# IConE: Batch Independent Collapse Prevention for Self-Supervised Representation Learning

Konstantinos Almpanakis and Anna Kreshuk

European Molecular Biology Laboratory, Heidelberg, Germany
{konstantinos.almpanakis,anna.kreshuk}@embl.de

**Abstract.** Self-supervised learning (SSL) has revolutionized representation learning, with Joint-Embedding Architectures (JEAs) emerging as an effective approach for capturing semantic features. Existing JEAs rely on implicit or explicit batch interaction – via negative sampling or statistical regularization – to prevent representation collapse. This reliance becomes problematic in regimes where batch sizes must be small, such as high-dimensional scientific data, where memory constraints and class imbalance make large, well-balanced batches infeasible. We introduce **IConE** (**I**nstance-**Con**trasted **E**mbeddings), a framework that decouples collapse prevention from the training batch size. Rather than enforcing diversity through batch statistics, IConE maintains a global set of learnable auxiliary instance embeddings regularized by an explicit diversity objective. This transfers the anti-collapse mechanism from the transient batch to a dataset-level embedding space, allowing stable training even when batch statistics are unreliable, down to batch size 1. Across diverse 2D and 3D biomedical modalities, IConE outperforms strong contrastive and non-contrastive baselines throughout the small-batch regime (from $B = 1$ to $B = 64$) and demonstrates marked robustness to severe class imbalance. Geometric analysis shows that IConE preserves high intrinsic dimensionality in the learned representations, preventing the collapse observed in existing JEAs as batch sizes shrink.

**Keywords:** Self-supervised Representation Learning · Joint Embedding Architectures · Biomedical Imaging

## 1 Introduction

The success of Self-Supervised Learning (SSL) has led to powerful general-purpose vision models, such as DINOv2 [31] and its successor DINOv3 [36], which were transformative for downstream tasks across diverse domains [2, 6, 14, 28, 29]. A major line of work in SSL is based on reconstruction, where representations are learned by recovering input content from a corrupted observation [20]. In contrast, multi-view or joint-embedding approaches learn by enforcing consistency between multiple augmented views of the same instance while preventing representational collapse [5, 9, 11, 19]. Empirically, joint-embedding methods have proven highly effective for learning discriminative representations in vision [38].

Unlike natural 2D images, most scientific domains lack foundation models as the Internet-scale paradigm [31] does not readily apply. With new specialized modalities continuously emerging and no ImageNet [15] equivalent for 3D medical volumes or hyperspectral data, waiting for massive datasets is impractical. Consequently, self-supervised learning must be effective not just at scale, but in the small-data regimes typical of scientific discovery.

These regimes impose not only data scarcity but also strict computational constraints. In many scientific applications (*e.g.* volumetric CT/MRI, remote sensing, materials science), a single instance – such as a 3D volume or video sequence – can saturate GPU memory, limiting training to batch sizes of 1-2 samples. Conversely, increasing batch size to match standard SSL practice (*e.g.* $B > 256$) becomes problematic when datasets are small (*e.g.* $N < 5000$), as large batches relative to dataset size reduce gradient stochasticity and weaken the implicit regularization relied upon by many objectives [25]. Severe class imbalance further exacerbates small-batch instability, as individual batches may fail to represent minority classes. As a result, existing SSL methods often degrade in small-batch, small-data regimes [1, 10].

Theoretical analysis [40] provides further motivation for our work: contrastive SSL can be understood as aligning augmentation manifolds. In higher-dimensional spaces, these manifolds are inherently sparse, requiring stronger augmentations to induce the semantic overlap for representation learning. This demands stronger regularization to prevent collapse, creating a paradox: high-dimensional data requires robust regularization, but its massive memory footprint dictates small batches, precisely where batch-dependent regularization becomes unreliable.

Current joint-embedding methods balance two coupled objectives: multi-view consistency and collapse prevention. In standard methods, both are implemented within the encoder's representation space and rely on batch-level interactions during the forward pass. This tight coupling makes collapse prevention dependent on reliable batch statistics, which deteriorate as batch size shrinks. In contrast, we show that decoupling these objectives enables robust optimization independent of batch size. Conceptually, this shifts collapse prevention from a stochastic batch-level interaction to a deterministic dataset-level geometric constraint.

**IConE.** We propose **ICon**E (**I**nstance-**Con**trasted **E**mbeddings), a framework designed to decouple the multi-view learning process from the anti-collapse mechanism (see Fig. 1) by introducing an auxiliary embedding for each training instance that provides repulsive regularization. These embeddings are produced by a learnable embedding table indexed by instance IDs, with regularization applied directly in this parameter space across the entire dataset rather than only within the current batch. The auxiliary embeddings are used only during training; at inference time the encoder operates as a standard feature extractor without requiring instance identifiers. The main encoder learns consistency under augmentation, but also aligns augmented instance views with these regularized targets via cosine similarity maximization. By applying a diversity penalty to the auxiliary embedding table, we ensure they remain well-separated on the

hypersphere, providing the geometric structure needed to prevent collapse without relying on batch composition. IConE therefore defines a repulsive objective that is invariant to the batch size, allowing the optimization to proceed stably regardless of memory constraints.

**Contributions.**

- We propose **IConE**, a minimal SSL method that explicitly separates multi-view invariance learning and the anti-collapse mechanism needed for joint-embedding architectures. Across diverse 2D and 3D modalities, IConE achieves state-of-the-art linear-probe performance in small-batch (1-64) and low-data regimes.
- Beyond standard downstream image classification, we provide a comprehensive geometric analysis of the learned representations. We show that IConE achieves higher rank and intrinsic dimensionality than baselines, indicating reduced dimensional collapse even when training on single-instance batches.
- We establish the robustness of IConE on highly unbalanced datasets, demonstrating that the instance embedding table provides a global regularization signal that prevents the poor feature learning for minority classes typically observed in batch-dependent methods.
- We extend the core principle of IConE to supervised representation learning. By applying the same parametric anchoring mechanism guided by class labels, we overcome the batch dependency of Supervised Contrastive Learning (SupCon) and demonstrate superior performance over both SupCon and standard Cross-Entropy baselines in constrained regimes.

## 2 Related Work

Modern Joint-Embedding Architectures (JEAs) learn representations by enforcing instance-level consistency across augmented views while preventing trivial outputs [39, 40]. To avoid representational collapse, existing methods utilize either explicit instance discrimination or implicit architectural and statistical constraints. Crucially, both paradigms rely on the training batch to implement the anti-collapse mechanisms. While highly effective at scale, this fundamental batch dependency creates the exact bottleneck that IConE is designed to resolve. We categorize existing approaches based on their mechanism for collapse prevention and their resulting dependency on batch size.

