Joint-Embedding vs Reconstruction: Provable Benefits of Latent Space Prediction for Self-Supervised Learning

Hugues Van Assel[*1,2], **Mark Ibrahim**[3], **Tommaso Biancalani**[1], **Aviv Regev**[1], **Randall Balestriero**[2,3]

[1] Genentech, [2] Brown University, [3] Meta AI, FAIR

Abstract

Reconstruction and joint-embedding have emerged as two leading paradigms in Self-Supervised Learning (SSL). Reconstruction methods focus on recovering the original sample from a different view in input space. On the other hand, joint-embedding methods align the representations of different views in latent space. Both approaches offer compelling advantages, yet practitioners lack clear guidelines for choosing between them. In this work, we unveil the core mechanisms that distinguish each paradigm. By leveraging closed-form solutions for both approaches, we precisely characterize how the view generation process, *e.g.* data augmentation, impacts the learned representations. We then demonstrate that, unlike supervised learning, both SSL paradigms require a minimal alignment between augmentations and irrelevant features to achieve asymptotic optimality with increasing sample size. Our findings indicate that in scenarios where these irrelevant features have a large magnitude, joint-embedding methods are preferable because they impose a strictly weaker alignment condition compared to reconstruction-based methods. These results not only clarify the trade-offs between the two paradigms but also substantiate the empirical success of joint-embedding approaches on real-world challenging datasets.

1 Introduction

Training deep neural networks to extract informative data representations is central to AI. Numerous families of methods pursue this goal [52]. In supervised learning, one does so by prescribing *labels* that encode what is considered informative in the data. While this has been the dominant approach to representation learning over the past decades, it has become clear that labels are often overly specialized. Such specialization prevents learning representations that transfer across an ever-increasing diversity of downstream tasks [24, 37]. Self-Supervised Learning (SSL) has emerged as an alternative that moves away from labels [4, 50, 14]. In SSL, one does not assume *a priori* what is informative; instead, one specifies which variations are uninformative and should be disregarded. Identifying, *a priori*, the invariances a representation should satisfy is a broadly applicable principle. For instance, many downstream tasks involving natural images, such as recognition, counting, and segmentation, are inherently robust to minor changes in color or illumination. Consequently, these tasks benefit from representations that exhibit such invariances, typically encoded through a data-augmentation process. Two primary families of methods have emerged to learn representations using this principle: *reconstruction*-based and *joint-embedding* approaches.

[*]Contact: van_assel.hugues@gene.com

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

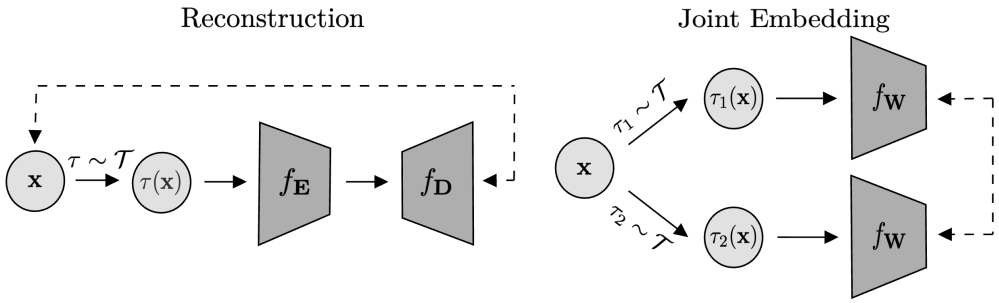

Figure 1: Two self-supervised learning paradigms studied in this work. Left: Reconstruction problem of Equation (SSL-RC): a random augmentation $\tau \sim \mathcal{T}$ is applied to $\mathbf{x}$ to form $\tau(\mathbf{x})$. An encoder $f_{\mathbf{E}}$ together with a decoder $f_{\mathbf{D}}$ is trained to recover $\mathbf{x}$ from $\tau(\mathbf{x})$. Right: Joint embedding problem of Equation (SSL-JE): two independent augmentations $\tau_1, \tau_2 \sim \mathcal{T}$ of the same $\mathbf{x}$ are mapped by $f_{\mathbf{W}}$ to nearby representations, while embeddings of different inputs are pushed apart.

1.1 The Reconstruction-based approach

Reconstruction-based approaches train models by augmenting an input signal, typically by adding noise or masking, and then training the model to restore the original input [41, 60, 31, 67] (left side of Figure 1). This process encourages the model to learn meaningful internal representations of the data's underlying structure and content to enable successful reconstruction. However, because the learning signal arises from minimizing reconstruction error in the input space, the model is naturally steered toward subspaces that explain the majority of the input's variance [65, 8]. Whether such variance-explaining features are also the most semantically discriminative or useful for downstream tasks depends strongly on the data modality.

In language, reconstruction-based learning, as used in large language models, is highly effective because textual tokens represent compact, semantically meaningful units that already abstract away most low-level variability. Although data quality may vary, language in itself is a highly compressed and rich modality where reconstruction can prove highly successful. Predicting a missing token provides a learning signal that operates directly in semantic space: to succeed, the model must infer the contextual and syntactic relationships that determine meaning, rather than replicate surface patterns. Consequently, minimizing reconstruction error encourages the emergence of abstract relational representations, capturing compositionality, long-range dependencies, and discourse coherence that align closely with human notions of meaning and reasoning [19, 30, 62].

In contrast, in visual domains, variance-explaining features often emphasize aspects of the data that are statistically dominant but semantically shallow. Unlike language, visual data are essentially sensorial recordings of the physical world, capturing raw information without inherent semantic compression. As a result, pixel-level reconstruction objectives tend to drive models toward capturing local statistics and textures that account for most of the input's variance, rather than the higher-order structures and object-level relationships that underpin semantic understanding. This local bias can result in representations that are well-suited for low-level perceptual fidelity but suboptimal for recognition, categorization, or other tasks that depend on global context and semantic abstraction [6, 28]. Consequently, purely reconstruction-based approaches in computer vision often struggle to produce features that generalize well across tasks without additional supervision or adaptation. Fine-tuning is thus frequently necessary to bridge the gap between variance-focused representations and those that encode meaningful, task-relevant semantics [31].

1.2 The Joint-embedding approach

Joint-embedding methods, in contrast, operate entirely in latent space (right side of Figure 1). Their objective is to produce similar representations for different augmented views of the

same input while ensuring that representations of distinct samples remain dissimilar. This separation can be enforced explicitly through a contrastive loss [14, 33], or implicitly via architectural mechanisms such as self-distillation, stop-gradient operations, momentum encoders, or predictor heads that stabilize training and prevent representational collapse even without negative samples [13, 26, 66, 38].

Unlike *reconstruction*-based approaches, joint-embedding methods do not predict in the input space and are therefore less biased toward capturing high-variance components of the signal. Empirically, joint-embedding frameworks have shown strong performance across domains where the input signal is high-dimensional and semantically diffuse. Successful applications span histopathology [73], Earth observation [61], and video representation learning [9]. Despite this progress, the mechanisms through which latent consistency objectives outperform reconstruction-based ones remain poorly understood, motivating the analysis presented in this work.

1.3 Contributions

The critical role of the prediction target in SSL, specifically whether to predict in the input space (*reconstruction*) or the latent representation space (*joint-embedding*), has been demonstrated numerous times [1, 6]. However, it remains unclear when to favor one approach over the other. This work clarifies when to prefer each. Our key findings can be summarized as follows.

1. We derive closed-form solutions for both *reconstruction*-based (Theorem 3.1) and *joint-embedding* (Theorem 3.2) SSL linear models. This enables a precise characterization of data augmentation impacts, analogous to well-known results in supervised learning [11].

2. We then leverage these results to show that optimally aligning the augmentations with the irrelevant components of the input signal can effectively eliminate these components and recover optimal performance for both families of methods (Propositions 4.3 and 4.4). However, in contrast to the supervised learning scenario (Proposition 4.2), simply increasing the sample size cannot overcome any misalignment between the augmentation and the noise (Propositions 4.3 and 4.4).

3. By inspecting the alignment requirements for both *reconstruction* and *joint-embedding* methods, we show that in settings with low-magnitude irrelevant noise features, *reconstruction* methods are preferable, as they require fewer tailored augmentations (Corollary 4.5). Conversely, in scenarios with *high-magnitude* irrelevant noise features, *i.e.* , where such features significantly impact the input signal, *joint-embedding* methods should be preferred, as they impose a strictly weaker alignment condition than reconstruction methods (Corollary 4.5).

4. In Section 5, we experimentally validate these findings on both vectorial and image data. We demonstrate that *joint-embedding* methods such as DINO [13] and BYOL [27] are considerably more robust to severe data corruption than *reconstruction*-based methods like MAE [31] (Section 5.2). In Appendix D, we further provide experimental validation for key results from our theoretical analysis. These experiments show that: (i) SSL methods exhibit significantly greater sensitivity to corruptions compared to supervised learning methods (Appendix D.2 and Figure 2); and (ii) SSL performance in noisy data settings is enhanced by aligning augmentations with the underlying noise (Appendix D.3 and Figure 2).

2 Background

Throughout, we consider n samples $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top \in \mathbb{R}^{n \times d}$ and associated labels $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)^\top \in \mathbb{R}^{n \times \ell}$. We consider a data augmentation distribution $\mathcal{T}$ defined as a distribution over transformations $\tau : \mathbb{R}^d \to \mathbb{R}^d$.

Supervised Learning. For the regression task of predicting labels $\mathbf{Y}$ from observations $\mathbf{X}$, the *augmented empirical risk minimization* problem is as follows:

$$\min_{\mathbf{V}} \frac{1}{n} \sum_{i \in [\![n]\!]} \mathbb{E}_{\tau \sim \mathcal{T}} \left[\| \mathbf{y}_i - f_{\mathbf{V}}(\tau(\mathbf{x}_i)) \|_2^2 \right] . \tag{SL}$$

Interestingly, when using a linear model $f_{\mathbf{V}} : \mathbf{x} \mapsto \mathbf{V}\mathbf{x}$ with $\mathbf{V} \in \mathbb{R}^{\ell \times d}$, the effect of data augmentation in Equation (SL) can be explicitly characterized as a Tikhonov regularization problem as shown in Lemma B.1 [7, 46, 11] which proof is provided in Appendix B:

$$\frac{1}{n}\sum_{i\in[\![n]\!]} \mathbb{E}_{\tau\sim\mathcal{T}}\left[\|\mathbf{y}_i - \mathbf{V}\tau(\mathbf{x}_i)\|_2^2\right] = \|\mathbf{V}\|_{\mathbf{\Sigma}}^2 + \frac{1}{n}\sum_{i\in[\![n]\!]} \|\mathbf{y}_i - \mathbf{V}\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]\|_2^2 \tag{1}$$

where $\|\mathbf{V}\|_{\mathbf{\Sigma}}^2 = \mathrm{Tr}(\mathbf{V}\mathbf{\Sigma}\mathbf{V}^\top)$ and

$$\mathbf{\Sigma} \coloneqq \frac{1}{n}\sum_i \mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\tau(\mathbf{x}_i)^\top\right] - \mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]^\top \tag{Cov}$$

denotes the covariance of the augmented samples. Therefore, the effect of data augmentation within supervised learning using linear models is well understood from a theoretical standpoint.

Lack of Foundations in Self-Supervised Learning. Similar results are lacking for SSL, where the explicit effect of data augmentation for linear models has not been rigorously studied. Despite recent efforts to elucidate the underlying principles [64, 56, 39, 36, 29, 22, 71, 63], these methods remain only superficially understood [53]. A robust statistical framework is still lacking to fully comprehend SSL methods and to position them relative to their supervised learning counterparts [3]. Key open questions involve precisely characterizing the role of data augmentation in shaping final representations within both *reconstruction* and *joint-embedding* frameworks. This work aims to lay the foundation for filling this gap.

3 Augmentation-Aware Closed-Form Solutions

In this section, we derive closed-form solutions for the two main families of SSL methods: *reconstruction-based* and *joint-embedding* approaches. To the best of our knowledge, the following results are the first instances of closed-form solutions for SSL that are directly parameterized by the data augmentation structure. This stands in contrast to previous solutions highlighted in [5], which focused on the dependency graph between augmented samples and were unaware of augmentations. These results will then allow us to analyze how augmentations affect the learned representations in both families of SSL methods.

In line with previous theoretical works focused on analytical tractability [12, 5, 47, 54, 57], we study models that are linear in their parameters. Note that these models can produce arbitrary nonlinear predictions via appropriate feature maps [3] and are known to describe regimes of wide neural networks [44].

3.1 Reconstruction-Based Self-Supervised Learning

We first consider *reconstruction*-based SSL models. The problem can be framed as follows, where $\mathcal{T}$ is the data augmentation distribution:

$$\min_{\mathbf{E},\mathbf{D}} \quad \frac{1}{n}\sum_{i\in[\![n]\!]} \mathbb{E}_{\tau\sim\mathcal{T}}\left[\|\mathbf{x}_i - f_{\mathbf{D}}(f_{\mathbf{E}}(\tau(\mathbf{x}_i)))\|_2^2\right] . \tag{SSL-RC}$$

In this formulation, each data sample is augmented and then passes through an encoder $f_{\mathbf{E}}$, followed by a decoder $f_{\mathbf{D}}$. The objective is to minimize the reconstruction error between the original sample and the reconstructed one. This methodology is analogous to the *Denoising Auto-Encoder* [59], *Masked Auto-Encoder* [32] and similar frameworks. Interestingly, the *reconstruction* problem can be solved in closed form when considering linear models for both encoder and decoder. All proofs can be found in Appendix A.

Theorem 3.1. *Let* $\overline{\mathbf{x}}_i \coloneqq \mathbb{E}_{\tau\sim\mathcal{T}}\big[\tau(\mathbf{x}_i)\big]$ *for each* $i \in [\![n]\!]$*, and define* $\overline{\mathbf{X}} \coloneqq (\overline{\mathbf{x}}_1, \dots, \overline{\mathbf{x}}_n)^\top$*. Assume that* $\frac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}} + \mathbf{\Sigma}$ *is positive definite where* $\mathbf{\Sigma}$ *is defined in Equation* (Cov)*. Consider the singular value decomposition:*

$$\tfrac{1}{n}\mathbf{X}^\top\overline{\mathbf{X}}\left(\tfrac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}} + \mathbf{\Sigma}\right)^{-\frac{1}{2}} = \mathbf{R}\mathbf{\Phi}\mathbf{P}^\top \tag{2}$$

where $\mathbf{R} \in \mathbb{R}^{d\times d}$ *and* $\mathbf{P} \in \mathbb{R}^{d\times d}$ *are orthogonal and* $\mathbf{\Phi} \coloneqq \operatorname{diag}(\phi_1, \dots, \phi_d)$ *with* $\phi_1 \geq \cdots \geq \phi_d \geq 0$. *Solutions of Equation* (SSL-RC) *for* $f_{\mathbf{E}} : \mathbf{x} \mapsto \mathbf{E}\mathbf{x}$ *and* $f_{\mathbf{D}} : \mathbf{x} \mapsto \mathbf{D}\mathbf{x}$ *take the form:*

$$\mathbf{E}^\star = \mathbf{T}\mathbf{P}_k^\top \left(\tfrac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}} + \mathbf{\Sigma}\right)^{-\frac{1}{2}} \quad \textit{and} \quad \mathbf{D}^\star = \mathbf{R}_k\mathbf{\Phi}_k\mathbf{T}^{-1}\,, \tag{3}$$

where $\mathbf{T}$ *is any invertible matrix in* $\mathbb{R}^{k\times k}$, $\mathbf{P}_k$ *and* $\mathbf{R}_k$ *are the first* k *columns of* $\mathbf{P}$ *and* $\mathbf{R}$ *respectively, and* $\mathbf{\Phi}_k = \operatorname{diag}(\phi_1, \dots, \phi_k)$.

3.2 Joint-Embedding-Based Self-Supervised Learning

We now consider a *joint-embedding* SSL problem formulated as follows, where $f_{\mathbf{W}}$ is the SSL model and $\mathcal{T}$ is the data augmentation distribution:

$$\begin{aligned} \min_{\mathbf{W}} \quad & \frac{1}{n}\sum_{i\in[\![n]\!]} \mathbb{E}_{\tau_1,\tau_2\sim\mathcal{T}}\left[\|f_{\mathbf{W}}(\tau_1(\mathbf{x}_i)) - f_{\mathbf{W}}(\tau_2(\mathbf{x}_i))\|_2^2\right]\,, \\ \text{subject to} \quad & \frac{1}{n}\sum_{i\in[\![n]\!]} \mathbb{E}_{\tau\sim\mathcal{T}}\left[f_{\mathbf{W}}(\tau(\mathbf{x}_i))f_{\mathbf{W}}(\tau(\mathbf{x}_i))^\top\right] = \mathbf{I}_k\,. \end{aligned} \tag{SSL-JE}$$

In the above Equation (SSL-JE), the objective represents the usual invariance term which ensures consistency between two augmented views of the same sample and is a common component of *joint-embedding* methods. The constraint enforces orthonormality in the learned representations, promoting diversity in the representation space [64] and thus preventing collapse. Most *joint-embedding* models incorporate a similar repulsion term, either explicitly within the loss function, such as in Barlow-Twins [70], SimCLR [14], VICReg [10], and MoCo [33], or implicitly through architectural design choices, as demonstrated by BYOL [26] and DINO [13]. In our case, we rely on the sum of the outer products of the representation vectors. This approach closely resembles VICReg [10], specifically its covariance regularization term. Interestingly, under the unifying formalism presented in [25], most popular *joint-embedding* methods can be framed with this simple repulsive term.

The problem of Equation (SSL-JE) can also be solved in closed form when considering a linear SSL model, as formalized below.

