the ground truth masks. This is similar to the object probes used in Li et al. (2025).

During training, images are augmented with random resized crops to $448 \times 448/512 \times 512$ pixels for patch size 14/16, i.e., square crops of 1024 tokens; no other augmentation is used. Targets are defined at the patch level by taking the majority semantic and instance label within each patch. As in the instance segmentation setup, the probe is trained with a semantic cross-entropy loss, in which void patches are ignored, and a binary cross-entropy loss on the pairwise similarity predictions, with equal weights. We sweep learning rates over $\left\{1 \cdot 10^{-4}, 3 \cdot 10^{-4}, 1 \cdot 10^{-3}\right\}$ and weight decays over $\left\{1 \cdot 10^{-3}, 1 \cdot 10^{-2}\right\}$, training all six probes jointly on the same frozen features. Each probe is trained with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) and a batch size of 16 for 20 epochs, with a cosine learning-rate decay to 10^{-6} and no warm-up.

At inference, images are resized to the same square resolution without cropping. Each patch predicted as a non-void class c yields a candidate soft mask, given by its pairwise similarities to all patches multiplied by their probabilities for class c. These candidate masks are filtered using class-wise mask

NMS (IoU threshold 0.7), upsampled to the original image resolution, and binarized at 0.3; regions belonging to the same "stuff" class are merged. Metrics are computed at the original image resolution. Probes are evaluated after every epoch, and we select the best combination of learning rate, weight decay and epoch by validation PQ.

Depth Estimation Monocular depth estimation is the task of predicting a dense, per-pixel metric depth map from a single RGB image. It tests whether frozen features encode 3D scene geometry and not only semantics. We evaluate on two standard benchmarks: NYU Depth v2 (Silberman et al., 2012), which contains indoor scenes at 480×640 resolution, and KITTI (Geiger et al., 2012), which contains outdoor driving scenes of about 375×1242 pixels with sparse LiDAR ground truth. We report the root mean squared error (RMSE, in meters). It is computed per image over valid pixels and averaged over the test set. Valid pixels are those whose ground truth lies strictly inside $[10^{-3}, d_{\text{max}}]$, with $d_{\text{max}} = 10\,\text{m}$ on NYUv2 and $80\,\text{m}$ on KITTI.

To predict monocular depth, we train a linear probe on top of the frozen backbone. The last-layer patch features, taken after the final layer normalization, are bilinearly upsampled by a factor of 4 and passed through a trained batch normalization layer and a 1×1 convolution with 256 output channels. To convert the outputs to depth, we predict it by classification over bins: the 256 outputs are passed through a ReLU, shifted by 0.1 and normalized into weights over 256 bins uniformly spaced in $[10^{-3}, d_{\text{max}}]$. The predicted depth is the weighted sum of the bin centers. During training, predictions are bilinearly upsampled to the label resolution and trained with the scale-invariant log loss (Eigen et al., 2014). Similar to the linear semantic segmentation, we sweep learning rates over $\{1 \cdot 10^{-4}, 3 \cdot 10^{-4}, 1 \cdot 10^{-3}\}$ and weight decays over $\{1 \cdot 10^{-3}, 1 \cdot 10^{-2}\}$, training all six probes jointly on the same frozen features. Each probe is trained with AdamW and a batch size of 16 for 38,400 iterations, with a linear learning-rate warm-up over 12,800 iterations followed by cosine decay, with gradient norms clipped at 35.

Inputs are prepared in three steps: a fixed crop, an aspect-preserving resize, and then random square crops (training) or sliding windows (inference). On NYUv2, no fixed crop is applied: we use the full 480×640 frame. On KITTI, we keep the bottom 352 rows and the central 1216 columns of each frame. We then resize the image so that its short side is $448/480$ pixels on NYUv2 and $308/352$ pixels on KITTI for patch size 14/16. During training, we apply random rotations of up to $\pm 2.5^\circ$ and horizontal flips, each with probability 0.5, jointly to the image and the depth map (resized with nearest-neighbor interpolation). We then take a random square crop whose side equals the resized short side (448^2/480^2 on NYUv2, 308^2/352^2 on KITTI), and apply color augmentation (random gamma, brightness and per-channel color scaling) with probability 0.5.