### 2.1 Batch-Dependent Representation Learning

**Explicit Interaction (Contrastive Learning).** Methods like SimCLR [11] and others [21, 30] minimize the InfoNCE loss by contrasting positive pairs against in-batch negative samples. This creates a strict dependency on large batches to adequately approximate the negative data distribution. At extremely small batch sizes, this limited negative pool causes learning to stall [10]. Furthermore, under

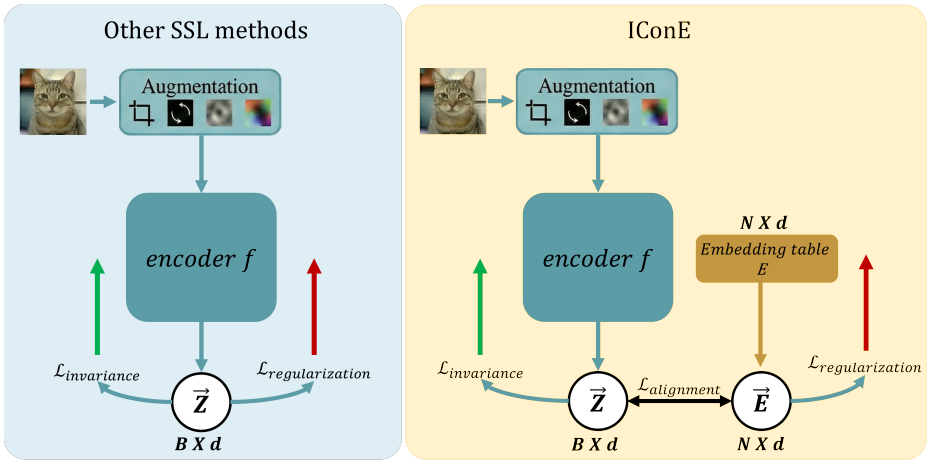


**Fig. 1: Decoupling invariance and anti-collapse gradients with IConE. Left:** Standard JEAs apply both the multi-view invariance objective ($\mathcal{L}_{\text{invariance}}$) and the anti-collapse mechanism ($\mathcal{L}_{\text{regularization}}$) directly within the encoder's dynamic representation space $\boldsymbol{Z}$. This couples the anti-collapse constraint to the current batch. **Right:** IConE transfers the anti-collapse mechanism to an explicitly different parameter space. A persistent embedding table yields instance-specific anchors $\boldsymbol{E}$, which are independently structured via $\mathcal{L}_{\text{regularization}}$. The encoder $f$ is optimized purely through attractive forces: local view consistency ($\mathcal{L}_{\text{invariance}}$) and alignment to its regularized target ($\mathcal{L}_{\text{alignment}}$). This structural decoupling removes the need for batch-dependent negative sampling or variance statistics, applying the repulsive force to the entire dataset at the same time.

significant class imbalance, minority classes are often entirely absent from small batches, depriving the model of the gradients needed to distinguish them and resulting in poor minority class representations [24].

**Implicit Interaction (Non-Contrastive Learning).** To avoid explicit negatives, methods like VICReg [5], Barlow Twins [45], and W-MSE [17] regularize statistical properties of batch embeddings. However, in very small batches, sample variance becomes ill-conditioned, rendering this anti-collapse mechanism unreliable. Asymmetric distillation approaches similarly rely on batch dynamics: DINO [9] centers representations using an exponential moving average of the batch mean, which becomes destabilized by individual samples in small batches. Likewise, while BYOL [19] can theoretically function without Batch Normalization [34], standard implementations exhibit lower performance as batch size decreases. Ultimately, because batch statistics are inherently unreliable in the small-batch limit, these methods require significant modification to function effectively [7,23]. Methods like SwAV [8] learn a set of prototypes and assign samples to them. While SwAV also learns embedding vectors (prototypes), these are shared across instances ($K \ll N$), requiring an assignment mechanism that depends on batch statistics. This step is computed over the batch using the Sinkhorn-Knopp

algorithm; consequently, SwAV requires sufficiently large batches to approximate the dataset distribution and perform meaningful clustering assignments.

### 2.2 Approaches for Reducing Batch Size Dependency

Several approaches aim to reduce the impact of smaller batches on downstream performance. MoCo [12, 13, 21] decouples negative sampling from batch size via a feature queue. However, ensuring consistency among these past encoded features requires a momentum encoder, doubling model parameters. Furthermore, queue dynamics assume older embeddings remain semantically relevant, which often fails during early training when the encoder evolves rapidly. IConE sidesteps these issues entirely: instead of caching potentially stale outputs, we optimize persistent instance embeddings directly via gradient descent, ensuring they remain synchronized with the evolving encoder. Beyond memory banks, several other works have recognized the batch-size sensitivity of SSL and proposed mitigation strategies, including loss decoupling, per-instance moving averages, or multi-view generation [27, 33, 43, 44, 46]. However, these approaches largely retain momentum encoders, memory banks, or in-batch negatives, and thus do not fully eliminate batch dependency. Moreover, they typically define small batch as 64–256 samples (infeasible for high-dimensional volumes) and evaluate on large-scale natural image benchmarks rather than data-scarce scientific regimes. To our knowledge, IConE is the first method to demonstrate robust performance in extremely small batch settings, providing comprehensive evaluation across 2D and 3D modalities for batch sizes from 1 to 64.

**Parametric Instance Discrimination.** Our work revisits Parametric Instance Discrimination (PID) [41], which treats each instance as a unique class. Recent work like DIET [3] has shown that PID can be competitive with modern SSL if tuned correctly. However, DIET's cross-entropy framework introduces two structural limitations: it enforces multi-view alignment only implicitly through shared instance labels, and it separates representations via implicit softmax competition rather than direct geometric constraints. IConE addresses these issues by augmenting instance discrimination with an explicit view-view consistency term, and by replacing softmax competition with a direct geometric regularizer on the auxiliary embedding space.

## 3 Method

Our framework is built on the insight that the geometric requirements of representation learning—uniformity and alignment [39]—can be decoupled into separate optimization processes. Alignment is learned in the encoder representation space, while uniformity is enforced in a persistent auxiliary embedding space. By offloading the uniformity constraint to a set of persistent instance embeddings, we free the encoder from the need to observe a representative set of negative samples within a single batch.