Theorem 3.2. *Let* $\mathbf{S} \coloneqq \frac{1}{n}\sum_i \mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\tau(\mathbf{x}_i)^\top\right]$, $\mathbf{G} \coloneqq \frac{1}{n}\sum_i \mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]^\top$. *Assume that* $\mathbf{S}$ *is positive definite. Consider the eigendecomposition:*

$$\mathbf{S}^{-\frac{1}{2}}\mathbf{G}\mathbf{S}^{-\frac{1}{2}} = \mathbf{Q}\mathbf{\Omega}\mathbf{Q}^\top \tag{4}$$

where $\mathbf{\Omega} = \operatorname{diag}(\omega_1, \dots, \omega_d)$ *with* $\omega_1 \geq \cdots \geq \omega_d$. *Solutions of Equation* (SSL-JE) *for a linear model* $f_{\mathbf{W}} : \mathbf{x} \mapsto \mathbf{W}\mathbf{x}$ *take the form:*

$$\mathbf{W}^\star = \mathbf{U}\mathbf{Q}_k^\top\mathbf{S}^{-\frac{1}{2}}, \tag{5}$$

where $\mathbf{Q}_k = (\mathbf{q}_1, \dots, \mathbf{q}_k)$ *and* $\mathbf{U}$ *is any orthogonal matrix of size* $k \times k$.

4 Augmentation Alignment Requirements in Self-Supervised Learning

In this section, we build on the results of Section 3 to evaluate the ability of both families of SSL models to reach optimal performance. To define such notion of optimality, we model our data as having k *important signal components* and $d-k$ *pure noise components*. These *noise components* represent the variations that SSL methods are typically designed to be invariant to, *e.g.* background noise for image classification tasks. An ideal SSL encoder would retain the k *informative important dimensions* while discarding the $d-k$ *noise components*.

We formalize this scenario in Section 4.1, where a parameter α is introduced to control the alignment between the irrelevant features and the augmentations. In Section 4.2, we demonstrate that supervised learning models can effectively achieve optimal performance either when augmentations are well aligned with the irrelevant noise features at finite sample

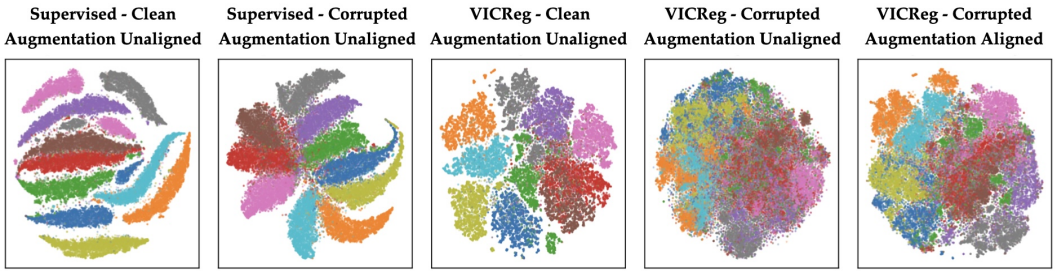

Figure 2: Injecting corruption-aligned noise into data augmentation improves SSL representation quality on corrupted CIFAR-10. Thus **aligning augmentations with the irrelevant components in the data is crucial in SSL**. t-SNE [58] visualizations of (left to right): (1) Supervised features (penultimate layer), clean data. (2) Supervised features (penultimate layer), fog-corrupted (severity 5). (3) VICReg representations, clean data. (4) VICReg representations, fog-corrupted (severity 5). (5) VICReg representations, fog-corrupted (severity 5) with fog noise (severity 1) injection during augmentation. Unlike supervised features, VICReg representations degrade significantly under corruption (compare 3 and 4). Injecting noise in the data augmentation (5) enhances class separability.

sizes or when the sample size is large, regardless of the augmentation employed. In Section 4.3, we show that, unlike supervised learning, SSL necessitates a sufficiently good alignment to achieve optimal performance, even in the infinite sample limit. Finally in Section 4.4, we compare the alignment requirements of *joint-embedding* and *reconstruction*-based SSL methods, thus providing insights into the characteristics of both families of methods.

4.1 Data, Noise and Augmentation

We consider an input dataset with two parts: *important features* and *irrelevant noise*. Optimal performance on downstream tasks is achieved when using only the *important features*. Let $\mathbf{X} = \mathbf{L}\mathbf{K}\mathbf{Q}^\top$ be the singular value decomposition of the *important features*, where $\mathbf{K} = \mathrm{diag}(\kappa_1, \dots, \kappa_d)$ is the diagonal matrix of singular values. Each sample $\mathbf{x}_i$ is corrupted by additive Gaussian noise constituting the *irrelevant* features:

$$\forall\, i \in [\![n]\!], \quad \widetilde{\mathbf{x}}_i = \mathbf{x}_i + \boldsymbol{\gamma}_i, \quad \boldsymbol{\gamma}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}), \tag{6}$$

with $\boldsymbol{\gamma}_i$ drawn independently across $i \in [\![n]\!]$ and where $\boldsymbol{\Gamma} \in \mathbb{R}^{d\times d}$ is positive semi-definite. For simplicity, we assume that $\boldsymbol{\Gamma}$ is diagonalized by the same orthonormal matrix $\mathbf{Q}$ from the SVD above *i.e.* $\boldsymbol{\Gamma} = \mathbf{Q}\boldsymbol{\Lambda}_{\boldsymbol{\Gamma}}\mathbf{Q}^\top$ where $\boldsymbol{\Lambda}_{\boldsymbol{\Gamma}} = \mathrm{diag}(\lambda_1^{\boldsymbol{\Gamma}}, \dots, \lambda_d^{\boldsymbol{\Gamma}})$. The matrix $\widetilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \dots, \tilde{\mathbf{x}}_n)^\top \in \mathbb{R}^{n\times d}$ then forms the *corrupted* input data. We assume that the *important features* are concentrated in exactly $k \geq 1$ components, meaning $\kappa_i > 0$ for $i \in [\![k]\!]$ and $\kappa_i = 0$ for $i > k$. These components are referred to as the *important components*. Additionally, we assume that the *irrelevant noise* are null in these k components, i.e. $\lambda_i^{\boldsymbol{\Gamma}} = 0$ for all $i \in [\![k]\!]$, and strictly positive otherwise, i.e. $\lambda_i^{\boldsymbol{\Gamma}} > 0$ for $i \in [\![k+1:d]\!]$. We refer to the $[\![k+1:d]\!]$ components as the *noise components*.

Data augmentation. Let $\boldsymbol{\Theta} \in \mathbb{R}^{d\times d}$ be positive semi-definite and diagonalized by $\mathbf{Q}$ *i.e.* $\boldsymbol{\Theta} = \mathbf{Q}\boldsymbol{\Lambda}_{\boldsymbol{\Theta}}\mathbf{Q}^\top$ where $\boldsymbol{\Lambda}_{\boldsymbol{\Theta}} = \mathrm{diag}(\lambda_1^{\boldsymbol{\Theta}}, \dots, \lambda_d^{\boldsymbol{\Theta}})$. We consider the augmentation distribution,

$$\forall\, \alpha \geq 0, \;\; \mathcal{T}(\alpha) \coloneqq \left\{ \tau : \mathbb{R}^d \to \mathbb{R}^d \;\middle|\; \tau(\mathbf{x}) = \mathbf{x} + \boldsymbol{\theta} + \alpha\,\boldsymbol{\gamma},\; \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}),\; \boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}) \right\}, \tag{7}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are drawn independently for each transformation. Note that the term $\boldsymbol{\gamma}$ follows the same distribution as the noisy irrelevant features. Increasing the magnitude of α thus aligns the data augmentation with these irrelevant features.

Remark 4.1. One can extend our result to augmentations beyond Gaussians by considering any augmentation of *covariance* (defined in Cov) $\boldsymbol{\Theta} + \alpha^2\boldsymbol{\Gamma}$ [46].

4.2 Supervised Learning Consistency Regardless of Augmentations

We first analyze the behavior of supervised learning models by identifying regimes in which the supervised model effectively disregards the *noisy irrelevant features in* $\widetilde{\mathbf{X}}$. This provides the foundation for pinpointing key differences between supervised and SSL models in Section 4.3. We rely on the data augmentation $\mathcal{T}(\alpha)$, where $\alpha \in \mathbb{R}_+$ controls the alignment between the data corruption and the augmentation process as presented in Section 4.1.

Proposition 4.2. *[Supervised Learning] Let* $\mathbf{V}^\star$ *(resp.* $\widetilde{\mathbf{V}}^\star$*) be the linear model solving Equation* (SL) *with augmentation* $\mathcal{T}(\alpha)$ *for* $\mathbf{X}$ *(resp. the corrupted* $\widetilde{\mathbf{X}}$*). The limit:*

$$\widetilde{\mathbf{V}}^\star \underset{a.s.}{\longrightarrow} \mathbf{V}^\star \tag{8}$$

holds almost surely in either of the following regimes:

- *as* $\alpha \to +\infty$ *(perfect augmentation-noise alignment) for any fixed sample size* $n \in \mathbb{N}$*.*
- *as* $n \to +\infty$ *(infinite samples) for any fixed alignment* $\alpha \geq 0$*.*

The above result shows that, when performing supervised learning with *corrupted* data, the model can achieve the same performance as if it were trained only on the *important features* (thus achieving optimal performance) if either of the following conditions holds: i) The data augmentation process is well aligned with the noise corrupting the inputs (α large). ii) A sufficiently large sample size is available to compensate for any misalignment between the augmentation and the input noise (n large).

4.3 Self-Supervised Learning Requires Aligned Augmentation and Noise

Building on the closed-form expressions for SSL provided in Theorems 3.1 and 3.2, we are now interested in studying the ability of SSL models to achieve optimal performance when trained on the *corrupted* dataset $\widetilde{\mathbf{X}}$, as defined in Section 4.1.

Proposition 4.3. *[Reconstruction] Let* $\mathbf{E}^\star$ *(resp.* $\widetilde{\mathbf{E}}^\star$*) be the linear (encoder) model solving Equation* (SSL-RC) *for* $\mathbf{X}$ *(resp. the corrupted* $\widetilde{\mathbf{X}}$*). The limit:*

$$\widetilde{\mathbf{E}}^\star \underset{a.s.}{\longrightarrow} \mathbf{E}^\star \tag{9}$$

holds[2] *almost surely in either of the following regimes:*

- *as* $\alpha \to +\infty$ *(perfect augmentation-noise alignment) for any fixed sample size* $n \in \mathbb{N}$*.*
- *as* $n \to +\infty$ *(infinite samples), if and only if the alignment* $\alpha \geq 0$ *satisfies:*

$$\alpha^2 \;>\; \alpha_{\mathrm{RC}}^2 \coloneqq \max_{i \in [\![k+1:d]\!]} \frac{\lambda_i^{\mathbf{\Gamma}}}{\eta^2} - \frac{\lambda_i^{\mathbf{\Theta}}}{\lambda_i^{\mathbf{\Gamma}}} - 1 \quad \textit{where} \quad \eta = \min_{i \in [\![k]\!]} \frac{\frac{1}{n}\kappa_i^2}{\sqrt{\frac{1}{n}\kappa_i^2 + \lambda_i^{\mathbf{\Theta}}}}\,. \tag{10}$$

Proposition 4.4. *[Joint-Embedding] Let* $\mathbf{W}^\star$ *(resp.* $\widetilde{\mathbf{W}}^\star$*) be the linear model solving Equation* (SSL-JE) *for* $\mathbf{X}$ *(resp. the corrupted* $\widetilde{\mathbf{X}}$*). The limit:*

$$\widetilde{\mathbf{W}}^\star \underset{a.s.}{\longrightarrow} \mathbf{W}^\star \tag{11}$$

holds[3] *almost surely in either of the following regimes:*

- *as* $\alpha \to +\infty$ *(perfect augmentation-noise alignment) for any fixed sample size* $n \in \mathbb{N}$*.*

[2] Up to an arbitrary invertible matrix (i.e., if $\mathbf{E}^\star$ is a solution, so is $\mathbf{T}\mathbf{E}^\star$ for any $k \times k$ invertible matrix $\mathbf{T}$).

[3] Up to an arbitrary orthogonal rotation (i.e., if $\mathbf{W}^\star$ is a solution, so is $\mathbf{U}\mathbf{W}^\star$ for any $k \times k$ orthogonal matrix $\mathbf{U}$).

- *as $n \to +\infty$ (infinite samples), if and only if the alignment $\alpha \geq 0$ satisfies:*

$$\alpha^2 > \alpha_{\text{JE}}^2 \coloneqq \max_{i \in [\![k+1:d]\!]} \frac{1-\delta}{\delta} - \frac{\lambda_i^{\mathbf{\Theta}}}{\lambda_i^{\mathbf{\Gamma}}} \quad \textit{where} \quad \delta = \min_{i \in [\![k]\!]} \frac{\frac{1}{n}\kappa_i^2}{\frac{1}{n}\kappa_i^2 + \lambda_i^{\mathbf{\Theta}}} \,. \tag{12}$$

The above Propositions 4.3 and 4.4 reveal that, when augmentations are well aligned with the noise *i.e.* when α is large enough, undesired *noise components* are removed from the obtained SSL representation for both families of models, even when combined with other augmentations. However, the alignment requirement persists even as the sample size n becomes arbitrarily large. Therefore, in SSL, achieving optimal performance requires that the data augmentation process be sufficiently well aligned with the *irrelevant noise features* in the data. This marks a key difference from supervised models. Unlike SSL, supervised models can overcome misalignment between augmentations and noise with enough samples, as it learns robustness by observing different noise realizations across identically labeled data. This underscores the critical role of augmentations in SSL. Figure 2 illustrates this by visualizing the embedding spaces of a supervised model and the VICReg [10] SSL model, revealing the latter's susceptibility to noise when augmentations lack proper alignment. This finding is consistent with an empirical study by [51], which concluded that improving augmentations is more impactful than altering architectural designs.

Interestingly, the above results reveal that *reconstruction* (Proposition 4.3) and *joint-embedding* (Proposition 4.4) methods exhibit different alignment requirements to achieve optimal performance. We next analyze these differences.

4.4 Comparison of Joint-Embedding and Reconstruction-Based Methods via Augmentation Alignment Requirement

Leveraging Propositions 4.3 and 4.4, we obtain the following key result.

Corollary 4.5. *Let α_{JE}, δ, α_{RC}, and η be defined as in Proposition 4.4 and Proposition 4.3.*

- *If $\max_{i \in [\![k+1:d]\!]} \lambda_i^{\mathbf{\Gamma}} < \dfrac{\eta^2}{\delta}$ (low noise), then $\alpha_{\text{JE}} > \alpha_{\text{RC}}$ (reconstruction is preferable).*
- *If $\min_{i \in [\![k+1:d]\!]} \lambda_i^{\mathbf{\Gamma}} > \dfrac{\eta^2}{\delta}$ (high noise), then $\alpha_{\text{JE}} < \alpha_{\text{RC}}$ (joint-embedding is preferable).*

This result shows that when the spectral norm of the noise covariance $\mathbf{\Gamma}$ is small, *reconstruction*-based methods impose a less stringent alignment compared to *joint-embedding* methods. Conversely, when the noise magnitude is large, *joint-embedding* methods exhibit lower sensitivity to the augmentation-noise alignment compared to their *reconstruction*-based counterparts. Ultimately, the goal is to minimize the alignment requirement, as the *irrelevant components* are typically unknown in real-world applications.

Reconstruction-based approaches are therefore well-suited for scenarios with weak, *irrelevant noise features*. Intuitively, the core *important components* in such settings possess the greatest magnitude and are consequently prioritized by the model during reconstruction. Due to their reconstruction objective, these methods depend less on data augmentations, which explains their superior performance and robustness over *joint-embedding* techniques in this particular context.

However, when strong *noise features* obscure the *important features* within the raw input signal, *joint-embedding* methods demonstrate greater robustness. This is because *joint-embedding* techniques prioritize latent space prediction, thereby bypassing the need to reconstruct *irrelevant noise components* as model outputs. Consequently, in real-world scenarios where the extent of *irrelevant features* (typically image backgrounds or experimental batch effects in scRNA-seq data) cannot be precisely quantified, *joint-embedding* approaches appear more reliable. This preference is further underscored by the principle that high-amplitude noise naturally exerts a more significant impact on the final representation than low-amplitude noise, making robustness crucial when noise levels are uncertain. This practical advantage also explains the community's preference for *joint-embedding* approaches in challenging datasets [73, 43, 20].

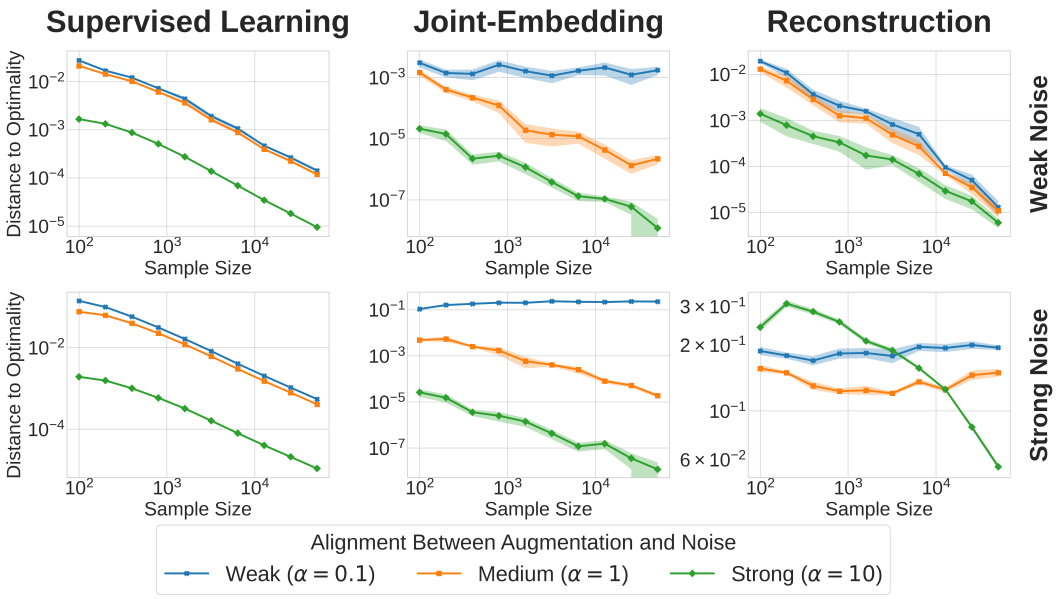

Figure 3: Performance of linear supervised and SSL models (Sections 3.1 and 3.2 and Theorems 3.1 and 3.2) on MNIST corrupted with synthetic Gaussian noise (Section 4.1) with various augmentation alignment α Section 4. Each subplot's y-axis is the absolute difference of supervised linear probing loss (on clean vs. corrupted data) and its x-axis is the sample size n. This figure highlights that *joint-embedding* is preferable to *reconstruction* in the presence of strong *irrelevant noise features*. Conversely, *reconstruction* requires less tailored augmentation when dealing with weak *irrelevant noise features*. Weak noise corresponds to $\lambda^{\mathbf{\Gamma}}_{\max} = 10^3$ and strong noise to $\lambda^{\mathbf{\Gamma}}_{\max} = 10^6$ (details in Appendix C).