At inference, the whole resized image is processed with square sliding windows of the training crop size, spanning the full image height, with strides of 298/319 pixels on NYUv2 and 205/234 pixels on KITTI for patch size 14/16. Bin logits are averaged where windows overlap, bilinearly upsampled to the ground-truth resolution, converted to depth, and averaged with the prediction for the horizontally flipped image. We report the configuration with the lowest RMSE, using the probe weights at the end of training.

A.4.2 Evaluated Model Checkpoints

In Tab. 2, we give an estimate of the total number of tokens used to train each model. For simplicity, we count full patch grids for all crops and mask variants, and exclude teacher and special tokens. Note that the I-JEPA ViT-g checkpoint is taken from epoch 44 as listed in the checkpoint file, even though it is denoted as `IN22K-vit.g.16-600e`.

Table 6 lists the external model checkpoints evaluated in this work. Parameter counts refer to the evaluated backbone; pretraining-only prediction heads and other auxiliary components are excluded.

A.5 Qualitative examples

We show visualizations of representations of a JEM ViT-g model trained on A140M in Fig. 9 and Fig. 10.

Model	Citation	Arch. / Variant	Training Data	Params.	External Checkpoint
MAE	He et al. (2022)	ViT-L/16	IN-1k	303M	`facebook/vit-mae-large`
MSN	Assran et al. (2022)	ViT-L/16	IN-1k	303M	`vitl16_600ep`
data2vec 2.0	Baevski et al. (2023)	ViT-L/16	IN-1k	303M	`large_imagenet`
I-JEPA	Assran et al. (2023)	ViT-g/16	IN-22k	1,011M	`IN22K-vit.g.16-600e`
DINOv2 (reg.)	Oquab et al. (2024)	ViT-g/14	LVD-142M	1,136M	`dinov2_vitg14_reg`
CAPI	Darcet et al. (2025)	ViT-L/14	IN-1k	303M	`capi_vitl14_in1k`
WebSSL	Fan et al. (2025)	ViT-7B/14	MC-8B	6,450M	`webssl_dino7b_full8b_224`
Franca	Venkataramanan et al. (2026)	ViT-g/14 (224 px; no RASA)	IN-22k	1,135M	`franca_vitg14_In21K.pth`
DINOv3	Siméoni et al. (2026)	ViT-7B/16	LVD-1689M	6,716M	`dinov3_vit7b16`
V-JEPA 2.1	Mur-Labadia et al. (2026)	ViT-G/16	VM-163M	1,843M	`vjepa2_1_vit_gigantic_384`
LingBot-Vision	Fu et al. (2026)	ViT-g/16	Unk.-163M	1,134M	`robbyant/lingbot-vision-vit-giant`

Table 6: External checkpoints evaluated in Tab. 2, grouped by method and ordered by first public appearance.