### 3.1 Preliminaries

Let $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be a dataset of $N$ unlabeled samples. We aim to learn an encoder $f_\theta : \mathcal{X} \to \mathbb{R}^d$ that maps high-dimensional inputs to discriminative representations. For a given sample $\mathbf{x}_i$, we generate $V$ augmented views $\{\mathbf{x}_i^{(v)}\}_{v=1}^{V}$ by sampling from a stochastic augmentation family $\mathcal{T}$. The encoder outputs are projected onto the unit hypersphere $\mathbb{S}^{d-1} = \{\mathbf{z} \in \mathbb{R}^d : \|\mathbf{z}\|_2 = 1\}$:

$$\mathbf{z}_i^{(v)} = \frac{f_\theta(\mathbf{x}_i^{(v)})}{\|f_\theta(\mathbf{x}_i^{(v)})\|_2}. \tag{1}$$

Unlike methods that employ a projector network between the backbone and the loss [5, 11], IConE operates directly on backbone representations. This eliminates the ambiguity of choosing between projected and backbone features at evaluation time. In this work, $B$ refers to the batch size of unique instances loaded from the dataset, resulting in $B \times V$ total forward passes per step.

### 3.2 Instance Embeddings as Geometric Anchors

To remove the batch-size dependency, we define the target geometry of the latent space explicitly using an auxiliary embedding set. We maintain a learnable embedding table $\mathbf{E} \in \mathbb{R}^{N \times d}$, where the row $\mathbf{e}_i$ serves as a persistent anchor for instance $i$. These embeddings are initialized from a normal distribution $\mathcal{N}$. During training, we normalize embeddings when computing similarities: $\tilde{\mathbf{e}}_i = \mathbf{e}_i / \|\mathbf{e}_i\|_2$. The unnormalized parameters $\mathbf{e}_i$ are updated directly by gradient descent. This table allows us to separate the optimization into two distinct objectives:

1. **Target Geometry (Uniformity):** We regularize the set of anchors $\{\tilde{\mathbf{e}}_i\}_{i=1}^{N}$ to be uniformly distributed on the hypersphere. Crucially, because $\mathbf{E}$ persists across iterations, this regularization enforces constraints over the entire dataset, utilizing the full $N^2$ Gram matrix rather than just the current batch.
2. **Learned Geometry (Alignment):** We train the encoder $f_\theta$ to maintain multi-view consistency, ensuring that different stochastic augmentations of the same sample $\mathbf{x}_i$ are mapped to nearby representations.

To align the two objectives, we train the encoder $f_\theta$ to map the stochastic augmented views of $\mathbf{x}_i$ to its corresponding stable anchor $\tilde{\mathbf{e}}_i$.

### 3.3 The IConE Objective

The total loss function is composed of three terms, each enforcing a specific geometric constraint.

**View-View Consistency ($\mathcal{L}_{\mathbf{vv}}$).** To ensure the representations capture the invariance defined by the augmentation pipeline, we enforce consistency between different views of the same instance:

$$\mathcal{L}_{\text{vv}} = \frac{1}{B} \sum_{i \in \mathcal{B}} \frac{2}{V(V-1)} \sum_{m<n} \left(1 - \langle \mathbf{z}_i^{(m)}, \mathbf{z}_i^{(n)} \rangle\right). \tag{2}$$

This term pulls views together in the representation space. Without the other terms, minimizing $\mathcal{L}_{\text{vv}}$ would lead to a trivial solution where all outputs map to a single point. Note that even if the batch size is $B = 1$ (a single unique instance), as long as $V \geq 2$ augmented views are generated, this term remains well-defined.

**Instance Diversity ($\mathcal{L}_{\text{div}}$).** To prevent the geometric anchors from collapsing, we apply an explicit repulsive regularizer to the embedding table. We compute the Gram matrix $\mathbf{G} = \tilde{\mathbf{E}}\tilde{\mathbf{E}}^\top$ of the entire embedding table and minimize the off-diagonal correlations using a squared hinge loss:

$$\mathcal{L}_{\text{div}} = \frac{1}{N(N-1)} \sum_{i \neq j} \left[\max(0, G_{ij})\right]^2 . \tag{3}$$

Since usually $N > d$ precludes strict orthogonality, we relax the constraint to penalize only positive correlations. The squared penalty makes gradients proportional to the violation, strongly repelling similar pairs while leaving nearly orthogonal ones stable. Although computing the $N \times N$ matrix is $\mathcal{O}(N^2)$, using lightweight vectors (*e.g.*, $d = 384$) keeps this compute overhead negligible relative to the backbone forward pass for our target dataset sizes. Likewise, the table's $\mathcal{O}(Nd)$ memory footprint (*e.g.*, $\sim$7.7 MB for $N = 5$k) is trivial compared to the model parameters.

However, for massive-scale datasets where computing an $N \times N$ matrix becomes a computational bottleneck, IConE's modularity provides a direct solution. Our explicit instance-to-instance repulsion can seamlessly be substituted with standard statistical anti-collapse mechanisms, such as Variance-Covariance [5] or SIGReg [4], applied directly to the persistent parameter space. Because these methods regularize the $d$-dimensional feature space rather than the $N \times N$ instance similarities, their computational complexity scales linearly with the dataset size. Thus, IConE offers a scalable, complementary architectural paradigm rather than a mutually exclusive competitor to existing regularization strategies.

**View-Instance Alignment ($\mathcal{L}_{\text{vi}}$).** We anchor the encoder outputs to their corresponding auxiliary instance embeddings by minimizing their cosine distance:

$$\mathcal{L}_{\text{vi}} = \frac{1}{B \cdot V} \sum_{i \in \mathcal{B}} \sum_{v=1}^{V} \left(1 - \langle \mathbf{z}_i^{(v)}, \tilde{\mathbf{e}}_i \rangle\right). \tag{4}$$

Since the target $\tilde{\mathbf{e}}_i$ is decoupled from the batch, this term exerts a purely attractive force on the encoder. It effectively asks the network to predict the instance's latent coordinate, which is positioned by global geometric constraints.