Key takeaway. When *irrelevant features* have low magnitude and there is limited prior information on effective augmentations, *reconstruction* is preferable. In contrast, when these *irrelevant features* are non-negligible (as is common with real-world data) or effective augmentations can be identified, *joint-embedding* is preferable.

5 Experiments

This section validates the theoretical findings of Section 4 through experiments on linear models (Section 5.1) and deep networks (Section 5.2). These experiments confirm that the results from the linear model are consistent with those observed in the nonlinear setting of deep networks.

5.1 Experiments With Linear Models

We first validate the theoretical results of Section 4 through experiments with linear models. Data features are corrupted by adding synthetic Gaussian noise, allowing us to precisely control the noise magnitude and its alignment with data augmentations. The details of our experimental design are provided in Appendix C. Our primary results are illustrated in Figure 3. This figure effectively illustrates contrasting behaviors between the various types of methods as sample size and noise magnitude vary. On the left, one can notice that the supervised model achieves optimal performance with either increasing sample size or increasing alignment, with any augmentation and regardless of the *noise magnitude*, confirming the result of Proposition 4.2. In contrast, SSL models exhibit different sensitivities, as predicted in Propositions 4.3 and 4.4. The middle panel shows that *joint-embedding* indeed requires a minimal alignment between augmentation and noise to reach optimal performance (as predicted in Proposition 4.4). Notably, *joint-embedding* maintains robustness even with increasing noise magnitude, as shown in Corollary 4.5. On the right panel, we can see

Table 1: Linear probing top1 accuracy scores of MAE, DINO, and SimCLR on ImageNet with various corruptions [35] and relative performance drop from severity 1 to 5.

	Pixelate Corruption				Gaussian Noise Corruption				Zoomblur Corruption			
Method	Sev. 1	Sev. 3	Sev. 5	Drop (%)	Sev. 1	Sev. 3	Sev. 5	Drop (%)	Sev. 1	Sev. 3	Sev. 5	Drop (%)
BYOL	66.7	61.3	58.7	12.0	67.2	63.1	56.4	16.1	70.1	67.0	63.8	9.0
DINO	68.7	64.9	60.2	12.4	67.6	62.4	59.0	12.7	69.4	67.2	64.9	6.5
MAE	64.9	52.3	46.8	27.9	61.6	46.7	44.8	27.3	64.1	58.4	51.3	20.0

that under weak noise conditions, *reconstruction* is robust to the choice of augmentation. However, as noise becomes stronger, *reconstruction* performance degrades and necessitates a strong alignment between augmentation and noise. These observations confirm the results of Propositions 4.3 and 4.4 and Corollary 4.5. These findings are further supported by experiments on other datasets, including Fashion-MNIST, Kuzushiji-MNIST and the single-cell RNA-seq data from [49], presented in Figures 4 to 7 in Appendix C. All experimental outcomes align with and confirm the theoretical insights detailed in Section 4.

5.2 Experiments with Deep Networks on Images with Various Corruptions

We conduct experiments using both ViT [21] and ResNet [34] architectures. Our evaluation focuses on top-1 linear probing accuracy on ImageNet-100 [18]. ImageNet images inherently contain non-negligible features irrelevant to the classification task [6] (*e.g.* background noise elements), which contributes to the superior performance of *joint-embedding* methods over *reconstruction* methods, as demonstrated in several benchmarks [17, 15]. To further introduce and control the magnitude of *irrelevant noise features*, we utilize the ImageNet-C corruptions [35], which offers corruptions at various severity levels. The results are presented in Table 1. The performance of MAE (ViT) [31], which uses a *reconstruction* objective, is much more affected by the increasing corruptions than DINO (ViT) [13] and BYOL (ResNet) [26], which use a *joint-embedding* objective. Indeed, there is a 25.1% average drop in accuracy for MAE when the severity of the corruption increases from 1 to 5, while DINO and BYOL only experience a 10.5% drop and 12.4% drop, respectively. All experimental details are provided in Appendix D.1.

We perform further ablation studies in the appendix (Appendix D) highlighting this paper's key results. Specifically, our experiments confirm that: (i) SSL is considerably more sensitive to the alignment between augmentations and noise than supervised learning (see Appendix D.2), and (ii) aligning augmentations with the underlying noise can enhance the performance of SSL models in noisy data settings (see Appendix D.3).

6 Conclusion

A growing body of work demonstrates that joint-embedding methods often outperform reconstruction methods on real-world datasets, particularly where extracting useful features for downstream tasks from raw signals is challenging [40, 2]. This is further supported by a consistent empirical finding: reconstruction methods typically necessitate fine-tuning to address the inherent misalignment between the features they learn and those that are perceptually useful for downstream applications [72, 6, 45, 68]. In this work, we have established a theoretical framework to explain this phenomenon. Our analysis provides clear guidelines for practitioners: opt for *reconstruction* when *irrelevant* components show low variance and prior information on effective augmentations is scarce. In contrast, prefer *joint-embedding* when these *irrelevant* components have high variance magnitude, as is common with real-world data, or when effective augmentations are either readily available or can be found through cross-validation.

This paper offers a basis for future work. One could consider extending these results to finite sample size settings to precisely characterize the interplay between sample complexity and augmentation in SSL.

References

[1] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[2] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3478–3488, 2021.

[3] Francis Bach. *Learning theory from first principles*. MIT press, 2024.

[4] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023.

[5] Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 26671–26685. Curran Associates, Inc., 2022.

[6] Randall Balestriero and Yann Lecun. How learning by reconstruction produces uninformative features for perception. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 2566–2585. PMLR, 21–27 Jul 2024.

[7] Randall Balestriero, Ishan Misra, and Yann LeCun. A data-augmentation is worth a thousand samples: Analytical moments and sampling-free training. *Advances in Neural Information Processing Systems*, 35:19631–19644, 2022.

[8] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimization of nonlinear transform codes for perceptual quality. In *2016 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2016.

[9] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023.

[10] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

[11] Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

[12] Vivien Cabannes, Bobak Kiani, Randall Balestriero, Yann LeCun, and Alberto Bietti. The ssl interplay: Augmentations, inductive bias, and generalization. In *International Conference on Machine Learning*, pages 3252–3298. PMLR, 2023.

[13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[15] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.

[16] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

[17] Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *J. Mach. Learn. Res.*, 23:56–1, 2022.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[20] Michael Doron, Théo Moutakanni, Zitong S Chen, Nikita Moshkov, Mathilde Caron, Hugo Touvron, Piotr Bojanowski, Wolfgang M Pernice, and Juan C Caicedo. Unbiased single-cell morphology with self-supervised vision transformers. *bioRxiv*, 2023.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[22] Yann Dubois, Tatsunori Hashimoto, Stefano Ermon, and Percy Liang. Improving self-supervised learning by characterizing idealized representations. *arXiv preprint arXiv:2209.06235*, 2022.

[23] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[24] Yutong Feng, Jianwen Jiang, Mingqian Tang, Rong Jin, and Yue Gao. Rethinking supervised pre-training for better downstream transferring. In *International Conference on Learning Representations*, 2022.

[25] Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann Lecun. On the duality between contrastive and non-contrastive self-supervised learning. *arXiv preprint arXiv:2206.02574*, 2022.

[26] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[27] Thomas Grill and Jan Schlüter. Two convolutional neural networks for bird detection in audio signals. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1764–1768. IEEE, 2017.

[28] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[29] Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

[30] David A Haslett and Zhenguang G Cai. How much semantic information is available in large language model tokens? *Transactions of the Association for Computational Linguistics*, 13:408–423, 2025.

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[32] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[33] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[35] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[36] Weiran Huang, Mingyang Yi, and Xuyang Zhao. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*, 2021.

[37] Ashraful Islam, Chun-Fu Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, and Rogerio Feris. A broad study on the transferability of visual representations with contrastive learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8825–8835. IEEE.

[38] Abhishek Jha, Matthew B Blaschko, Yuki M Asano, and Tinne Tuytelaars. The common stability mechanism behind most self-supervised learning approaches. *arXiv preprint arXiv:2402.14957*, 2024.

[39] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

[40] Bidur Khanal, Binod Bhattarai, Bishesh Khanal, and Cristian Linte. How does self-supervised pretraining improve robustness against noisy labels across various medical image classification datasets? *arXiv preprint arXiv:2401.07990*, 2024.

[41] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[42] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[43] Bipasha Kundu, Bidur Khanal, Richard Simon, and Cristian A Linte. Assessing the performance of the dinov2 self-supervised learning vision transformer model for the segmentation of the left atrium from mri images. *arXiv preprint arXiv:2411.09598*, 2024.

[44] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

[45] Feng Liang, Yangguang Li, and Diana Marculescu. Supmae: Supervised masked autoencoders are efficient vision learners. *arXiv preprint arXiv:2205.14540*, 2022.

[46] Chi-Heng Lin, Chiraag Kaushik, Eva L Dyer, and Vidya Muthukumar. The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective. *Journal of Machine Learning Research*, 25(91):1–85, 2024.

[47] Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How jepa avoids noisy features: The implicit bias of deep linear self distillation networks. *arXiv preprint arXiv:2407.03475*, 2024.

[48] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[49] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[50] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[51] Warren Morningstar, Alex Bijamov, Chris Duvarney, Luke Friedman, Neha Kalibhat, Luyang Liu, Philip Mansfield, Renan Rojas-Gomez, Karan Singhal, Bradley Green, et al. Augmentations vs algorithms: What works in self-supervised learning. *arXiv preprint arXiv:2403.05726*, 2024.

[52] Amirreza Payandeh, Kourosh T. Baghaei, Pooya Fayyazsanavi, Somayeh Bakhtiari Ramezani, Zhiqian Chen, and Shahram Rahimi. Deep representation learning: Fundamentals, technologies, applications, and open challenges. *IEEE Access*, 11:137621–137659, 2023.

[53] Patrik Reizinger, Randall Balestriero, David Klindt, and Wieland Brendel. An empirically grounded identifiability theory will accelerate self-supervised learning research. *arXiv preprint arXiv:2504.13101*, 2025.

[54] James B Simon, Maksis Knutins, Liu Ziyin, Daniel Geisz, Abraham J Fetterman, and Joshua Albrecht. On the stepwise nature of self-supervised learning. In *International Conference on Machine Learning*, pages 31852–31876. PMLR, 2023.

[55] Igor Susmelj, Matthias Heller, Philipp Wirth, Jeremy Prescott, Malte Ebner, and et al. Lightly.

[56] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020.

[57] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR, 2021.

[58] Hugues Van Assel, Nicolas Courty, Rémi Flamary, Aurélien Garivier, Mathurin Massias, Titouan Vayer, and Cédric Vincent-Cuaz. TorchDR: Pytorch Dimensionality Reduction.

[59] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[60] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[61] Leonard Waldmann, Ando Shah, Yi Wang, Nils Lehmann, Adam J Stewart, Zhitong Xiong, Xiao Xiang Zhu, Stefan Bauer, and John Chuang. Panopticon: Advancing any-sensor foundation models for earth observation. *arXiv preprint arXiv:2503.10845*, 2025.

[62] Chenxi Wang, Tianle Gu, Zhongyu Wei, Lang Gao, Zirui Song, and Xiuying Chen. Word form matters: Llms' semantic reconstruction under typoglycemia. *arXiv preprint arXiv:2503.01714*, 2025.

[63] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021.

[64] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[65] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[66] Zixin Wen and Yuanzhi Li. The mechanism of prediction head in non-contrastive self-supervised learning. *Advances in Neural Information Processing Systems*, 35:24794–24809, 2022.

[67] Johann Wenckstern, Eeshaan Jain, Kiril Vasilev, Matteo Pariset, Andreas Wicki, Gabriele Gut, and Charlotte Bunne. Ai-powered virtual tissues from spatial proteomics for clinical diagnostics and biomedical discovery. *arXiv preprint arXiv:2501.06039*, 2025.

[68] Sang Michael Xie, Tengyu Ma, and Percy Liang. Composed fine-tuning: Freezing pre-trained denoising autoencoders for improved generalization. In *International Conference on Machine Learning*, pages 11424–11435. PMLR, 2021.

[69] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[70] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

[71] Wenzheng Zhang and Karl Stratos. Understanding hard negatives in noise contrastive estimation. *arXiv preprint arXiv:2104.06245*, 2021.

[72] Hongling Zheng, Li Shen, Anke Tang, Yong Luo, Han Hu, Bo Du, Yonggang Wen, and Dacheng Tao. Learning from models beyond fine-tuning. *Nature Machine Intelligence*, pages 1–12, 2025.

[73] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide a clear overview of the paper's contributions in the introduction and abstract.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Addressed in the conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Proofs are provided in the appendix Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Details about experiments are provided in the appendix Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to

the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Provided in the supplemental material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We refer to the appendix Appendix D for details on the experimental setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Error bars are provided when relevant.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Discussed in the appendix Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and our work conforms to it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

Justification: This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Outline of the appendix:

- Appendix A: provide the proofs for the theoretical results.
  - Appendix A.1 provides the proof of Theorem 3.1 on the closed-form solution for the reconstruction-based SSL problem.
  - Appendix A.2 provides the proof of Theorem 3.2 on the closed-form solution for the joint-embedding SSL problem.
  - Appendix A.3 provides the proof of Proposition 4.2 on the asymptotic behavior of supervised learning.
  - Appendix A.4 provides the proof of Proposition 4.3 on the asymptotic behavior of reconstruction-based SSL.
  - Appendix A.5 provides the proof of Proposition 4.4 on the asymptotic behavior of joint-embedding SSL.
  - Appendix A.6 provides the proof of Corollary 4.5 on the comparison between joint-embedding and reconstruction-based SSL.
- Appendix B provides the proof of Lemma B.1 on the equivalence between regression with augmented samples and ridge regression.
- Appendix C provides details on the experiments using linear models.
- Appendix D provides details on the experiments using deep networks.

A Proofs

A.1 Proof of Theorem 3.1

Theorem 3.1. *Let* $\overline{\mathbf{x}}_i \coloneqq \mathbb{E}_{\tau\sim\mathcal{T}}\big[\tau(\mathbf{x}_i)\big]$ *for each* $i \in [\![n]\!]$*, and define* $\overline{\mathbf{X}} \coloneqq (\overline{\mathbf{x}}_1, \dots, \overline{\mathbf{x}}_n)^\top$*. Assume that* $\frac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}} + \mathbf{\Sigma}$ *is positive definite where* $\mathbf{\Sigma}$ *is defined in Equation* (Cov)*. Consider the singular value decomposition:*

$$\tfrac{1}{n}\mathbf{X}^\top\overline{\mathbf{X}}\left(\tfrac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}} + \mathbf{\Sigma}\right)^{-\frac{1}{2}} = \mathbf{R}\mathbf{\Phi}\mathbf{P}^\top \tag{2}$$

where $\mathbf{R} \in \mathbb{R}^{d\times d}$ *and* $\mathbf{P} \in \mathbb{R}^{d\times d}$ *are orthogonal and* $\mathbf{\Phi} \coloneqq \operatorname{diag}(\phi_1, \dots, \phi_d)$ *with* $\phi_1 \geq \cdots \geq \phi_d \geq 0$*. Solutions of Equation* (SSL-RC) *for* $f_{\mathbf{E}} : \mathbf{x} \mapsto \mathbf{E}\mathbf{x}$ *and* $f_{\mathbf{D}} : \mathbf{x} \mapsto \mathbf{D}\mathbf{x}$ *take the form:*

$$\mathbf{E}^\star = \mathbf{T}\mathbf{P}_k^\top\left(\tfrac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}} + \mathbf{\Sigma}\right)^{-\frac{1}{2}} \quad \textit{and} \quad \mathbf{D}^\star = \mathbf{R}_k\mathbf{\Phi}_k\mathbf{T}^{-1}\,, \tag{3}$$

where $\mathbf{T}$ *is any invertible matrix in* $\mathbb{R}^{k\times k}$*,* $\mathbf{P}_k$ *and* $\mathbf{R}_k$ *are the first* k *columns of* $\mathbf{P}$ *and* $\mathbf{R}$ *respectively, and* $\mathbf{\Phi}_k = \operatorname{diag}(\phi_1, \dots, \phi_k)$*.*

Proof. Relying on the result of Lemma B.1, we can rewrite the objective as:

$$\tfrac{1}{n}\sum_{i\in[\![n]\!]}\mathbb{E}_{\tau\sim\mathcal{T}}\left[\|\mathbf{x}_i - \mathbf{D}\mathbf{E}\tau(\mathbf{x}_i)\|_2^2\right] = \tfrac{1}{n}\sum_{i\in[\![n]\!]}\|\mathbf{x}_i - \mathbf{D}\mathbf{E}\mathbb{E}_{\tau\sim\mathcal{T}}[\tau(\mathbf{x}_i)]\|_2^2 + \|\mathbf{D}\mathbf{E}\|_{\mathbf{\Sigma}}^2\,, \tag{13}$$

where $\mathbf{\Sigma} = \frac{1}{n}\sum_i \mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\tau(\mathbf{x}_i)^\top\right] - \mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]^\top$. Define per-sample means $\overline{\mathbf{x}}_i \coloneqq \mathbb{E}_{\tau\sim\mathcal{T}}[\tau(\mathbf{x}_i)]$ and stack them into $\overline{\mathbf{X}} \coloneqq (\overline{\mathbf{x}}_1, \dots, \overline{\mathbf{x}}_n)^\top$. The objective can be rewritten as:

$$\min_{\mathbf{E}\in\mathbb{R}^{k\times d}, \mathbf{D}\in\mathbb{R}^{d\times k}} \quad \tfrac{1}{n}\|\mathbf{X} - \overline{\mathbf{X}}\mathbf{E}^\top\mathbf{D}^\top\|_F^2 + \|\mathbf{D}\mathbf{E}\|_{\mathbf{\Sigma}}^2\,. \tag{14}$$