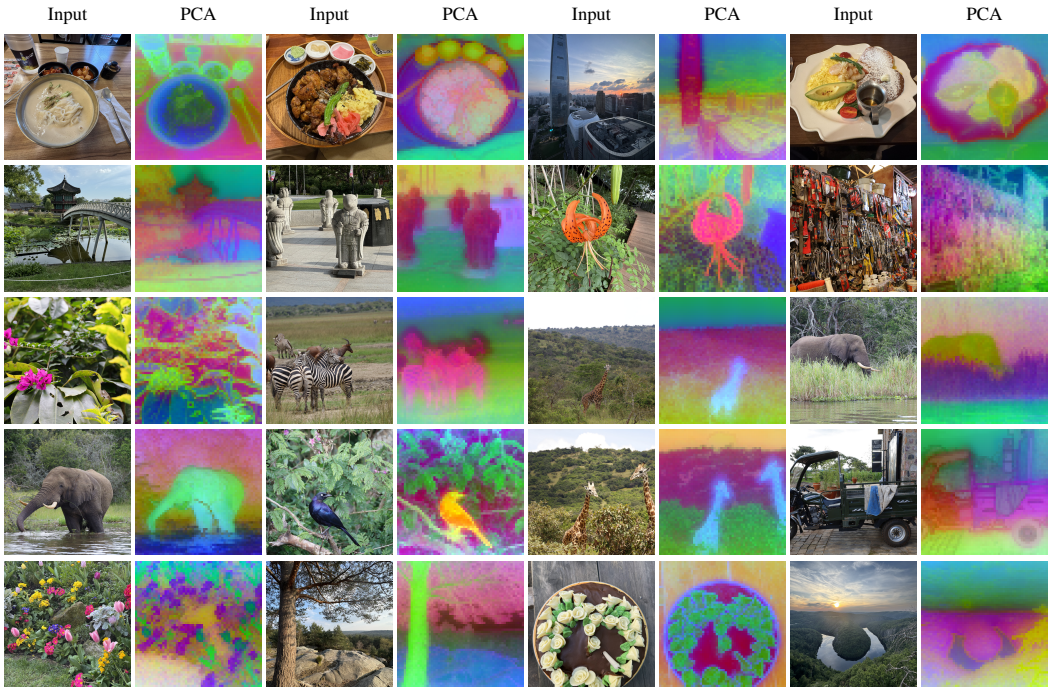


Figure 9: PCA visualization of final-layer patch representations from the multi-layer model at $4\times$ evaluation resolution. For each of the 20 images, we show the input and its RGB projection onto the first three principal components.

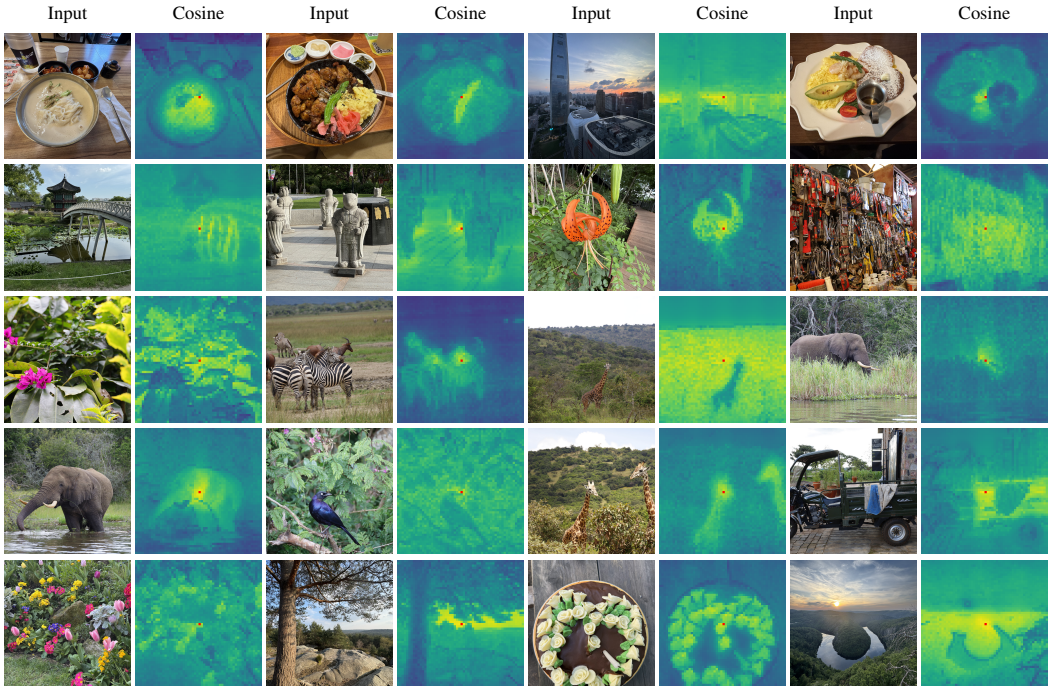


Figure 10: Cosine similarity between final-layer patch representations and the center-patch representation for the multi-layer model at $4\times$ evaluation resolution. Red markers identify the reference patch.