**Total Loss.** The complete objective is a sum:

$$\mathcal{L} = \mathcal{L}_{\text{vi}} + \mathcal{L}_{\text{vv}} + \mathcal{L}_{\text{div}}. \tag{5}$$

where each term plays a distinct, complementary role, which we isolate empirically in Appendix F. Notably, IConE does not require temperature scaling. Contrastive

**Algorithm 1** IConE Training

**Require:** Dataset $\mathcal{D}$, encoder $f_\theta$, embedding table $\mathbf{E}$, augmentation family $\mathcal{T}$
Initialize $\mathbf{E} \sim \mathcal{N}(0, 0.02)$
**for** each iteration **do**
  Sample batch $\mathcal{B}$ of $B$ instances
  **for** $i \in \mathcal{B}$ **do**
    Generate $V$ views: $\mathbf{x}_i^{(v)} \sim \mathcal{T}(\mathbf{x}_i)$ for $v = 1, \ldots, V$
    $\mathbf{z}_i^{(v)} \leftarrow \text{Normalize}(f_\theta(\mathbf{x}_i^{(v)}))$ for $v = 1, \ldots, V$
  **end for**
  $\tilde{\mathbf{E}} \leftarrow \text{RowNormalize}(\mathbf{E})$
  Compute $\mathcal{L}_{\text{vi}}$ using $\{\mathbf{z}_i^{(v)}\}$ and $\{\tilde{\mathbf{e}}_i\}_{i \in \mathcal{B}}$
  Compute $\mathcal{L}_{\text{vv}}$ using $\{\mathbf{z}_i^{(v)}\}$
  Compute $\mathcal{L}_{\text{div}}$ over full $\tilde{\mathbf{E}}$ $\{O(N^2)\}$
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{vi}} + \mathcal{L}_{\text{vv}} + \mathcal{L}_{\text{div}}$
  Update $\theta$ with learning rate $\eta$
  Update $\mathbf{E}$ with learning rate $\eta$
**end for**

SSL use temperature to control the sharpness of the softmax distribution over negatives; IConE replaces this implicit competition with direct geometric losses and eliminates this sensitive hyperparameter. The gradient structure reveals why IConE's anti-collapse mechanism is mathematically batch-agnostic. The encoder parameters $\theta$ receive gradients only from the attractive terms ($\mathcal{L}_{\text{vi}}$ and $\mathcal{L}_{\text{vv}}$).

$$\nabla_\theta \mathcal{L} = \nabla_\theta \mathcal{L}_{\text{vi}} + \nabla_\theta \mathcal{L}_{\text{vv}}. \tag{6}$$

The encoder is never explicitly pushed away from other instances in the batch; it is only pulled toward its specific anchor. This is a fundamental departure from contrastive learning, where the gradient for $\mathbf{z}_i$ depends on the summation over negatives $\sum_{j \neq i} \exp(\mathbf{z}_i \cdot \mathbf{z}_j)$.

The repulsive forces are handled entirely by the update to the table $\mathbf{E}$:

$$\nabla_{\mathbf{e}_i} \mathcal{L} = \underbrace{\nabla_{\mathbf{e}_i} \mathcal{L}_{\text{vi}}}_{\text{attraction}} + \underbrace{\nabla_{\mathbf{e}_i} \mathcal{L}_{\text{div}}}_{\text{repulsion}}. \tag{7}$$

Notably, $\mathcal{L}_{\text{vi}}$ bidirectionally updates the encoder and table $\mathbf{E}$: views are mapped toward their current anchor, while the anchor shifts toward the mean of its views. Since $\mathbf{E}$ persists across iterations, it stores the dataset's global repulsive pressure. Even at $B = 1$, $\mathcal{L}_{\text{div}}$ repels $\mathbf{e}_i$ from all other embeddings $\mathbf{e}_{j \neq i}$, guaranteeing a non-collapsed target geometry without relying on in-batch negatives.

## 4 Experiments

We evaluate IConE across diverse imaging modalities, assessing its stability in small-batch and small-dataset regimes, its efficacy on high-dimensional 3D volumes, its robustness under class imbalance, and its generalization to supervised

learning. We further analyze the geometric properties of the learned representations to validate our theoretical motivations. Unless otherwise noted, all results presented in this section are aggregated over all datasets within each benchmark (7 datasets for 2D and 6 datasets for 3D); per-dataset results and additional plots are provided in Appendix B.

### 4.1 Experimental Setup

**Training and Evaluation Protocol.** All SSL methods are pretrained using identical encoder architectures (ViT-Small [16] for 2D, ResNet-18 [22] for 3D), augmentation pipelines, optimizers, learning-rate schedules, and training epochs. We evaluate learned representations via linear probing on frozen features, reporting top-1 balanced accuracy. Results are averaged over 5 random seeds, and all methods share the same dataset splits and subsampled subsets for fair comparison. Full training details and hyperparameters are provided in Appendix A.

**Datasets.** We utilize the MedMNIST+(v2) benchmark collection [42], which provides standardized medical imaging datasets spanning diverse clinical domains, imaging modalities, and classification tasks. Unlike synthetic benchmarks, these datasets are derived from real clinical sources and capture the heterogeneity encountered in practice including varying image quality, and domain-specific visual patterns. This diversity makes them well-suited for evaluating method generalization across biomedical imaging applications. Each subset is sampled once and reused across all methods.

- **2D Benchmarks:** Blood (blood cell microscopy), Breast (ultrasound), OrganA/OrganC/OrganS (abdominal CT, different views), Path (colon pathology), and Pneumonia (chest X-ray). Image resolution of all datasets is $224 \times 224$. To simulate data-scarce clinical scenarios, we create stratified subsets with $N \in \{500, 1\text{k}, 2\text{k}, 5\text{k}\}$ total training samples.
- **3D Benchmarks:** Organ3D (abdominal CT), Nodule3D (chest CT), Adrenal3D (abdominal CT), Fracture3D (knee MRI), Synapse3D (electron microscopy), and Vessel3D (brain MRA). All volumes are $64^3$ voxels. These datasets represent the memory-constrained regime where small batches are often unavoidable.

**Baselines.** We compare against contrastive [11, 21], clustering-based [8], non-contrastive [5, 9, 19], and parametric instance discrimination [3] baselines, using default hyperparameters for our encoders. Because standard methods require in-batch negatives or variance statistics, they inherently fail at $B = 1$; thus, we report their results only where their losses are well-defined. To ensure fair comparison, we train IConE with $V = 2$. Appendix C details further experiments accommodating alternative regularization methods (VCReg, SIGReg) used for IConE's diversity loss, alongside ablations on the number of views and the Gaussian initialization of instance embeddings.

### 4.2 Batch-Size Stability and Performance

We evaluate performance across batch sizes $B \in \{1, 2, 4, 8, 16, 32, 64\}$ to characterize each method's dependence on batch statistics.