We consider the equivalent problem on $\mathbf{M} = \mathbf{D}\mathbf{E} \in \mathbb{R}^{d\times d}$:

$$\min_{\mathbf{M}\in\mathbb{R}^{d\times d}} \quad \tfrac{1}{n}\|\mathbf{X} - \overline{\mathbf{X}}\mathbf{M}^\top\|_F^2 + \|\mathbf{M}\|_{\mathbf{\Sigma}}^2 \quad \text{s.t.} \quad \operatorname{rank}(\mathbf{M}) \leq k\,. \tag{15}$$

In the above, the rank constraint captures the fact that $\mathbf{M}$ has the form $\mathbf{M} = \mathbf{D}\mathbf{E}$ with $\mathbf{D} \in \mathbb{R}^{d\times k}$ and $\mathbf{E} \in \mathbb{R}^{k\times d}$. The objective can be developed as

$$\tfrac{1}{n}\operatorname{Tr}(\mathbf{X}^\top\mathbf{X}) + \tfrac{1}{n}\operatorname{Tr}(\mathbf{M}\overline{\mathbf{X}}^\top\overline{\mathbf{X}}\mathbf{M}^\top) - \tfrac{2}{n}\operatorname{Tr}(\mathbf{M}\overline{\mathbf{X}}^\top\mathbf{X}) + \operatorname{Tr}(\mathbf{M}\mathbf{\Sigma}\mathbf{M}^\top)\,. \tag{16}$$

Keeping only the terms that depend on $\mathbf{M}$, we can rewrite the problem as

$$\min_{\mathbf{M}\in\mathbb{R}^{d\times d}} \quad \operatorname{Tr}\left(\mathbf{M}\left(\tfrac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}}+\mathbf{\Sigma}\right)\mathbf{M}^\top\right) - \tfrac{2}{n}\operatorname{Tr}(\mathbf{M}\overline{\mathbf{X}}^\top\mathbf{X}) \quad \text{s.t.} \quad \operatorname{rank}(\mathbf{M})\leq k\,. \tag{17}$$

We now consider the change of variable $\mathbf{M}' = \mathbf{M}\left(\frac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}}+\mathbf{\Sigma}\right)^{\frac{1}{2}}$. The above problem then becomes

$$\min_{\mathbf{M}'\in\mathbb{R}^{d\times d}} \quad \operatorname{Tr}(\mathbf{M}'\mathbf{M}'^\top) - 2\operatorname{Tr}\left(\mathbf{M}'\left(\tfrac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}}+\mathbf{\Sigma}\right)^{-\frac{1}{2}}\overline{\mathbf{X}}^\top\mathbf{X}\right) \quad \text{s.t.} \quad \operatorname{rank}(\mathbf{M}')\leq k\,. \tag{18}$$

This problem is equivalent to

$$\min_{\mathbf{M}'\in\mathbb{R}^{d\times d}} \quad \left\|\mathbf{M}' - \tfrac{1}{n}\mathbf{X}^\top\overline{\mathbf{X}}(\tfrac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}}+\mathbf{\Sigma})^{-\frac{1}{2}}\right\|_F^2 \quad \text{s.t.} \quad \operatorname{rank}(\mathbf{M}')\leq k\,. \tag{19}$$

Therefore the optimal $\mathbf{M}'$ is the Euclidean projection of $\frac{1}{n}\mathbf{X}^\top\overline{\mathbf{X}}\left(\frac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}}+\mathbf{\Sigma}\right)^{-\frac{1}{2}}$ onto the set of matrices of rank at most k. This projection has a well-known closed-form solution [23]. Consider the SVD of $\frac{1}{n}\mathbf{X}^\top\overline{\mathbf{X}}\left(\frac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}}+\mathbf{\Sigma}\right)^{-\frac{1}{2}}$:

$$\tfrac{1}{n}\mathbf{X}^\top\overline{\mathbf{X}}\left(\tfrac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}}+\mathbf{\Sigma}\right)^{-\frac{1}{2}} = \mathbf{R}\mathbf{\Phi}\mathbf{P}^\top \tag{20}$$

where $\mathbf{R}\in\mathbb{R}^{d\times d}$ is an orthogonal matrix, $\mathbf{\Phi}\in\mathbb{R}^{d\times d}$ is a diagonal matrix with singular values $\phi_1\geq\cdots\geq\phi_d\geq 0$ on the diagonal, and $\mathbf{P}\in\mathbb{R}^{d\times d}$ is an orthogonal matrix. The optimal $\mathbf{M}'$ is then given by:

$$\mathbf{M}'^\star = \mathbf{R}_k\mathbf{\Phi}_k\mathbf{P}_k^\top\,, \tag{21}$$

where $\mathbf{R}_k\in\mathbb{R}^{d\times k}$ and $\mathbf{P}_k\in\mathbb{R}^{d\times k}$ contain the first k columns of $\mathbf{R}$ and $\mathbf{P}$, respectively, and $\mathbf{\Phi}_k\in\mathbb{R}^{k\times k}$ is a diagonal matrix with the first k largest singular values $\phi_1,\ldots,\phi_k$ on the diagonal.

Then the optimal $\mathbf{M}$ is recovered as

$$\mathbf{M}^\star = \mathbf{M}'^\star\left(\tfrac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}}+\mathbf{\Sigma}\right)^{-\frac{1}{2}} = \mathbf{R}_k\mathbf{\Phi}_k\mathbf{P}_k^\top\left(\tfrac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}}+\mathbf{\Sigma}\right)^{-\frac{1}{2}}\,. \tag{22}$$

It follows that the optimal decoder $\mathbf{D}$ and encoder $\mathbf{E}$ are given by

$$\mathbf{D}^\star = \mathbf{R}_k\mathbf{\Phi}_k\mathbf{T}^{-1} \tag{23}$$

$$\mathbf{E}^\star = \mathbf{T}\mathbf{P}_k^\top\left(\tfrac{1}{n}\overline{\mathbf{X}}^\top\overline{\mathbf{X}}+\mathbf{\Sigma}\right)^{-\frac{1}{2}}\,, \tag{24}$$

where $\mathbf{T}$ is any invertible matrix in $\mathbb{R}^{k\times k}$. □

A.2 Proof of Theorem 3.2

Theorem 3.2. *Let* $\mathbf{S} \coloneqq \frac{1}{n}\sum_i \mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\tau(\mathbf{x}_i)^\top\right]$, $\mathbf{G} \coloneqq \frac{1}{n}\sum_i \mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]^\top$. *Assume that* $\mathbf{S}$ *is positive definite. Consider the eigendecomposition:*

$$\mathbf{S}^{-\frac{1}{2}}\mathbf{G}\mathbf{S}^{-\frac{1}{2}} = \mathbf{Q}\mathbf{\Omega}\mathbf{Q}^\top \tag{4}$$

where $\mathbf{\Omega} = \operatorname{diag}(\omega_1,\ldots,\omega_d)$ *with* $\omega_1\geq\cdots\geq\omega_d$. *Solutions of Equation* (SSL-JE) *for a linear model* $f_{\mathbf{W}}:\mathbf{x}\mapsto\mathbf{W}\mathbf{x}$ *take the form:*

$$\mathbf{W}^\star = \mathbf{U}\mathbf{Q}_k^\top\mathbf{S}^{-\frac{1}{2}}, \tag{5}$$

where $\mathbf{Q}_k = (\mathbf{q}_1,\ldots,\mathbf{q}_k)$ *and* $\mathbf{U}$ *is any orthogonal matrix of size* $k\times k$.

Proof. First, let us consider the constraint which ensures that the learned representations have orthonormal features. Using that $f_{\mathbf{W}} : \mathbf{x} \mapsto \mathbf{W}\mathbf{x}$ and $\mathbf{S} \coloneqq \frac{1}{n}\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\tau(\mathbf{x}_i)^\top\right]$, we obtain:

$$\tfrac{1}{n}\sum_{i\in[\![n]\!]} \mathbf{W}\,\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\tau(\mathbf{x}_i)^\top\right]\mathbf{W}^\top = \mathbf{W}\mathbf{S}\mathbf{W}^\top = \mathbf{I}_k\,. \tag{25}$$

Then, the invariance term measures the consistency between positive views and is given by:

$$\frac{1}{2n}\sum_{i\in[\![n]\!]} \mathbb{E}_{\tau_1,\tau_2\sim\mathcal{T}}\left[\|\mathbf{W}(\tau_1(\mathbf{x}_i)-\tau_2(\mathbf{x}_i))\|^2\right] = \operatorname{Tr}\left(\mathbf{W}\mathbf{\Sigma}\mathbf{W}^\top\right) \tag{26}$$

where

$$\mathbf{\Sigma} \coloneqq \frac{1}{2n}\sum_{i\in[\![n]\!]} \mathbb{E}_{\tau_1,\tau_2\sim\mathcal{T}}\left[(\tau_1(\mathbf{x}_i)-\tau_2(\mathbf{x}_i))(\tau_1(\mathbf{x}_i)-\tau_2(\mathbf{x}_i))^\top\right] \tag{27}$$

$$= \tfrac{1}{n}\sum_{i\in[\![n]\!]} \mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\tau(\mathbf{x}_i)^\top\right] - \mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]^\top \tag{28}$$

$$= \mathbf{S} - \mathbf{G}\,. \tag{29}$$

Thus we have

$$\operatorname{Tr}\left(\mathbf{W}\mathbf{\Sigma}\mathbf{W}^\top\right) = \operatorname{Tr}\left(\mathbf{W}\mathbf{S}\mathbf{W}^\top\right) - \operatorname{Tr}\left(\mathbf{W}\mathbf{G}\mathbf{W}^\top\right) \tag{30}$$

$$= \operatorname{Tr}\left(\mathbf{I}_k\right) - \operatorname{Tr}\left(\mathbf{W}\mathbf{G}\mathbf{W}^\top\right) \tag{31}$$

$$= k - \operatorname{Tr}\left(\mathbf{W}\mathbf{G}\mathbf{W}^\top\right)\,. \tag{32}$$

Therefore, the SSL problem simplifies to:

$$\max_{\mathbf{W}\in\mathbb{R}^{k\times d}} \quad \operatorname{Tr}\left(\mathbf{W}\mathbf{G}\mathbf{W}^\top\right) \tag{33}$$

$$\text{s.t.} \quad \mathbf{W}\mathbf{S}\mathbf{W}^\top = \mathbf{I}_k\,. \tag{34}$$

To solve this constrained non-convex optimization problem, we rely on the KKT conditions that are necessary conditions for optimality.

First-order condition. We introduce the Lagrangian:

$$\mathcal{L} = \operatorname{Tr}\left(\mathbf{W}\mathbf{G}\mathbf{W}^\top\right) - \operatorname{Tr}\left(\mathbf{\Lambda}\left(\mathbf{W}\mathbf{S}\mathbf{W}^\top - \mathbf{I}_k\right)\right), \tag{35}$$

where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1,\ldots,\lambda_k)$ is the diagonal matrix of Lagrange multipliers. Taking the gradient of the Lagrangian with respect to $\mathbf{W}$ and setting it to zero yields:

$$\mathbf{W}^\star\mathbf{G} = \mathbf{\Lambda}\mathbf{W}^\star\mathbf{S}\,. \tag{36}$$

$\mathbf{S}$ is positive definite thus invertible and we can write the above as the following eigenvalue problem:

$$\mathbf{W}^\star\mathbf{G}\mathbf{S}^{-1} = \mathbf{\Lambda}\mathbf{W}^\star\,. \tag{37}$$

Since $\mathbf{S}$ is positive definite, it admits a unique positive definite square root $\mathbf{S}^{1/2}$, and we can write:

$$\mathbf{G}\mathbf{S}^{-1} = \mathbf{S}^{1/2}(\mathbf{S}^{-1/2}\mathbf{G}\mathbf{S}^{-1/2})\mathbf{S}^{-1/2}\,. \tag{38}$$

The matrix $\mathbf{S}^{-1/2}\mathbf{G}\mathbf{S}^{-1/2}$ is symmetric thus admits an eigendecomposition:

$$\mathbf{S}^{-1/2}\mathbf{G}\mathbf{S}^{-1/2} = \mathbf{Q}\mathbf{\Omega}\mathbf{Q}^\top, \tag{39}$$

where $\mathbf{Q}$ is orthogonal and $\mathbf{\Omega} = \operatorname{diag}(\omega_1,...,\omega_d)$ with $\omega_1 \geq ... \geq \omega_d$. Substituting this into the expression for $\mathbf{G}\mathbf{S}^{-1}$, we obtain:

$$\mathbf{G}\mathbf{S}^{-1} = \mathbf{S}^{1/2}\mathbf{Q}\mathbf{\Omega}\mathbf{Q}^\top\mathbf{S}^{-1/2}. \tag{40}$$

Let $\mathbf{P} = \mathbf{S}^{1/2}\mathbf{Q}$. Then $\mathbf{G}\mathbf{S}^{-1}$ admits the decomposition:

$$\mathbf{G}\mathbf{S}^{-1} = \mathbf{P}\mathbf{\Omega}\mathbf{P}^{-1}. \tag{41}$$

In order to maximize the objective, the optimal choice for $\mathbf{W}^\star$ is to pick the eigenvectors associated with the k largest eigenvalues of $\mathbf{G}\mathbf{S}^{-1}$. Therefore, the rows of $\mathbf{W}^\star$ lie in the span of $\{\mathbf{p}_1,\ldots,\mathbf{p}_k\}$. Precisely, there exists a matrix $\mathbf{U}\in\mathbb{R}^{k\times k}$ such that

$$\mathbf{W}^\star = \mathbf{U}\mathbf{P}_k^\top = \mathbf{U}\mathbf{Q}_k^\top\mathbf{S}^{-\frac{1}{2}} \tag{42}$$

where $\mathbf{P}_k = (\mathbf{p}_1,\ldots,\mathbf{p}_k)$ and $\mathbf{Q}_k = (\mathbf{q}_1,\ldots,\mathbf{q}_k)$.

Primal feasibility. Finally, solutions must satisfy the primal constraint: $\mathbf{W}^\star \mathbf{S} \mathbf{W}^{\star\top} = \mathbf{I}_k$. Using the first-order condition and the fact that $\mathbf{Q}_k^\top \mathbf{Q}_k = \mathbf{I}_k$, we obtain:

$$\mathbf{W}^\star \mathbf{S} \mathbf{W}^{\star\top} = \mathbf{U} \mathbf{Q}_k^\top \mathbf{S}^{-\frac{1}{2}} \mathbf{S} \mathbf{S}^{-\frac{1}{2}} \mathbf{Q}_k \mathbf{U}^\top = \mathbf{U} \mathbf{Q}_k^\top \mathbf{Q}_k \mathbf{U}^\top = \mathbf{U}\mathbf{U}^\top \,. \tag{43}$$

Therefore, to satisfy the condition $\mathbf{W}^\star \mathbf{S} \mathbf{W}^{\star\top} = \mathbf{I}_k$, it follows that $\mathbf{U}$ is an orthogonal matrix.

□

A.3 Proof of Proposition 4.2

Proposition 4.2. *[Supervised Learning] Let $\mathbf{V}^\star$ (resp. $\widetilde{\mathbf{V}}^\star$) be the linear model solving Equation* (SL) *with augmentation $\mathcal{T}(\alpha)$ for $\mathbf{X}$ (resp. the corrupted $\widetilde{\mathbf{X}}$). The limit:*

$$\widetilde{\mathbf{V}}^\star \underset{a.s.}{\longrightarrow} \mathbf{V}^\star \tag{8}$$

holds almost surely in either of the following regimes:

- *as $\alpha \to +\infty$ (perfect augmentation-noise alignment) for any fixed sample size $n \in \mathbb{N}$.*
- *as $n \to +\infty$ (infinite samples) for any fixed alignment $\alpha \geq 0$.*

Proof. Let $\mathbf{Q} = (\mathbf{Q}_1 | \mathbf{Q}_2)$ where $\mathbf{Q}_1 \in \mathbb{R}^{d \times k}$ contains the first k columns of $\mathbf{Q}$ and $\mathbf{Q}_2 \in \mathbb{R}^{d \times (d-k)}$ contains the remaining $d-k$ columns spanning the null directions of $\mathbf{X}$. All columns of $\mathbf{X}$ lie in the column space of $\mathbf{Q}_1$ and have no component along $\mathbf{Q}_2$ *i.e.* $\kappa_i > 0$ for $i \in [\![k]\!]$ and $\kappa_i = 0$ for $i > k$. Formally, $\mathbf{X} = \mathbf{X}_1 \mathbf{Q}_1^\top$ where $\mathbf{X}_1 = \mathbf{X}\mathbf{Q}_1 \in \mathbb{R}^{n \times k}$ has full column rank k and $\mathbf{X}\mathbf{Q}_2 = \mathbf{0}$.

Recall that *noise components* are orthogonal to the column span of $\mathbf{Q}_1$. Indeed, $\mathbf{\Gamma} = \mathbf{Q} \mathbf{\Lambda}_{\mathbf{\Gamma}} \mathbf{Q}^\top$ with $\mathbf{\Lambda}_{\mathbf{\Gamma}} = \mathrm{diag}(\lambda_1^{\mathbf{\Gamma}}, \dots, \lambda_d^{\mathbf{\Gamma}})$ such that $\lambda_i^{\mathbf{\Gamma}} = 0$ for any $i \in [\![k]\!]$. Finally, $\lambda_i^{\mathbf{\Gamma}} > 0$ for all $i \in [\![k+1 : d]\!]$.

Uncorrupted data. We consider the problem regularized by data augmentation given by Lemma B.1:

$$\min_{\mathbf{V} \in \mathbb{R}^{\ell \times d}} \; \tfrac{1}{n} \|\mathbf{Y} - \mathbf{X}\mathbf{V}^\top\|_F^2 + \|\mathbf{V}\|_{\mathbf{\Theta} + \alpha^2 \mathbf{\Gamma}}^2 \,. \tag{44}$$

Differentiating the objective with respect to $\mathbf{V}$ and setting the gradient to zero leads to the first-order optimality condition:

$$\tfrac{1}{n} (\mathbf{X}\mathbf{V}^{\star\top} - \mathbf{Y})^\top \mathbf{X} + \mathbf{V}^\star (\mathbf{\Theta} + \alpha^2 \mathbf{\Gamma}) = \mathbf{0}. \tag{45}$$

Given that the matrix $\mathbf{X}^\top \mathbf{X} + n(\mathbf{\Theta} + \alpha^2 \mathbf{\Gamma})$ is invertible, the closed form solution is given by:

$$\mathbf{V}^\star = \tfrac{1}{n} \mathbf{Y}^\top \mathbf{X} \left(\tfrac{1}{n} \mathbf{X}^\top \mathbf{X} + \mathbf{\Theta} + \alpha^2 \mathbf{\Gamma} \right)^{-1} \,. \tag{46}$$

Since $\mathbf{X} = \mathbf{X}_1 \mathbf{Q}_1^\top$ it gives:

$$\mathbf{V}^\star = \tfrac{1}{n} \mathbf{Y}^\top \mathbf{X}_1 \left(\tfrac{1}{n} \mathbf{X}_1^\top \mathbf{X}_1 + \mathbf{\Lambda}_{\mathbf{\Theta},1} \right)^{-1} \mathbf{Q}_1^\top \tag{47}$$

where $\mathbf{\Lambda}_{\mathbf{\Theta},1} = \mathrm{diag}(\lambda_1^{\mathbf{\Theta}}, \dots, \lambda_k^{\mathbf{\Theta}})$ is the block of $\mathbf{\Lambda}_{\mathbf{\Theta}}$ corresponding to the subspace spanned by $\mathbf{Q}_1$.

Corrupted data. We now consider the corrupted data $\widetilde{\mathbf{X}} = \mathbf{X} + \mathbf{N}$ where $\mathbf{N} = (\mathbf{n}_1, \dots, \mathbf{n}_n)^\top$. The $\{\mathbf{n}_i\}_{i \in [\![n]\!]}$ are independent variables such that for any $i \in [\![n]\!]$, $\mathbf{n}_i = \mathbf{\Gamma}^{\frac{1}{2}} \mathbf{z}_i$ where the $\{\mathbf{z}_i\}_{i \in [\![n]\!]}$ are independent $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ vectors.

Using Lemma B.1 gives the following problem:

$$\min_{\widetilde{\mathbf{V}} \in \mathbb{R}^{\ell \times d}} \; \tfrac{1}{n} \|\mathbf{Y} - \widetilde{\mathbf{X}}\widetilde{\mathbf{V}}^\top\|_F^2 + \|\widetilde{\mathbf{V}}\|_{\mathbf{\Theta} + \alpha^2 \mathbf{\Gamma}}^2 \,. \tag{48}$$

Splitting $\widetilde{\mathbf{V}}$ into two parts: $\widetilde{\mathbf{V}}_1 = \widetilde{\mathbf{V}}\mathbf{Q}_1$ and $\widetilde{\mathbf{V}}_2 = \widetilde{\mathbf{V}}\mathbf{Q}_2$, one has that the optimal $\widetilde{\mathbf{V}}_1^\star$ is identical to the one in the uncorrupted case *i.e.*

$$\widetilde{\mathbf{V}}_1^\star = \tfrac{1}{n}\mathbf{Y}^\top\mathbf{X}_1\left(\tfrac{1}{n}\mathbf{X}_1^\top\mathbf{X}_1 + \mathbf{\Lambda}_{\mathbf{\Theta},1}\right)^{-1} . \tag{49}$$

We define $\mathbf{N}_2 = \mathbf{N}\mathbf{Q}_2$. For $\widetilde{\mathbf{V}}_2$, canceling the gradient of the objective we get:

$$\tfrac{1}{n}\widetilde{\mathbf{V}}_2^\star\mathbf{N}_2^\top\mathbf{N}_2 - \tfrac{1}{n}\mathbf{Y}^\top\mathbf{N}_2 + \widetilde{\mathbf{V}}_2^\star(\mathbf{\Lambda}_{\mathbf{\Theta},2} + \alpha^2\mathbf{\Lambda}_{\mathbf{\Gamma},2}) = \mathbf{0} . \tag{50}$$

where $\mathbf{\Lambda}_{\mathbf{\Theta},2} = \mathrm{diag}(\lambda_{k+1}^{\mathbf{\Theta}}, \dots, \lambda_d^{\mathbf{\Theta}})$ and $\mathbf{\Lambda}_{\mathbf{\Gamma},2} = \mathrm{diag}(\lambda_{k+1}^{\mathbf{\Gamma}}, \dots, \lambda_d^{\mathbf{\Gamma}})$.

a) Asymptotic $\alpha \to +\infty$. In this regime, the above condition is equivalent to:

$$\widetilde{\mathbf{V}}_2^\star\mathbf{\Lambda}_{\mathbf{\Gamma},2} = \mathbf{0} \tag{51}$$

Since $\mathbf{\Lambda}_{\mathbf{\Gamma},2}$ is invertible, the first-order condition implies that $\widetilde{\mathbf{V}}_2^\star = \mathbf{0}$.

b) Asymptotic $n \to +\infty$. We obtain by the strong law of large numbers,

$$\tfrac{1}{n}\mathbf{Y}^\top\mathbf{N}_2 \underset{\text{a.s.}}{\longrightarrow} \mathbf{0} \quad \text{and} \quad \tfrac{1}{n}\mathbf{N}_2^\top\mathbf{N}_2 \underset{\text{a.s.}}{\longrightarrow} \mathbf{\Lambda}_{\mathbf{\Gamma},2} . \tag{52}$$

Hence at the limit we have,

$$\widetilde{\mathbf{V}}_2^\star(\mathbf{\Lambda}_{\mathbf{\Theta},2} + (1+\alpha^2)\mathbf{\Lambda}_{\mathbf{\Gamma},2}) = \mathbf{0} . \tag{53}$$

The matrix $\mathbf{\Lambda}_{\mathbf{\Theta},2} + (1+\alpha^2)\mathbf{\Lambda}_{\mathbf{\Gamma},2}$ is invertible for any $\alpha \geq 0$, thus it follows that $\widetilde{\mathbf{V}}_2^\star = \mathbf{0}$.

Conclusion. Therefore, both in the large-sample limit for a fixed α and in the large α limit for a fixed sample size, the optimal coefficients along the $\mathbf{Q}_2$-subspace vanish, *i.e.* $\widetilde{\mathbf{V}}_2^\star = \mathbf{0}$. Since $\widetilde{\mathbf{V}}_1^\star$ coincides with $\mathbf{V}_1^\star$, we recover the same solution as in the noiseless setting in both asymptotic regimes. □

A.4 Proof of Proposition 4.3

Proposition 4.3. *[Reconstruction] Let $\mathbf{E}^\star$ (resp. $\widetilde{\mathbf{E}}^\star$) be the linear (encoder) model solving Equation* (SSL-RC) *for $\mathbf{X}$ (resp. the corrupted $\widetilde{\mathbf{X}}$). The limit:*

$$\widetilde{\mathbf{E}}^\star \underset{a.s.}{\longrightarrow} \mathbf{E}^\star \tag{9}$$

holds[4] *almost surely in either of the following regimes:*

- *as $\alpha \to +\infty$ (perfect augmentation-noise alignment) for any fixed sample size $n \in \mathbb{N}$.*
- *as $n \to +\infty$ (infinite samples), if and only if the alignment $\alpha \geq 0$ satisfies:*

$$\alpha^2 > \alpha_{\mathrm{RC}}^2 \coloneqq \max_{i \in [\![k+1:d]\!]} \frac{\lambda_i^{\mathbf{\Gamma}}}{\eta^2} - \frac{\lambda_i^{\mathbf{\Theta}}}{\lambda_i^{\mathbf{\Gamma}}} - 1 \quad \textit{where} \quad \eta = \min_{i \in [\![k]\!]} \frac{\frac{1}{n}\kappa_i^2}{\sqrt{\frac{1}{n}\kappa_i^2 + \lambda_i^{\mathbf{\Theta}}}} . \tag{10}$$

Proof. Using Theorem 3.1, the closed-form solution for the encoder of the reconstruction SSL problem of Equation (SSL-RC) applied to $\mathbf{X}$ takes the form $\mathbf{E}^\star = \mathbf{T}\mathbf{P}_k^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \mathbf{\Sigma}\right)^{-\frac{1}{2}}$ where $\mathbf{T}$ is any invertible matrix in $\mathbb{R}^{k\times k}$ and $\mathbf{P}_k$ is the matrix containing the k columns of $\mathbf{P}$ associated with the k largest singular values of the matrix $\frac{1}{n}\mathbf{X}^\top\mathbf{X}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \mathbf{\Sigma}\right)^{-\frac{1}{2}}$. Given the construction of $\mathcal{T}(\alpha)$ (Section 4.1), if follows that $\mathbf{\Sigma} = \mathbf{\Theta} + \alpha^2\mathbf{\Gamma}$.

Let $\mathbf{Q} = (\mathbf{Q}_1|\mathbf{Q}_2)$ where $\mathbf{Q}_1 \in \mathbb{R}^{d\times k}$ contains the k columns of $\mathbf{Q}$ corresponding to the *important components* and $\mathbf{Q}_2 \in \mathbb{R}^{d\times(d-k)}$ contains the remaining $d-k$ columns corresponding to the *noise components*. We denote $\mathbf{\Lambda}_{\mathbf{\Theta},1} = \mathrm{diag}(\lambda_1^{\mathbf{\Theta}}, \dots, \lambda_k^{\mathbf{\Theta}})$, $\mathbf{\Lambda}_{\mathbf{\Theta},2} = \mathrm{diag}(\lambda_{k+1}^{\mathbf{\Theta}}, \dots, \lambda_d^{\mathbf{\Theta}})$ and $\mathbf{\Lambda}_{\mathbf{\Gamma},2} = \mathrm{diag}(\lambda_{k+1}^{\mathbf{\Gamma}}, \dots, \lambda_d^{\mathbf{\Gamma}})$.

[4] Up to an arbitrary invertible matrix (i.e., if $\mathbf{E}^\star$ is a solution, so is $\mathbf{T}\mathbf{E}^\star$ for any $k \times k$ invertible matrix $\mathbf{T}$).

Uncorrupted data. Decomposing $\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \mathbf{\Sigma}$ and $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ in $(\mathbf{Q}_1|\mathbf{Q}_2)$ gives:

$$\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \mathbf{\Sigma} = \mathbf{Q}_1\left(\frac{1}{n}\mathbf{K}_1^2 + \mathbf{\Lambda}_{\mathbf{\Theta},1}\right)\mathbf{Q}_1^\top + \mathbf{Q}_2\left(\mathbf{\Lambda}_{\mathbf{\Theta},2} + \alpha^2\mathbf{\Lambda}_{\mathbf{\Gamma},2}\right)\mathbf{Q}_2^\top \tag{54}$$

$$\frac{1}{n}\mathbf{X}^\top\mathbf{X} = \frac{1}{n}\mathbf{Q}_1\mathbf{K}_1^2\mathbf{Q}_1^\top\,. \tag{55}$$

Then using the same reasoning as in the proof of Proposition 4.4 (Appendix A.5) for the uncorrupted data case, we have that the eigenvalues of $\frac{1}{n}\mathbf{X}^\top\mathbf{X}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X} + \mathbf{\Sigma}\right)^{-\frac{1}{2}}$ on the *noise components* $\mathbf{Q}_2$ are all null and they are strictly positive on the *important components*. Therefore the largest eigenvalues are found on the *important components* $\mathbf{Q}_1$ and we obtain $\mathbf{P}_k = \mathbf{Q}_1$.

The solution of the reconstruction SSL problem is then given by

$$\mathbf{E}^\star = \mathbf{T}\left(\frac{1}{n}\mathbf{K}_1^2 + \mathbf{\Lambda}_{\mathbf{\Theta},1}\right)^{-\frac{1}{2}}\mathbf{Q}_1^\top \tag{56}$$

where $\mathbf{T}$ is any invertible matrix of size $k \times k$.

Corrupted data. Decomposing $\frac{1}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \mathbf{\Sigma}$ and $\frac{1}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}$ in $(\mathbf{Q}_1|\mathbf{Q}_2)$ gives:

$$\frac{1}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \mathbf{\Sigma} = \mathbf{Q}_1\left(\frac{1}{n}\mathbf{K}_1^2 + \mathbf{\Lambda}_{\mathbf{\Theta},1}\right)\mathbf{Q}_1^\top + \mathbf{Q}_2\left(\frac{1}{n}\mathbf{N}_2^\top\mathbf{N}_2 + \mathbf{\Lambda}_{\mathbf{\Theta},2} + \alpha^2\mathbf{\Lambda}_{\mathbf{\Gamma},2}\right)\mathbf{Q}_2^\top \tag{57}$$

$$\frac{1}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} = \frac{1}{n}\mathbf{Q}_1\mathbf{K}_1^2\mathbf{Q}_1^\top + \frac{1}{n}\mathbf{Q}_2\mathbf{N}_2^\top\mathbf{N}_2\mathbf{Q}_2^\top\,. \tag{58}$$

a) Asymptotic $\alpha \to +\infty$. In this asymptotic regime the term $\alpha^2\mathbf{\Lambda}_{\mathbf{\Gamma},2}$ dominates thus

$$\mathbf{Q}_2^\top\frac{1}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\left(\frac{1}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \mathbf{\Sigma}\right)^{-\frac{1}{2}}\mathbf{Q}_2 \;\to\; \mathbf{0}\,. \tag{59}$$

Therefore the singular values of $\frac{1}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\left(\frac{1}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \mathbf{\Sigma}\right)^{-\frac{1}{2}}$ on the *noise components* $\mathbf{Q}_2$ all converge to 0 almost surely. On the *important components* $\mathbf{Q}_1$, the smallest singular value is positive. Hence at the limit $\alpha \to +\infty$, we have $\mathbf{P}_k = \mathbf{Q}_1$.

b) Asymptotic $n \to +\infty$. Again using the strong law of large numbers we obtain

$$\frac{1}{n}\mathbf{N}_2^\top\mathbf{N}_2 \xrightarrow[\text{a.s.}]{} \mathbf{\Lambda}_{\mathbf{\Gamma},2}\,. \tag{60}$$

We denote by

$$\eta = \min_{i\in[\![k]\!]} \frac{\frac{1}{n}\kappa_i^2}{\sqrt{\frac{1}{n}\kappa_i^2 + \lambda_i^{\mathbf{\Theta}}}}\,. \tag{61}$$

In this asymptotic regime, ensuring that all eigenvalues of $\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\left(\frac{1}{n}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \mathbf{\Sigma}\right)^{-\frac{1}{2}}$ on *noise components* $\mathbf{Q}_2$ are smaller than eigenvalues on *important components* $\mathbf{Q}_1$ boils down to

$$\forall i \in [\![k+1:d]\!], \quad \eta \;>\; \frac{\lambda_i^{\mathbf{\Gamma}}}{\sqrt{\lambda_i^{\mathbf{\Theta}} + (1+\alpha^2)\lambda_i^{\mathbf{\Gamma}}}}\,. \tag{62}$$

Rearranging this inequality gives

$$\alpha^2 \;>\; \max_{i\in[\![k+1:d]\!]} \frac{\lambda_i^{\mathbf{\Gamma}}}{\eta^2} - \frac{\lambda_i^{\mathbf{\Theta}}}{\lambda_i^{\mathbf{\Gamma}}} - 1\,. \tag{63}$$

Thus, if this condition is satisfied, we obtain $\mathbf{P}_k = \mathbf{Q}_1$.

Conclusion. In the large α limit for a fixed sample size, and in the large-sample limit under the condition of Equation (10), the SSL reconstruction problem is solved by

$$\widetilde{\mathbf{E}}^\star = \mathbf{T}\left(\frac{1}{n}\mathbf{K}_1^2 + \mathbf{\Lambda}_{\mathbf{\Theta},1}\right)^{-\frac{1}{2}}\mathbf{Q}_1^\top = \mathbf{E}^\star\,, \tag{64}$$

where $\mathbf{T}$ is any invertible matrix of size $k \times k$.

□

A.5 Proof of Proposition 4.4

Proposition 4.4. *[Joint-Embedding] Let* $\mathbf{W}^\star$ *(resp.* $\widetilde{\mathbf{W}}^\star$*) be the linear model solving Equation* (SSL-JE) *for* $\mathbf{X}$ *(resp. the corrupted* $\widetilde{\mathbf{X}}$*). The limit:*

$$\widetilde{\mathbf{W}}^\star \xrightarrow[a.s.]{} \mathbf{W}^\star \tag{11}$$

holds[5] *almost surely in either of the following regimes:*

- *as* $\alpha \to +\infty$ *(perfect augmentation-noise alignment) for any fixed sample size* $n \in \mathbb{N}$*.*
- *as* $n \to +\infty$ *(infinite samples), if and only if the alignment* $\alpha \geq 0$ *satisfies:*

$$\alpha^2 \; > \; \alpha_{\text{JE}}^2 \coloneqq \max_{i \in [\![k+1:d]\!]} \frac{1-\delta}{\delta} - \frac{\lambda_i^{\mathbf{\Theta}}}{\lambda_i^{\mathbf{\Gamma}}} \quad \textit{where} \quad \delta = \min_{i \in [\![k]\!]} \frac{\frac{1}{n}\kappa_i^2}{\frac{1}{n}\kappa_i^2 + \lambda_i^{\mathbf{\Theta}}} \, . \tag{12}$$

Proof. Let $\mathbf{Q} = (\mathbf{Q}_1 | \mathbf{Q}_2)$ where $\mathbf{Q}_1 \in \mathbb{R}^{d \times k}$ contains the k columns of $\mathbf{Q}$ corresponding to the *important components* and $\mathbf{Q}_2 \in \mathbb{R}^{d \times (d-k)}$ contains the remaining $d-k$ columns corresponding to the *noise components*.