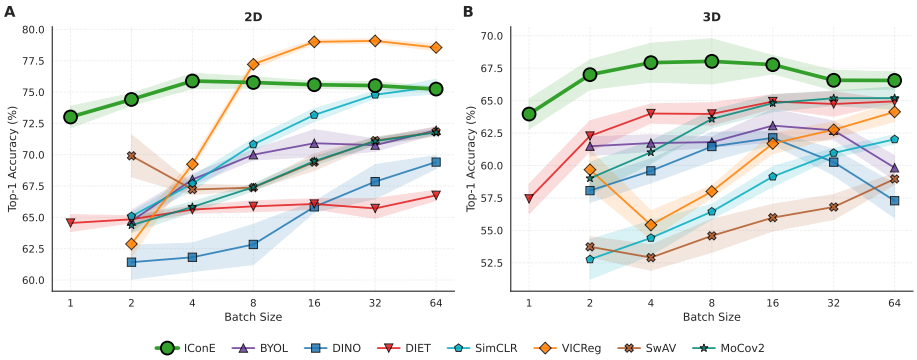


**Fig. 2: Batch-size stability.** Mean and standard deviation linear-probe top-1 balanced accuracy aggregated over **(A)** 2D datasets and **(B)** 3D datasets.

**Table 1: Self-supervised k-NN performance.** Mean k-NN balanced accuracy (%) across 6 3D datasets. Best results in **bold**.

| | $B$=4 | | | $B$=16 | | | $B$=64 | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | kNN-1 | kNN-5 | kNN-20 | kNN-1 | kNN-5 | kNN-20 | kNN-1 | kNN-5 | kNN-20 |
| BYOL | 54.6 | 53.7 | 49.6 | 58.1 | 56.3 | 49.7 | 58.1 | 55.3 | 53.0 |
| DIET | 55.6 | 55.6 | 54.7 | 55.3 | 55.3 | 55.3 | 57.1 | 57.0 | 54.4 |
| DINO | 55.8 | 54.6 | 50.5 | 57.7 | 54.9 | 51.9 | 57.4 | 56.6 | 54.6 |
| MoCo-v2 | 50.8 | 51.1 | 49.7 | 56.7 | 56.2 | 55.4 | 57.2 | 56.3 | 53.3 |
| SimCLR | 49.9 | 48.4 | 47.7 | 52.0 | 51.6 | 51.8 | 53.1 | 54.5 | 53.9 |
| SwAV | 48.0 | 46.5 | 44.2 | 50.8 | 47.8 | 46.6 | 50.6 | 48.9 | 45.8 |
| VICReg | 50.8 | 47.3 | 45.2 | 55.9 | 56.7 | 55.7 | 59.1 | 57.8 | 57.0 |
| IConE (Ours) | **63.6** | **62.2** | **60.6** | **63.3** | **62.8** | **62.4** | **62.5** | **60.6** | **58.8** |

**2D Results.** Fig. 2A presents aggregated top-1 balanced accuracy across 2D datasets. IConE exhibits stability, maintaining a flat curve. Conversely, other methods degrade characteristically as batch size decreases. While VICReg achieves the best accuracy at larger batches, it drops sharply in smaller regimes. Overall, IConE outperforms all other baselines across the spectrum, with SimCLR catching up at $B = 64$, and maintains strictly higher absolute performance than DIET, which shares a parametric approach but yields lower accuracy.

**3D Results.** Fig. 2B demonstrates that IConE's advantages are even more pronounced in 3D, where memory constraints render small batches unavoidable. IConE consistently achieves the highest accuracy across all batch sizes. Interestingly, some baselines exhibit non-monotonic behavior, with performance dropping at the largest batch sizes, possibly showing optimization difficulties because of the use of large batch sizes in small datasets [25]. To complement the linear probe, we evaluate the representations with a non-parametric k-NN classifier in Tab. 1.

**Batch Size Sensitivity Analysis.** Fig. 3 quantifies this dependence via absolute performance drop and correlation between batch size and accuracy. The left panel shows IConE maintaining high performance with minimal degradation, avoiding

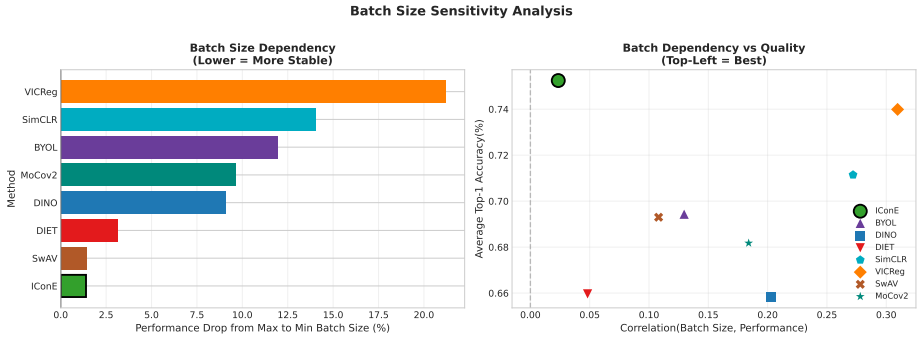


**Fig. 3: Batch sensitivity analysis. Left:** Performance drop from largest to smallest batch size (lower is better). **Right:** Pearson Correlation between batch size and performance versus absolute top-1 balanced accuracy.

the collapse of baselines. The right panel plots the trade-off between batch-size-agnosticism (low correlation, left) and representation quality (high accuracy, top). By uniquely occupying the optimal top-left corner with high accuracy and near-zero correlation, IConE achieves true batch-size-agnostic learning without sacrificing representational power. To confirm these trends extend beyond the biomedical domain, we additionally evaluate in the same small-data, small-batch regime on a natural-image subset of 20 ImageNet [15] classes (250 images each); IConE exhibits the same batch-size stability there (Fig. 14, Appendix E).

### 4.3 Analysis of Learned Representations

We complement the evaluation with geometric analysis of the learned representations. A single linear probe evaluates only one specific task, whereas robust features must support diverse applications. A collapsed representation might therefore satisfy a specific probe yet carry limited information.

**Qualitative Analysis.** One intuitive way to assess representation quality is through direct visualization. Fig. 4 displays UMAP projections of learned representations for OrganMNIST3D across batch sizes and methods. At larger batch sizes, some methods achieve reasonable class separation. However, as batch size decreases, baselines exhibit various forms of representational collapse, with embeddings losing their class structure. In contrast, IConE preserves well-separated, compact clusters across all batch sizes.

**Quantitative Analysis.** We complement the qualitative analysis with established representation quality metrics. Fig. 5 presents four complementary views.

**(A) Alignment vs. Uniformity.** Following [39], we plot alignment loss (an invariance measure) against uniformity loss (an entropy measure). Good representations should minimize both. IConE points cluster in the favorable low-alignment, low-uniformity region and are colored yellow, indicating high accuracy.

**(B) Effective Rank vs. Accuracy.** Effective rank [35] measures how many embedding dimensions are actively utilized. We observe a strong correlation with

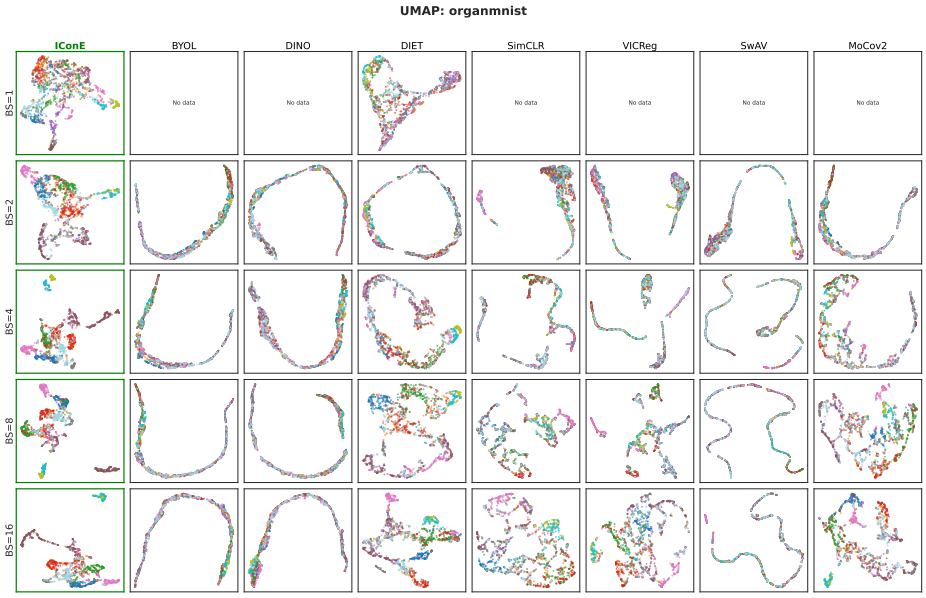


**Fig. 4: UMAP visualization of learned representations.** Embeddings for OrganMNIST3D across batch sizes (rows) and methods (columns). IConE (leftmost column, green border) maintains well-separated class clusters at all batch sizes, while batch-dependent methods show progressive collapse as batch size decreases.

downstream performance: the optimal top-right region represents high utilization and high accuracy. IConE consistently occupies this region across all batch sizes. **(C–D) Collapse Indicators.** While evaluating quality without a priori information about downstream tasks is challenging, task-agnostic metrics like RankMe [18] and LiDAR [37] quantify dimensional collapse via the singular value spectrum. Higher values indicate greater dimensional utilization; on these metrics, IConE scores dramatically higher than all baselines across the entire batch-size range, confirming its robustness against collapse.

### 4.4 Robustness to Class Imbalance

Class imbalance poses a critical challenge for batch-dependent methods: minority classes are frequently absent from individual small batches, depriving the model of the continuous regularization signal needed to learn distinct, discriminative features for them. To test this regime, we construct class-imbalanced variants of the 2D and 3D datasets by designating half of the classes as minority classes and subsampling them by a factor $k$ (the imbalance ratio). We evaluate robustness with increasing imbalance ratios $k$, using $B = 64$ to alleviate the collapse phenomenon. **Results.** Fig. 6A demonstrates that IConE maintains the highest performance across all imbalance ratios. This advantage is particularly pronounced for minority classes (Fig. 6B), where IConE dominates baselines. This robustness stems from

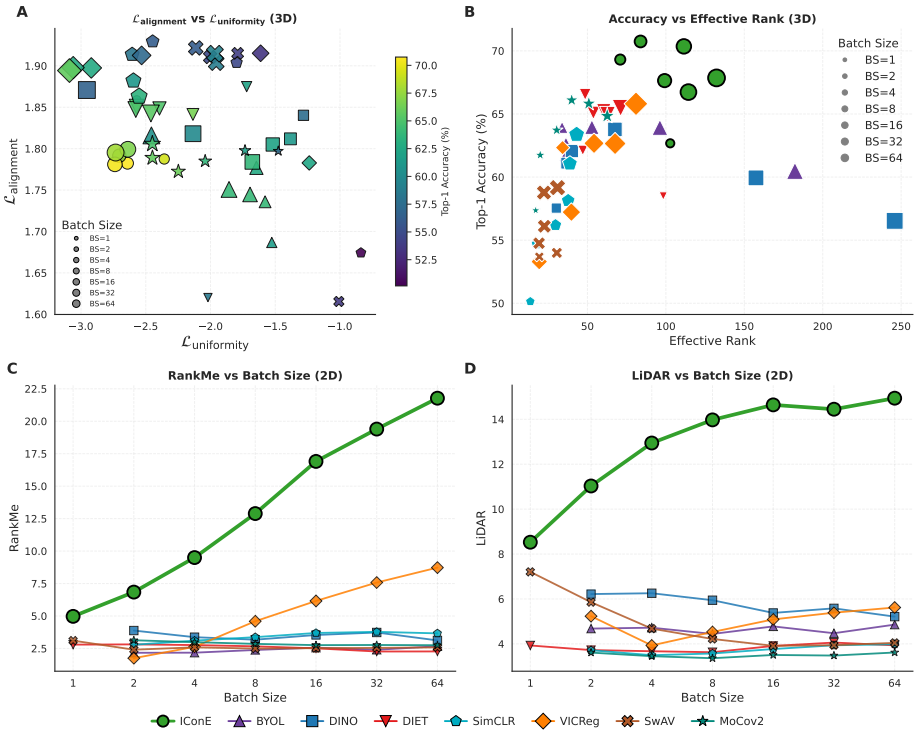


**Fig. 5: Representation geometry analysis. (A)** Alignment vs. uniformity loss for 3D datasets, colored by downstream accuracy. **(B)** Accuracy vs. effective rank (3D). **(C)** RankMe vs. batch size (2D). **(D)** LiDAR vs. batch size (2D). IConE achieves superior geometry metrics across all measures.

the persistent embedding table acting as an implicit memory: even when minority classes are absent from a batch, their corresponding anchors remain active in the diversity loss, preventing the model from forgetting underrepresented concepts.