Using Theorem 3.2, the closed-form solution to the joint-embedding SSL problem of Equation (SSL-JE) applied to $\mathbf{X}$ takes the form $\mathbf{W}^\star = \mathbf{U}\mathbf{Q}_k^\top \mathbf{S}^{-\frac{1}{2}}$, where $\mathbf{U}$ is any orthogonal matrix of size $k \times k$. Recall that $\mathbf{Q}_k$ contains the k columns of $\mathbf{Q}$ associated with the k eigenvectors with largest eigenvalues of the matrix $\mathbf{S}^{-\frac{1}{2}}\mathbf{G}\mathbf{S}^{-\frac{1}{2}}$. Given the construction of $\mathcal{T}(\alpha)$ (Section 4.1), if follows that $\mathbf{S} = \frac{1}{n}\mathbf{X}^\top\mathbf{X} + \mathbf{\Theta} + \alpha^2\mathbf{\Gamma}$ and $\mathbf{G} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$.

We denote $\mathbf{\Lambda}_{\mathbf{\Theta},1} = \operatorname{diag}(\lambda_1^{\mathbf{\Theta}}, \dots, \lambda_k^{\mathbf{\Theta}})$, $\mathbf{\Lambda}_{\mathbf{\Theta},2} = \operatorname{diag}(\lambda_{k+1}^{\mathbf{\Theta}}, \dots, \lambda_d^{\mathbf{\Theta}})$ and $\mathbf{\Lambda}_{\mathbf{\Gamma},2} = \operatorname{diag}(\lambda_{k+1}^{\mathbf{\Gamma}}, \dots, \lambda_d^{\mathbf{\Gamma}})$.

Uncorrupted data. Decomposing $\mathbf{S}$ and $\mathbf{G}$ in $(\mathbf{Q}_1|\mathbf{Q}_2)$ gives:

$$\mathbf{S} = \mathbf{Q}_1 \left(\tfrac{1}{n}\mathbf{K}_1^2 + \mathbf{\Lambda}_{\mathbf{\Theta},1}\right)\mathbf{Q}_1^\top + \mathbf{Q}_2 \left(\mathbf{\Lambda}_{\mathbf{\Theta},2} + \alpha^2\mathbf{\Lambda}_{\mathbf{\Gamma},2}\right)\mathbf{Q}_2^\top \tag{65}$$

$$\mathbf{G} = \tfrac{1}{n}\mathbf{Q}_1\mathbf{K}_1^2\mathbf{Q}_1^\top \, . \tag{66}$$

It holds

$$\mathbf{Q}_1^\top \mathbf{S}^{-\frac{1}{2}}\mathbf{G}\mathbf{S}^{-\frac{1}{2}}\mathbf{Q}_2 = \mathbf{0}_{d-k} \, . \tag{67}$$

Hence on the *noise components* $\mathbf{Q}_2$, the eigenvalues of $\mathbf{S}^{-\frac{1}{2}}\mathbf{G}\mathbf{S}^{-\frac{1}{2}}$ are all null.

On the *important components* $\mathbf{Q}_1$, the smallest eigenvalue of $\mathbf{S}^{-\frac{1}{2}}\mathbf{G}\mathbf{S}^{-\frac{1}{2}}$ satisfies

$$\max_{i \in [\![k]\!]} \tfrac{1}{n}\kappa_i^2 \left(\tfrac{1}{n}\kappa_i^2 + \lambda_i^{\mathbf{\Theta}}\right)^{-1} \; > \; 0 \, , \tag{68}$$

since $\kappa_i > 0$ for any $i \in [\![k]\!]$. Therefore in the clean data setting, the largest eigenvalues of $\mathbf{S}^{-\frac{1}{2}}\mathbf{G}\mathbf{S}^{-\frac{1}{2}}$ are found on the *important component* thus it follows that $\mathbf{Q}_k = \mathbf{Q}_1$.

The solution of the joint-embedding SSL problem is then given by

$$\mathbf{W}^\star = \mathbf{U}\left(\tfrac{1}{n}\mathbf{K}_1^2 + \mathbf{\Lambda}_{\mathbf{\Theta},1}\right)^{-\frac{1}{2}}\mathbf{Q}_1^\top \tag{69}$$

where $\mathbf{U}$ is any orthogonal matrix of size $k \times k$.

[5] Up to an arbitrary orthogonal rotation (i.e., if $\mathbf{W}^\star$ is a solution, so is $\mathbf{U}\mathbf{W}^\star$ for any $k \times k$ orthogonal matrix $\mathbf{U}$).

Corrupted data. Let us denote $\mathbf{N}_2 = \mathbf{N}\mathbf{Q}_2$ with $\mathbf{N}$ being a $n \times d$ matrix whose rows are $\mathbf{n}_i \in \mathbb{R}^d$, where each $\mathbf{n}_i$ is drawn independently from $\mathcal{N}(\mathbf{0}, \mathbf{\Gamma})$.

We define

$$\widetilde{\mathbf{S}} \coloneqq \frac{1}{n} \sum_{i \in [\![n]\!]} \mathbb{E}_{\tau \sim \mathcal{T}} \left[\tau(\tilde{\mathbf{x}}_i) \tau(\tilde{\mathbf{x}}_i)^\top \right] , \tag{70}$$

$$\widetilde{\mathbf{G}} \coloneqq \frac{1}{n} \sum_{i \in [\![n]\!]} \mathbb{E}_{\tau \sim \mathcal{T}} \left[\tau(\tilde{\mathbf{x}}_i) \right] \mathbb{E}_{\tau \sim \mathcal{T}} \left[\tau(\tilde{\mathbf{x}}_i) \right]^\top \tag{71}$$

which is the equivalent of matrix $\mathbf{S}$ when replacing the clean dataset $\mathbf{X}$ by the noisy dataset $\widetilde{\mathbf{X}}$. Decomposing $\widetilde{\mathbf{S}}$ and $\widetilde{\mathbf{G}}$ in $(\mathbf{Q}_1 | \mathbf{Q}_2)$ gives:

$$\widetilde{\mathbf{S}} = \mathbf{Q}_1 \left(\tfrac{1}{n} \mathbf{K}_1^2 + \mathbf{\Lambda}_{\mathbf{\Theta},1} \right) \mathbf{Q}_1^\top + \mathbf{Q}_2 \left(\tfrac{1}{n} \mathbf{N}_2^\top \mathbf{N}_2 + \mathbf{\Lambda}_{\mathbf{\Theta},2} + \alpha^2 \mathbf{\Lambda}_{\mathbf{\Gamma},2} \right) \mathbf{Q}_2^\top \tag{72}$$

$$\widetilde{\mathbf{G}} = \tfrac{1}{n} \mathbf{Q}_1 \mathbf{K}_1^2 \mathbf{Q}_1^\top + \tfrac{1}{n} \mathbf{Q}_2 \mathbf{N}_2^\top \mathbf{N}_2 \mathbf{Q}_2^\top . \tag{73}$$

a) Asymptotic $\alpha \to +\infty$. In this asymptotic regime the term $\alpha^2 \mathbf{\Lambda}_{\mathbf{\Gamma},2}$ dominates thus

$$\mathbf{Q}_2^\top \widetilde{\mathbf{S}}^{-\frac{1}{2}} \widetilde{\mathbf{G}} \widetilde{\mathbf{S}}^{-\frac{1}{2}} \mathbf{Q}_2 \ \to \ \mathbf{0} . \tag{74}$$

Therefore the eigenvalues of $\widetilde{\mathbf{S}}^{-\frac{1}{2}} \widetilde{\mathbf{G}} \widetilde{\mathbf{S}}^{-\frac{1}{2}}$ on the *noise components* $\mathbf{Q}_2$ all converge to 0 almost surely. On the *important components* $\mathbf{Q}_1$, using the same argument as in the above uncorrupted data case, the smallest eigenvalue is strictly greater than 0. Therefore at the limit $\alpha \to +\infty$, the largest eigenvalues are found on the *important components* and we obtain $\mathbf{Q}_k = \mathbf{Q}_1$.

b) Asymptotic $n \to +\infty$. By the strong law of large numbers,

$$\tfrac{1}{n} \mathbf{N}_2^\top \mathbf{N}_2 \underset{\text{a.s.}}{\longrightarrow} \mathbf{\Lambda}_{\mathbf{\Gamma},2} . \tag{75}$$

Let us denote

$$\delta = \min_{i \in [\![k]\!]} \ \frac{\frac{1}{n} \kappa_i^2}{\frac{1}{n} \kappa_i^2 + \lambda_i^{\mathbf{\Theta}}} . \tag{76}$$

In this asymptotic regime, ensuring that all eigenvalues of $\widetilde{\mathbf{S}}^{-\frac{1}{2}} \widetilde{\mathbf{G}} \widetilde{\mathbf{S}}^{-\frac{1}{2}}$ on *noise components* $\mathbf{Q}_2$ are smaller than eigenvalues on *important components* $\mathbf{Q}_1$ gives the condition:

$$\forall i \in [\![k+1 : d]\!], \quad \delta \ > \ \frac{\lambda_i^{\mathbf{\Gamma}}}{\lambda_i^{\mathbf{\Theta}} + (1+\alpha^2) \lambda_i^{\mathbf{\Gamma}}} . \tag{77}$$

Rearranging this inequality, we obtain

$$\alpha^2 \ > \ \max_{i \in [\![k+1:d]\!]} \ \frac{1-\delta}{\delta} - \frac{\lambda_i^{\mathbf{\Theta}}}{\lambda_i^{\mathbf{\Gamma}}} . \tag{78}$$

Thus, if this condition is satisfied, the eigenvalues corresponding to the noise components are strictly smaller than those on the important components. Consequently, the k largest eigenvalues of $\widetilde{\mathbf{S}}^{-\frac{1}{2}} \widetilde{\mathbf{G}} \widetilde{\mathbf{S}}^{-\frac{1}{2}}$ come solely from the data subspace *i.e.* $\mathbf{Q}_k = \mathbf{Q}_1$.

Conclusion. Both in the large α limit for a fixed sample size, and in the large-sample limit under the condition given by Equation (12), we obtain that the optimal solution of the SSL problem takes the form:

$$\widetilde{\mathbf{W}^\star} = \mathbf{U} \left(\tfrac{1}{n} \mathbf{K}_1^2 + \mathbf{\Lambda}_{\mathbf{\Theta},1} \right)^{-\frac{1}{2}} \mathbf{Q}_1^\top = \mathbf{W}^\star , \tag{79}$$

where $\mathbf{U}$ is any orthogonal matrix of size $k \times k$. Therefore, in these two regimes the solution has the same form as in the uncorrupted data setting.

□

A.6 Proof of Corollary 4.5

Corollary 4.5. *Let α_{JE}, δ, α_{RC}, and η be defined as in Proposition 4.4 and Proposition 4.3.*

- *If $\max_{i\in[\![k+1:d]\!]} \lambda_i^{\mathbf{\Gamma}} < \dfrac{\eta^2}{\delta}$ (low noise), then $\alpha_{\mathrm{JE}} > \alpha_{\mathrm{RC}}$ (reconstruction is preferable).*
- *If $\min_{i\in[\![k+1:d]\!]} \lambda_i^{\mathbf{\Gamma}} > \dfrac{\eta^2}{\delta}$ (high noise), then $\alpha_{\mathrm{JE}} < \alpha_{\mathrm{RC}}$ (joint-embedding is preferable).*

Proof. Recall the definition of α_{JE} and α_{RC}:

$$\alpha_{\mathrm{JE}}^2 \coloneqq \max_{i\in[\![k+1:d]\!]} \frac{1-\delta}{\delta} - \frac{\lambda_i^{\mathbf{\Theta}}}{\lambda_i^{\mathbf{\Gamma}}} \quad \text{where} \quad \delta = \min_{i\in[\![k]\!]} \frac{\frac{1}{n}\kappa_i^2}{\frac{1}{n}\kappa_i^2 + \lambda_i^{\mathbf{\Theta}}}\,, \tag{80}$$

$$\alpha_{\mathrm{RC}}^2 \coloneqq \max_{i\in[\![k+1:d]\!]} \frac{\lambda_i^{\mathbf{\Gamma}}}{\eta^2} - \frac{\lambda_i^{\mathbf{\Theta}}}{\lambda_i^{\mathbf{\Gamma}}} - 1 \quad \text{where} \quad \eta = \min_{i\in[\![k]\!]} \frac{\frac{1}{n}\kappa_i^2}{\sqrt{\frac{1}{n}\kappa_i^2 + \lambda_i^{\mathbf{\Theta}}}}\,. \tag{81}$$

A sufficient condition for $\alpha_{\mathrm{JE}} > \alpha_{\mathrm{RC}}$ is the following, for any $i \in [\![k+1:d]\!]$:

$$\frac{1-\delta}{\delta} > \frac{\lambda_i^{\mathbf{\Gamma}}}{\eta^2} - 1\,. \tag{82}$$

Rearranging this inequality gives:

$$\forall i \in [\![k+1:d]\!], \quad \lambda_i^{\mathbf{\Gamma}} < \frac{\eta^2}{\delta}\,. \tag{83}$$

Then, a sufficient condition for $\alpha_{\mathrm{JE}} < \alpha_{\mathrm{RC}}$ is the following inequality, for any $i \in [\![k+1:d]\!]$:

$$\frac{1-\delta}{\delta} < \frac{\lambda_i^{\mathbf{\Gamma}}}{\eta^2} - 1\,. \tag{84}$$

It gives:

$$\forall i \in [\![k+1:d]\!], \quad \lambda_i^{\mathbf{\Gamma}} > \frac{\eta^2}{\delta}\,. \tag{85}$$

□

B Effect of Data Augmentation for Supervised Learning

We recall a well-known result that establishes the equivalence between the effect of data augmentation and ridge regularization. We provide the proof for completeness.

Lemma B.1. *[7, 46] For any $\mathbf{V} \in \mathbb{R}^{\ell\times d}$, it holds:*

$$\frac{1}{n}\sum_{i\in[\![n]\!]} \mathbb{E}_{\tau\sim\mathcal{T}}\left[\|\mathbf{y}_i - \mathbf{V}\tau(\mathbf{x}_i)\|_2^2\right] = \|\mathbf{V}\|_{\mathbf{\Sigma}}^2 + \frac{1}{n}\sum_{i\in[\![n]\!]} \|\mathbf{y}_i - \mathbf{V}\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]\|_2^2\,, \tag{86}$$

where

$$\mathbf{\Sigma} \coloneqq \frac{1}{n}\sum_{i\in[\![n]\!]} \mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\tau(\mathbf{x}_i)^\top\right] - \mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]^\top\,. \tag{87}$$

Proof.

$$\frac{1}{n}\sum_{i\in[\![n]\!]}\mathbb{E}_{\tau\sim\mathcal{T}}\left[\|\mathbf{y}_i-\mathbf{V}\tau(\mathbf{x}_i)\|_2^2\right] \tag{88}$$

$$=\frac{1}{n}\sum_{i\in[\![n]\!]}\|\mathbf{y}_i\|_2^2+\mathbb{E}_{\tau\sim\mathcal{T}}\left[\|\mathbf{V}\tau(\mathbf{x}_i)\|_2^2\right]-2\mathbb{E}_{\tau\sim\mathcal{T}}\left[\mathrm{Tr}\left(\mathbf{y}_i^\top\mathbf{V}\tau(\mathbf{x}_i)\right)\right] \tag{89}$$

$$=\frac{1}{n}\sum_{i\in[\![n]\!]}\|\mathbf{y}_i\|_2^2+\mathrm{Tr}\left(\mathbf{V}\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\tau(\mathbf{x}_i)^\top\right]\mathbf{V}^\top\right)-2\mathrm{Tr}\left(\mathbf{y}_i^\top\mathbf{V}\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]\right) \tag{90}$$

$$=\mathrm{Tr}\left(\mathbf{V}\mathbf{\Sigma}\mathbf{V}^\top\right)+\frac{1}{n}\sum_{i\in[\![n]\!]}\|\mathbf{y}_i\|_2^2-2\mathrm{Tr}\left(\mathbf{y}_i^\top\mathbf{V}\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]\right)+\|\mathbf{V}\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]\|_2^2 \tag{91}$$

$$=\|\mathbf{V}\|_{\mathbf{\Sigma}}^2+\frac{1}{n}\sum_{i\in[\![n]\!]}\|\mathbf{y}_i-\mathbf{V}\mathbb{E}_{\tau\sim\mathcal{T}}\left[\tau(\mathbf{x}_i)\right]\|_2^2\,. \tag{92}$$

□

C Experiments with Linear Models

In this section, we experiment with linear models and synthetic noise in order to validate the theoretical results of Section 4.

To process the data, we apply a PCA of dimension 50. We then add 50 components of noise to create a corrupted version of the dataset.

The eigenvalues for $\mathbf{\Gamma}$ and $\mathbf{\Theta}$ (as defined in Section 4.1) are randomly sampled from uniform distributions. Specifically, eigenvalues for $\mathbf{\Gamma}$ are drawn from the range $[0,\lambda_{\max}^{\mathbf{\Gamma}}]$, and eigenvalues for $\mathbf{\Theta}$ are drawn from $[0,\lambda_{\max}^{\mathbf{\Theta}}]$.

Across all experiments, we set a constant value for $\lambda_{\max}^{\mathbf{\Theta}}=10^4$. To investigate the impact of input noise levels on model performance, we vary the value of $\lambda_{\max}^{\mathbf{\Gamma}}$. In the experiments presented in Figure 3, we define a weak noise case with $\lambda_{\max}^{\mathbf{\Gamma}}=10^3$ and a strong noise case with $\lambda_{\max}^{\mathbf{\Gamma}}=10^6$.

Joint-embedding and reconstruction solutions are obtained using the closed forms provided in Theorems 3.1 and 3.2. For evaluation, we compute the supervised linear probing score:

$$\min_{\mathbf{V}\in\mathbb{R}^{\ell\times k}}\ \frac{1}{n}\sum_{i\in[\![n]\!]}\|\mathbf{y}_i-\mathbf{V}\mathbf{z}_i\|_2^2\,, \tag{93}$$

where $\mathbf{z}_i$ is the output of the SSL model on the i-th sample; *i.e.* $\mathbf{z}_i=\mathbf{W}^\star\mathbf{x}_i$ for joint embedding where $\mathbf{W}^\star$ is the optimal joint-embedding model of Theorem 3.2 and $\mathbf{z}_i=\mathbf{E}^\star\mathbf{x}_i$ for reconstruction where $\mathbf{E}^\star$ is the optimal encoder of the reconstruction model of Theorem 3.1. We then compute the absolute difference between the score of the model trained on clean data (*i.e.* , composed only of *important features*), and the score of the model trained on corrupted data (with the added *irrelevant noisy features*). As the model trained only on the *important features* naturally selects these features as SSL representations, this absolute difference directly quantifies the model's ability to filter out the *irrelevant noisy features*.

Figures 4, 5, 6, and 7 illustrate the experimental results, all of which support the intuitions outlined in Section 5.1. Notably, supervised models consistently discard noisy irrelevant components with increasing sample size or alignment strength, confirming the findings of Section A.3. However, SSL models demonstrate varied success depending on the setting. The *reconstruction* SSL model fails to retrieve *important components* from data with high noise magnitude, even with larger sample sizes and important alignment strength. Yet, with low noise, it successfully identifies these components and remains robust to alignment strength. In contrast, the *joint-embedding* SSL model requires a certain minimum alignment strength to filter out noisy irrelevant components, even in low-noise settings, but it exhibits strong robustness to increasing noise magnitude (bottom figures).

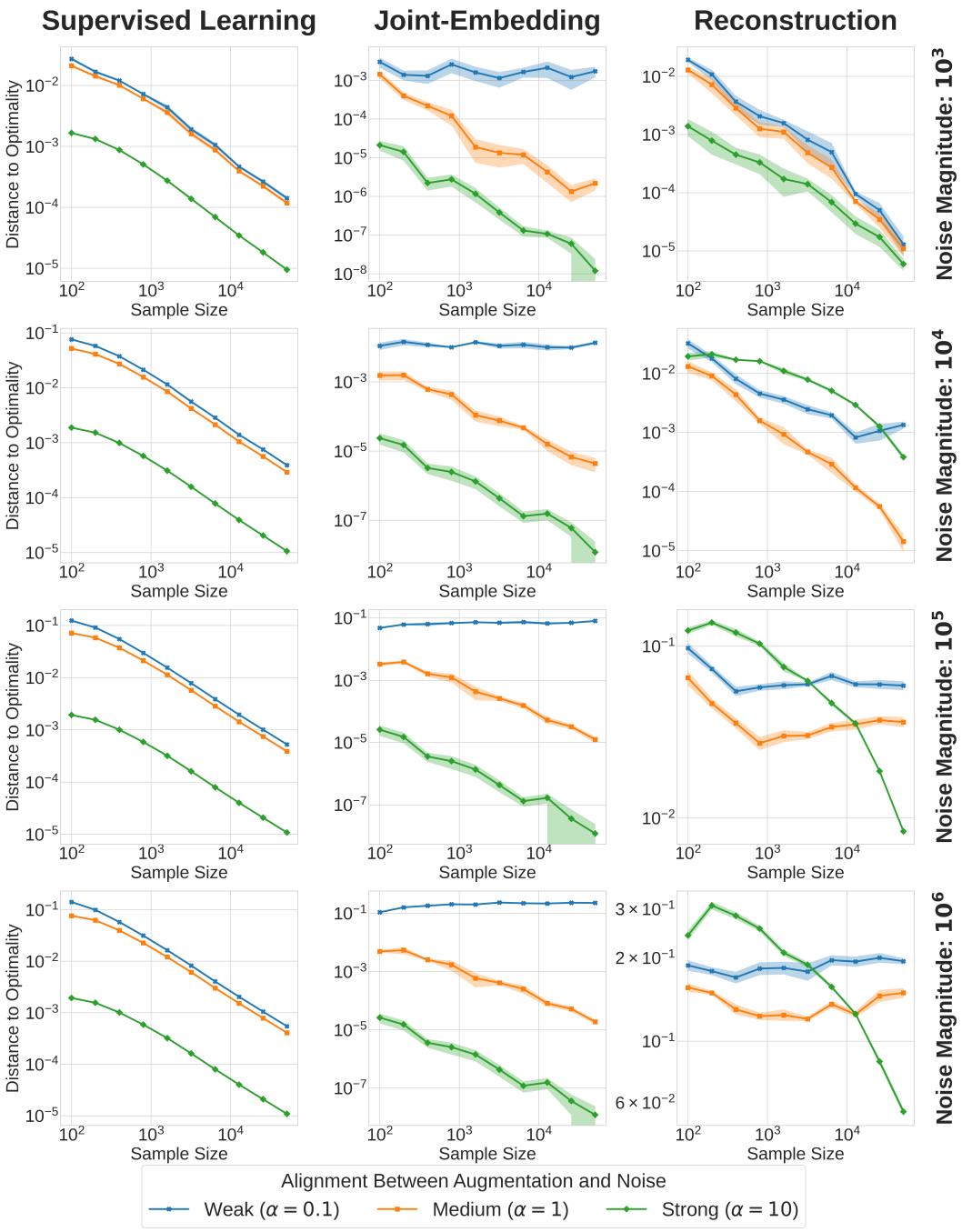

Figure 4: Performance of linear supervised and SSL models (Sections 3.1 and 3.2 and Theorems 3.1 and 3.2) with synthetic noise (Section 4.1 and Appendix C) on **MNIST**. Each subplot's y-axis is the absolute difference of supervised linear probing loss (on clean vs. corrupted data) and its x-axis is the sample size n.

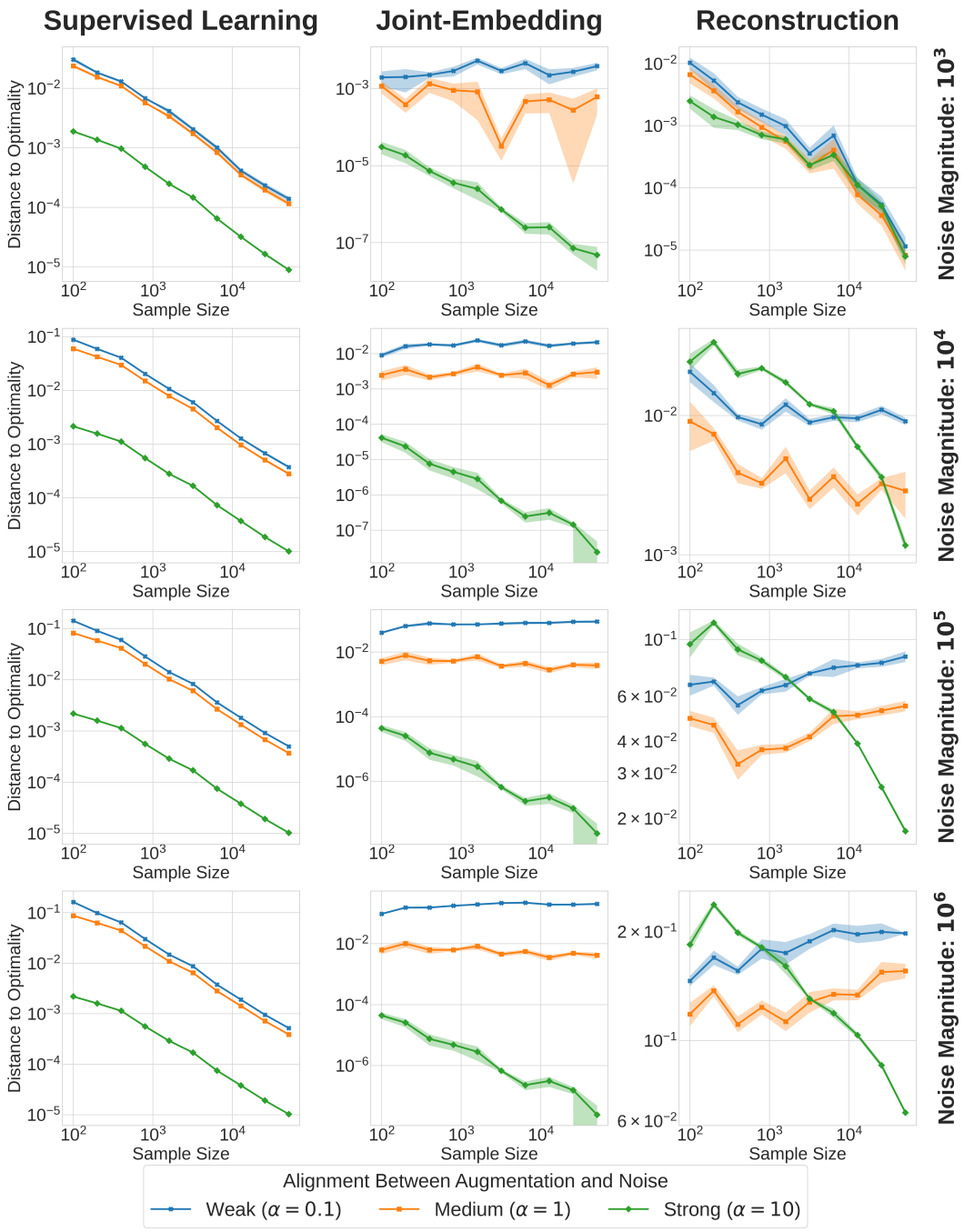

Figure 5: Performance of linear supervised and SSL models (Sections 3.1 and 3.2 and Theorems 3.1 and 3.2) with synthetic noise (Section 4.1 and Appendix C) on **Fashion-MNIST**. Each subplot's y-axis is the absolute difference of supervised linear probing loss (on clean vs. corrupted data) and its x-axis is the sample size n.

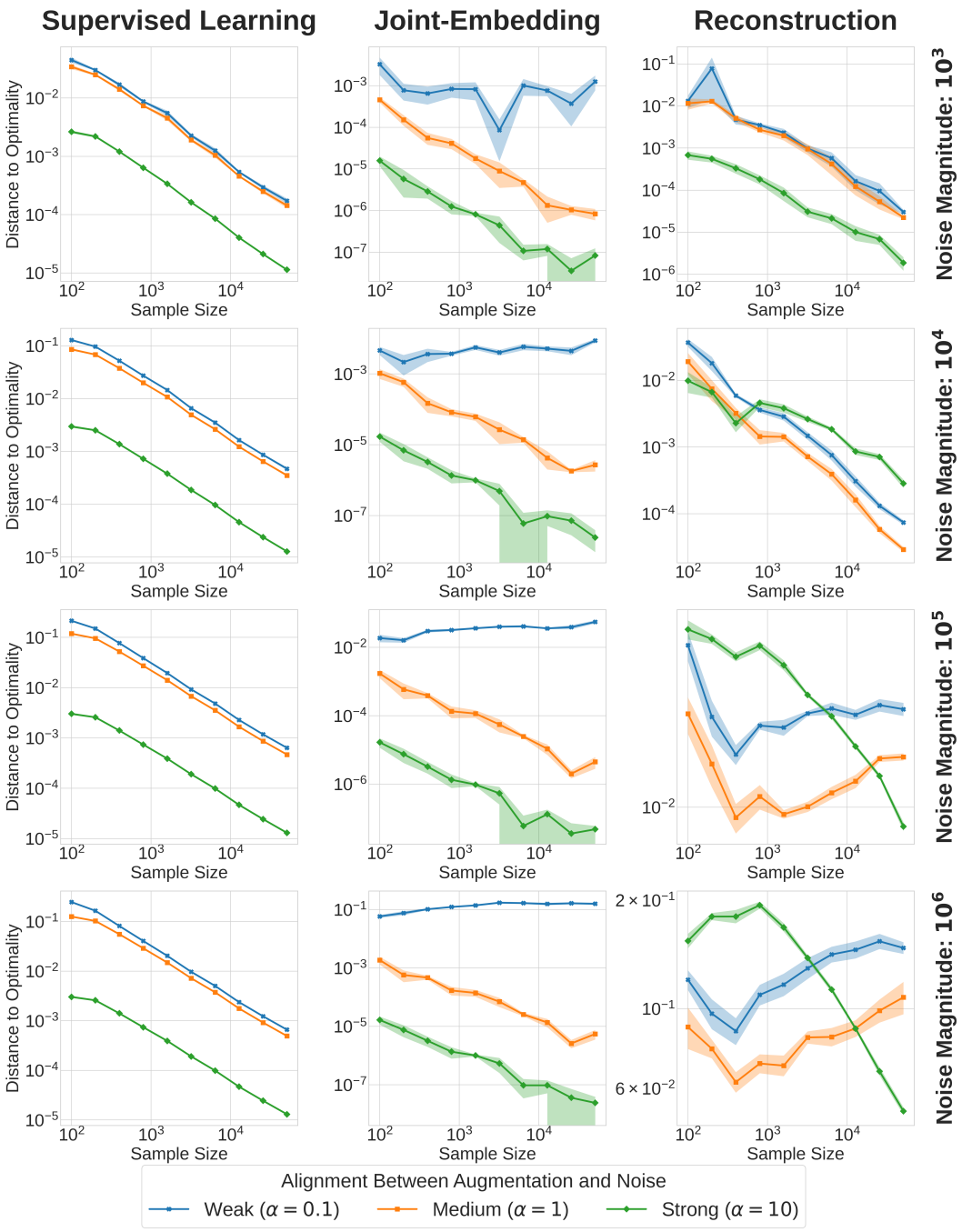

Figure 6: Performance of linear supervised and SSL models (Sections 3.1 and 3.2 and Theorems 3.1 and 3.2) with synthetic noise (Section 4.1 and Appendix C) on **Kuzushiji-MNIST** characters [16]. Each subplot's y-axis is the absolute difference of supervised linear probing loss (on clean vs. corrupted data) and its x-axis is the sample size n.

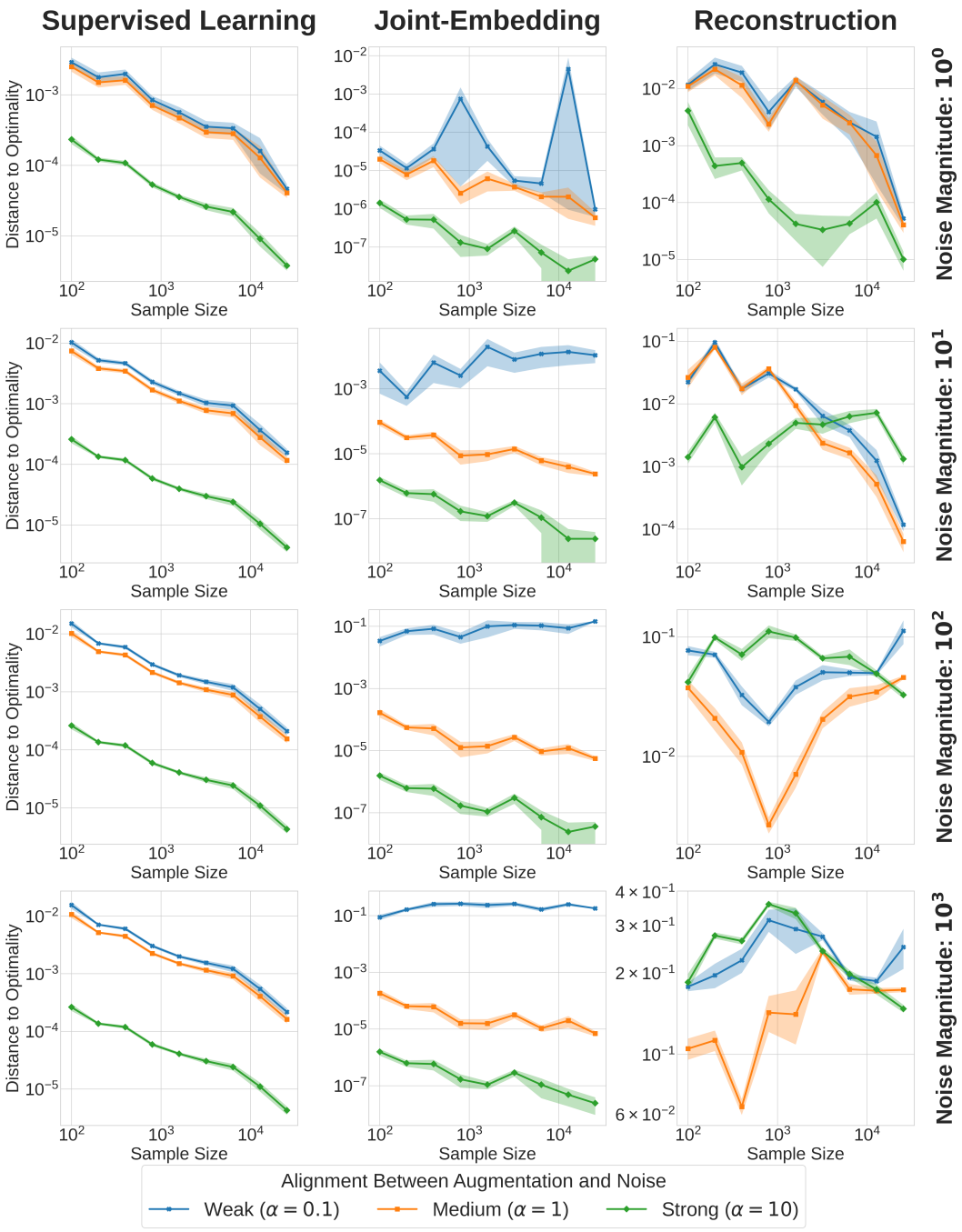

Figure 7: Performance of linear supervised and SSL models (Sections 3.1 and 3.2 and Theorems 3.1 and 3.2) with synthetic noise (Section 4.1 and Appendix C) on a **single-cell RNA seq** dataset from [49]. Each subplot's y-axis is the absolute difference of supervised linear probing loss (on clean vs. corrupted data) and its x-axis is the sample size n.