### 4.5 Extension to Supervised Learning

We investigate whether the IConE mechanism transfers to supervised learning. We compare against standard Cross-Entropy (CE) loss and Supervised Contrastive Learning (SupCon [26]), and introduce two supervised variants of our method:

- **IConE-Class:** Replace the instance-level embedding table with a class prototype table $\mathbf{E} \in \mathbb{R}^{C \times d}$. The encoder aligns views to their class prototype, while prototypes are regularized for orthogonality.
- **IConE-Instance:** Retain instance-level anchors but use labels: the alignment loss pulls views toward same-class anchors, while the diversity loss repels cross-class anchors.

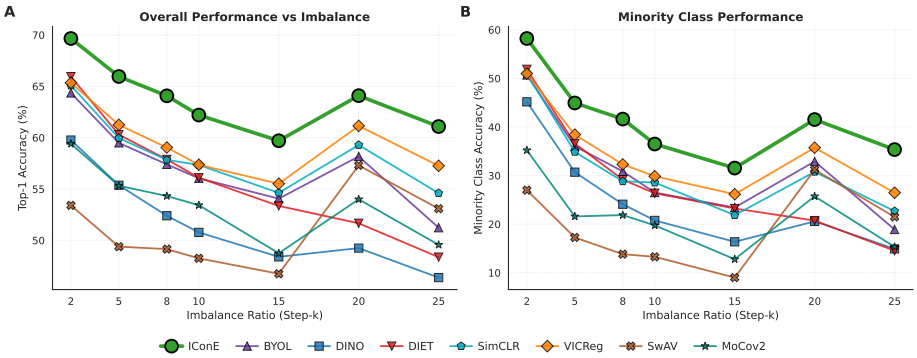


**Fig. 6: Robustness to class imbalance. (A)** Overall top-1 balanced accuracy vs. imbalance ratio. **(B)** Minority class accuracy vs. imbalance ratio. IConE maintains superior performance on both metrics, with particularly strong advantages on minority classes at high imbalance ratios. Results are averaged over all 2D and 3D datasets.

**Table 2: Supervised learning performance.** Aggregated top-1 balanced accuracy (%) across batch sizes for 2D and 3D datasets. Best in **bold**, second best underlined.

| | 2D | | | | | | 3D | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Method** | $B$=1 | $B$=4 | $B$=8 | $B$=16 | $B$=32 | $B$=64 | $B$=1 | $B$=4 | $B$=8 | $B$=16 | $B$=32 | $B$=64 |
| Cross-Entropy | 76.20 | 78.17 | 79.72 | 79.87 | 80.67 | 80.32 | 66.61 | 72.70 | 71.81 | 69.65 | 73.03 | 72.85 |
| SupCon | 72.20 | 74.08 | 76.32 | 79.70 | 79.93 | 79.73 | 59.49 | 66.50 | 72.97 | 72.59 | 67.99 | 71.55 |
| IConE-Instance | 76.81 | 78.43 | 78.56 | 79.53 | 79.58 | 79.63 | 68.77 | **75.16** | **73.14** | 76.56 | **75.10** | 75.31 |
| IConE-Class | **78.71** | **82.42** | **81.97** | **82.16** | **82.28** | **82.31** | **69.59** | 74.25 | 72.13 | **77.14** | 74.34 | **76.21** |

**Results.** Tab. 2 shows both IConE variants outperform SupCon across all batch sizes, as SupCon inherits the same batch dependency as its unsupervised counterpart (SimCLR) due to reliance on in-batch positives. More notably, IConE-Class surpasses even standard CE. We hypothesize this stems from CE's reliance on prolonged training to converge toward a maximally separated class geometry [32], a phenomenon known as Neural Collapse. IConE-Class circumvents this by directly imposing inter-class separation via $\mathcal{L}_{\text{div}}$ and augmentation invariance via $\mathcal{L}_{\text{vv}}$, neither of which has an analogue in standard CE training.

### 4.6 Modularity and Scalability of $\mathcal{L}_{\text{div}}$

The diversity term $\mathcal{L}_{\text{div}}$ is modular: the default $\mathcal{O}(N^2)$ orthogonality penalty can be replaced by any anti-collapse mechanism on the embedding table. Tab. 3 shows that substituting VCReg [5] or SIGReg [4] regularization leaves accuracy within $\sim$1–2 points across batch sizes, confirming the mechanism is agnostic to the form of $\mathcal{L}_{\text{div}}$, while scaling much better in terms of GPU memory (Appendix C).

**Table 3: Regularizer ablation.** Top-1 accuracy (%) for IConE with default orthogonality, VCReg, SIGReg regularization on 2D and 3D datasets across batch sizes.

| | 2D | | | | | | 3D | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Regularizer** | $B$=1 | $B$=2 | $B$=4 | $B$=8 | $B$=16 | $B$=32 | $B$=1 | $B$=2 | $B$=4 | $B$=8 | $B$=16 | $B$=32 |
| IConE (default) | 59.6 | 59.5 | 60.6 | 60.2 | 62.0 | 60.1 | 62.1 | 67.9 | 69.7 | 66.7 | 68.8 | 65.4 |
| IConE-VCReg | 59.3 | 60.2 | 60.8 | 60.9 | 60.4 | 61.2 | 63.0 | 66.1 | 68.8 | 69.2 | 69.1 | 67.9 |
| IConE-SIGReg | 57.3 | 58.4 | 58.3 | 58.5 | 58.5 | 58.4 | 62.9 | 67.5 | 68.4 | 68.9 | 67.5 | 68.0 |

## 5 Conclusion

While self-supervised learning is highly effective, standard methods rely heavily on batch statistics, causing performance degradation at the small batch sizes necessitated by high-dimensional data. To resolve this, we introduced **IConE**, a batch-size-robust approach that replaces dynamic batch interactions with a persistent, learnable embedding table. By aligning augmented views to these stable anchors and applying an explicit orthogonality regularizer, IConE prevents representation collapse independently of batch composition, maintaining consistent performance from batch size 1 to 64. Empirically, IConE achieves state-of-the-art small-batch performance across diverse 2D and 3D tasks, with geometric analysis confirming it successfully prevents dimensional collapse. Furthermore, the framework demonstrates superior robustness to class imbalance, and its core mechanism generalizes effectively to supervised representation learning.

**Limitations and Future Work.** While standard methods may outperform IConE when massive batches are computationally feasible, high-dimensional scientific data fundamentally restricts batch size, making our small-batch focus practically necessary. Moreover, eliminating the need for large batches frees significant memory, enabling the training of much larger encoder architectures. Regarding dataset scalability, while our full-table orthogonality loss scales as $\mathcal{O}(N^2)$, this overhead is negligible in our target domains. For massive-scale future applications, IConE's modularity accommodates $\mathcal{O}(N \cdot d^2)$ statistical regularizers without compromising its core decoupling principle. Ultimately, by removing large-batch dependencies, IConE provides a robust, scalable path forward for high-dimensional self-supervised learning.

## Acknowledgements

RT-SuperES has received funding from the European Innovation Council (EIC) under grant agreement No. 101099654.

## References

1. Assran, M., Balestriero, R., Duval, Q., Bordes, F., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., Ballas, N.: The hidden uniform cluster prior in self-supervised learning. arXiv preprint arXiv:2210.07277 (2022)

2. Ayzenberg, L., Giryes, R., Greenspan, H.: Dinov2 based self supervised learning for few shot medical image segmentation. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
3. Balestriero, R.: Unsupervised learning on a diet: Datum index as target free of self-supervision, reconstruction, projector head. arXiv preprint arXiv:2302.10260 (2023)
4. Balestriero, R., LeCun, Y.: Lejepa: Provable and scalable self-supervised learning without the heuristics. arXiv preprint arXiv:2511.08544 (2025)
5. Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906 (2021)
6. Brondolo, F., Beaussant, S.: Dinov2 rocks geological image analysis: Classification, segmentation, and interpretability. Journal of Rock Mechanics and Geotechnical Engineering (2025)
7. Buglakova, E., Archit, A., D'Imprima, E., Mahamid, J., Pape, C., Kreshuk, A.: Tiling artifacts and trade-offs of feature normalization in the segmentation of large biological images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13109–13118 (2025)
8. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in neural information processing systems **33**, 9912–9924 (2020)
9. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
10. Chen, C., Zhang, J., Xu, Y., Chen, L., Duan, J., Chen, Y., Tran, S., Zeng, B., Chilimbi, T.: Why do we need large batchsizes in contrastive learning? a gradient-bias perspective. Advances in Neural Information Processing Systems **35**, 33860–33875 (2022)
11. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
12. Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)
13. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9640–9649 (2021)
14. Chen, X., Zhang, C., Fu, C., Yang, Z., Zhou, K., Zhang, Y., He, J., Zhang, Y., Sun, M., Wang, Z., et al.: Driving with dino: Vision foundation features as a unified bridge for sim-to-real generation in autonomous driving. arXiv preprint arXiv:2602.06159 (2026)
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
17. Ermolov, A., Siarohin, A., Sangineto, E., Sebe, N.: Whitening for self-supervised representation learning. In: International conference on machine learning. pp. 3015–3024. PMLR (2021)

18. Garrido, Q., Balestriero, R., Najman, L., Lecun, Y.: Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In: International conference on machine learning. pp. 10929–10974. PMLR (2023)
19. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
20. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
21. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
23. Ioffe, S.: Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. Advances in neural information processing systems **30** (2017)
24. Jiang, Z., Chen, T., Mortazavi, B.J., Wang, Z.: Self-damaging contrastive learning. In: International conference on machine learning. pp. 4927–4939. PMLR (2021)
25. Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M., Tang, P.T.P.: On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836 (2016)
26. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
27. Kuang, Y., Guan, J., Liu, H., Chen, F., Wang, Z., Wang, W.: Piccl: A lightweight multiview contrastive learning framework for image classification. PloS one **20**(8), e0329273 (2025)
28. Moutakanni, T., Couprie, C., Yi, S., Doron, M., Chen, Z.S., Moshkov, N., Gardes, E., Caron, M., Touvron, H., Joulin, A., et al.: Cell-dino: Self-supervised image-based embeddings for cell fluorescent microscopy. PLOS Computational Biology **21**(12), e1013828 (2025)
29. Neuschmied, H., Winter, M., Bailer, W.: Improving few-shot object detection using visual explanations of dinov2 features. In: International Conference on Multimedia Modeling. pp. 59–71. Springer (2026)
30. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
31. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
32. Papyan, V., Han, X., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences **117**(40), 24652–24663 (2020)
33. Pototzky, D., Sultan, A., Schmidt-Thieme, L.: Fastsiam: Resource-efficient self-supervised learning on a single gpu. In: DAGM German Conference on Pattern Recognition. pp. 53–67. Springer (2022)
34. Richemond, P.H., Grill, J.B., Altché, F., Tallec, C., Strub, F., Brock, A., Smith, S., De, S., Pascanu, R., Piot, B., et al.: Byol works even without batch statistics. arXiv preprint arXiv:2010.10241 (2020)

35. Roy, O., Vetterli, M.: The effective rank: A measure of effective dimensionality. In: 2007 15th European signal processing conference. pp. 606–610. IEEE (2007)
36. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
37. Thilak, V., Huang, C., Saremi, O., Dinh, L., Goh, H., Nakkiran, P., Susskind, J.M., Littwin, E.: Lidar: Sensing linear probing performance in joint embedding ssl architectures. arXiv preprint arXiv:2312.04000 (2023)
38. Van Assel, H., Ibrahim, M., Biancalani, T., Regev, A., Balestriero, R.: Joint embedding vs reconstruction: Provable benefits of latent space prediction for self supervised learning. arXiv preprint arXiv:2505.12477 (2025)
39. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International conference on machine learning. pp. 9929–9939. PMLR (2020)
40. Wang, Y., Zhang, Q., Wang, Y., Yang, J., Lin, Z.: Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. arXiv preprint arXiv:2203.13457 (2022)
41. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
42. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
43. Yeh, C.H., Hong, C.Y., Hsu, Y.C., Liu, T.L., Chen, Y., LeCun, Y.: Decoupled contrastive learning. In: European conference on computer vision. pp. 668–684. Springer (2022)
44. Yuan, Z., Wu, Y., Qiu, Z.H., Du, X., Zhang, L., Zhou, D., Yang, T.: Provable stochastic optimization for global contrastive learning: Small batch does not harm performance. In: International Conference on Machine Learning. pp. 25760–25782. PMLR (2022)
45. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International conference on machine learning. pp. 12310–12320. PMLR (2021)
46. Zhang, J., Shen, L., Liu, P.: From pretext to purpose: Batch-adaptive self-supervised learning. arXiv preprint arXiv:2311.09974 (2023)