D Experiments with Deep Networks on Image Classification

D.1 Details about the Experiments

This section details the experimental setup and hyperparameters used in the experiments.

For the evaluation of SSL methods, we freeze the learned representations and then conduct linear probing, reporting the top-1 accuracy on the test set. For the data augmentation pipelines and SSL modules, we use default modules from the `lightly` library [55].

To introduce *irrelevant noisy features* in the data, we use the ImageNet-C corruptions [35]. This benchmark provides a comprehensive suite of corruptions designed to assess model robustness. ImageNet-C corruptions span categories such as noise, blur, weather, and digital distortions, each available at multiple levels of severity. For each experiment, we create a *fixed corrupted dataset* by assigning a deterministic corruption to each image, ensuring every time an image is loaded it undergoes the same corruption. This effectively simulates a corrupted dataset, consistent across epochs. For the CIFAR-10 dataset, these corruptions were adapted to the specific image dimensions.

D.1.1 ImageNet Experiments

For the ImageNet [18] experiments, we use a batchsize of 256 and train the model for 500 epochs. We use a ResNet-50 backbone [34] for BYOL [26] with a learning rate of 0.45 and a weight decay of 10^{-6}, with the LARS optimizer [69]. We use a cosine scheduler from 0.99 to 1 for the student momentum parameter. For MAE and DINO, we use a ViT-B/16 backbone [21], and the AdamW optimizer [48] with a learning rate of 1.5×10^{-4} and a weight decay of 0.05.

For all methods, the hyperparameters, including architectural choice, were set to the values for ImageNet presented in their respective original papers. For augmentations, we use the default augmentations associated to BYOL, DINO and MAE from the lightly library [55] with default parameters.

D.1.2 CIFAR10 Experiments

In Appendices D.2 and D.3, we perform experiments on the CIFAR-10 [42] dataset. For these experiments, we use a ResNet-50 backbone [34], a batch size of 256, and train the model for 1000 epochs. The LARS optimizer [69] is employed with a learning rate of 5 and a weight decay of 10^{-6}. Supervised training is conducted using the same architecture and optimization parameters. All experiments are run with 5 different random seeds.

We conduct experiments using three SSL methods: SimCLR [14], BYOL [26], and VICReg [10]. For SimCLR, we set the temperature parameter to $\tau = 0.5$. For VICReg, we use the following default hyperparameters: a scaling coefficient of 25 for the invariance term of the loss, 25 for the variance term, and 1 for the covariance term. For BYOL, we use a cosine scheduler from 0.99 to 1 for the student momentum parameter. For augmentations, we use the default augmentations associated to SimCLR, BYOL and VICReg from the `lightly` library [55] adapted to the CIFAR-10 dataset as follows: the random resized crop is set to 32 and Gaussian blur is removed.

D.2 Self-Supervised Learning is much more Sensitive to Corruptions than Supervised Learning

In this section, we compare the performances of SimCLR against a supervised model.

Scores for SimCLR and supervised learning with the same augmentations on corrupted datasets at various corruption strengths are shown in Figure 8. We observe that the performance of SimCLR tends to degrade rapidly as the level of noise in the data increases. A similar trend is observed for supervised learning, but the decline is significantly less steep. For instance, performance on corruptions such as fog and frost remains relatively stable across corruption strengths for supervised learning, whereas for SSL, performance can drop by a factor of two (e.g., from approximately 0.8 to 0.4 in top-1 accuracy).

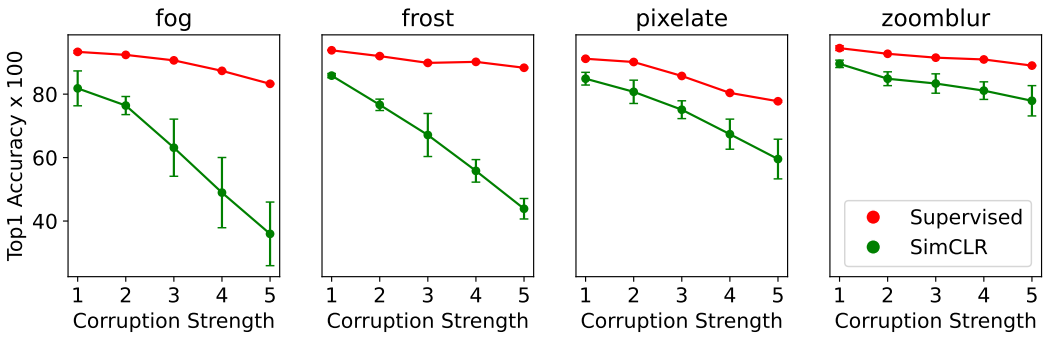

Figure 8: Top-1 accuracy for supervised learning and self-supervised learning methods on CIFAR-10 under various corruptions with severity levels ranging from 1 to 5. SSL is performed using SimCLR with default augmentations while supervised learning uses the same augmentations as SimCLR. We observe that supervised learning exhibits greater robustness to data corruption compared to SSL. This confirms the theoretical results of Section 4.

This phenomenon is confirmed by examining the learned representations of VICReg in Figure 2, without and with noisy corruptions, also on CIFAR10. While the supervised model maintains a clear separation of clusters even under noisy conditions, the VICReg representations degrade significantly and fail to distinguish clusters effectively when data is strongly corrupted.

These observations corroborate the results proved in Section 4. Indeed, the lack of labels prevents the model from compensating for any misalignment between augmentation and noise. Consequently, SSL performance deteriorates rapidly in the presence of corruptions if the data augmentations are not appropriately adapted.

D.3 Aligning Augmentation with Known Noise

In this section, we validate the findings of Propositions 4.3 and 4.4 and demonstrate empirically the impact of aligning the data augmentation with the *irrelevant noisy features* in the data. To do so, we artificially create misalignment with usual augmentations by applying a known corruption to the dataset. We then investigate whether augmenting the data with the same type of noise can help the model learn better representations.

Specifically, for each view, we append a transform of the same corruption type but not necessarily the same strength, at the end of the data augmentation pipeline. We refer to the corruption applied to the dataset as the *Corruption Strength* and the additional corruption appended during augmentation as the *Augmentation Strength*. For each corruption tested, we evaluate the *Corruption Strength* over the set $\{1, 2, 3, 4, 5\}$ and the *Augmentation Strength* over the set $\{0, 1, 2\}$.

Experiments are conducted on CIFAR10 and results are displayed in Figure 9. For each heatmap, we are interested in verifying if the top two rows corresponding to *augmentation strength* $= 1$ and 2 yield better results than the bottom row without noise injection (*augmentation strength* $= 0$). Looking at SimCLR runs, we observe that *noise injection* generally boosts performance in most settings. This holds true for all 16 corruptions, except for spatter and brightness, where an augmentation strength of 0 performs best. Some corruptions clearly demonstrate that noise injection improves performance regardless of the corruption strength. For instance, this is evident for fog, Gaussian blur, and glass blur. For other corruptions, the utility of noise injection depends on the strength of the noise. For example, frost, saturate, and snow benefit from noise injection when the corruption in the input data is strong, whereas for impulse noise, JPEG compression, and defocus blur, noise injection in the augmentation pipeline is more effective when the corruption in the input data is not severe. In some cases, the improvements are substantial; for instance, for fog with a corruption strength of 4, adding noise injection with a strength of 2 increases the top-1 accuracy from 49.0 to 67.1. Across all configurations of noise and corruption strength, a total

of 80 combinations are tested with SimCLR, comprising 16 corruption types at 5 strength levels. Noise injection leads to an improvement in top-1 accuracy in 67.5% of these cases.

We observe that these trends are quite consistent across various SSL methods. For VICReg, noise injection improves top-1 accuracy in 85% of the configurations, while for BYOL, this improvement is observed in 60% of the tested configurations. For VICReg, noise injection results in significant improvements; for instance, in the case of fog with *corruption strength* = 5, the top-1 accuracy increases dramatically from 43.5 to 70.9. This performance improvement is evident in Figure 2, where the clusters corresponding to class labels are more clearly separated when noise injection is applied. For completeness, we provide in Figure 10 the top-1 linear probing accuracy scores evaluated on a clean test set (standard CIFAR-10 test set). We observe that for SimCLR, noise injection improves top-1 accuracy in 68.75% of the configurations, while for both BYOL and VICReg, it enhances performance in 85% of the configurations.

These outcomes confirm the results of Proposition 4.4. Aligning augmentations with noise directs the model's focus to important features, thereby enhancing SSL representations. Note that our theory does not cover cases where data and noise components are intertwined, as seen with *e.g.* spatter, where strong noise may discard important data features and render *noise injection* less effective. Finally, one key observation is that, even under substantial data corruption, a small injection of noise during augmentation can still yield benefits. Hence the augmentation noise strength does not need to be precisely tuned to match the severity of the corruption.

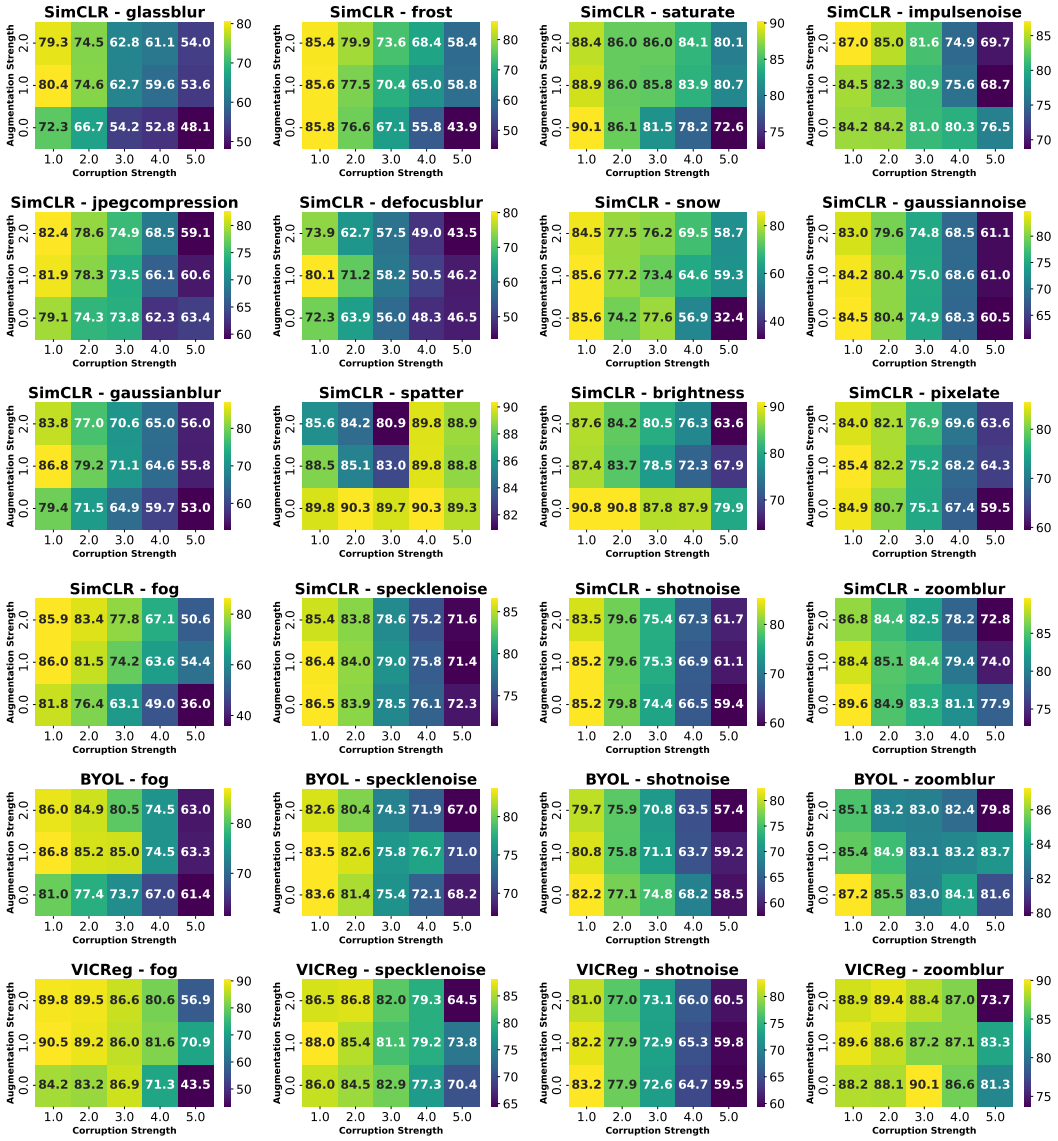

Figure 9: Top-1 linear probing accuracy of SSL methods evaluated on CIFAR10 test set with the same corruption type and severity as the training set. The experiments investigate varying levels of corruption severity in the input data (Corruption Strength) and different levels of noise injection severity in the augmentation pipeline (Augmentation Strength). The noise injection corruption type matches the data corruption type. Each reported value is an average over 5 random seeds. The first four rows present SimCLR results across 16 distinct corruptions: glassblur, frost, saturate, impulsenoise, jpegcompression, defocusblur, snow, gaussiannoise, gaussianblur, spatter, brightness, pixelate, fog, specklenoise, shotnoise and zoomblur. These transformations are sourced from Imagenet-C [35]. The bottom two rows display results for BYOL and VICReg on a subset of 4 corruptions: fog, specklenoise, shotnoise and zoomblur.

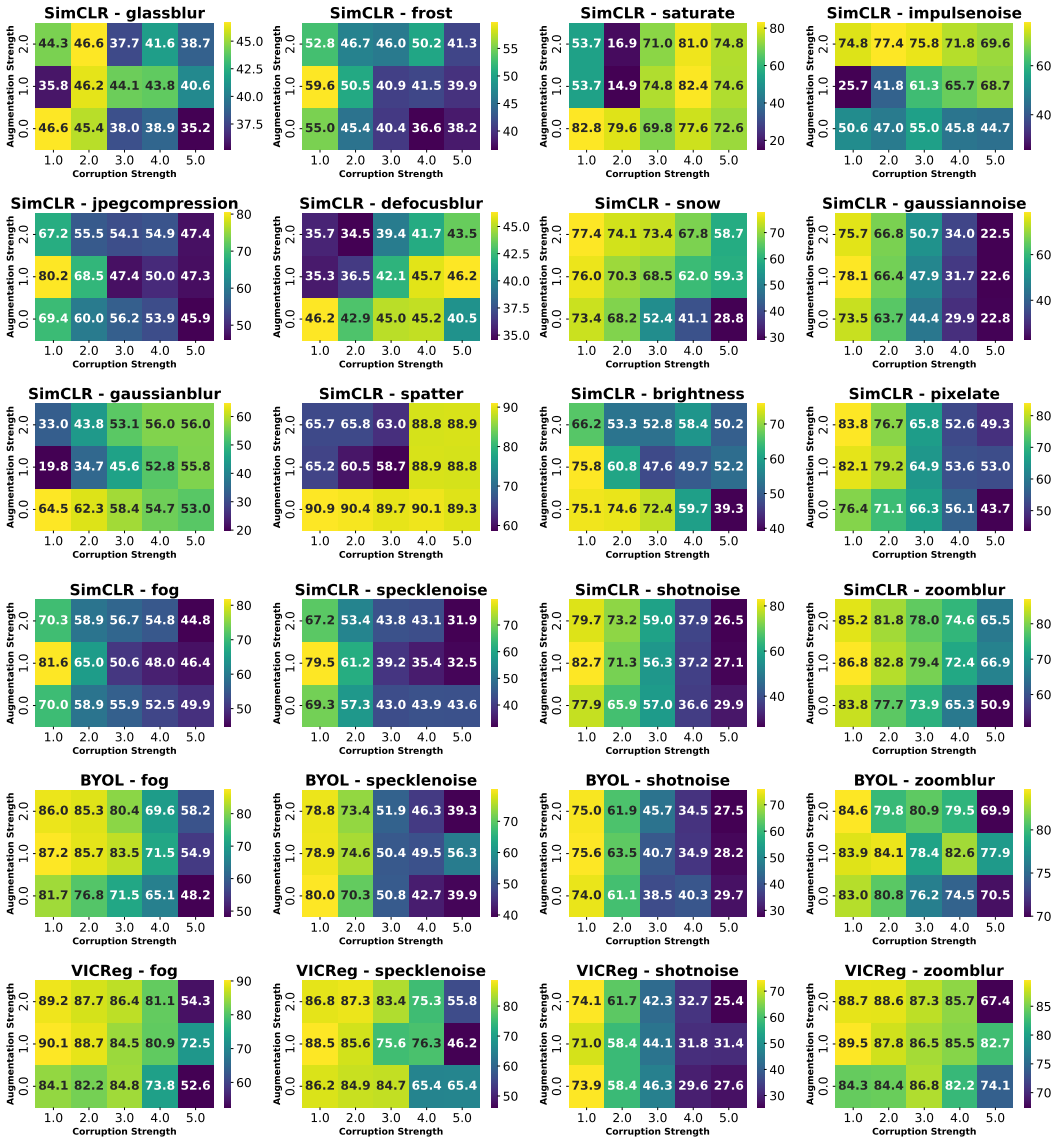

Figure 10: Top-1 linear probing accuracy for SSL methods evaluated on the clean (uncorrupted) CIFAR10 test set. The experiments investigate varying levels of corruption severity in the input data (Corruption Strength) and different levels of noise injection severity in the augmentation pipeline (Augmentation Strength). The noise injection corruption type matches the data corruption type. Each reported value is an average over 5 random seeds. The first four rows present SimCLR results across 16 distinct corruptions: glassblur, frost, saturate, impulsenoise, jpegcompression, defocusblur, snow, gaussiannoise, gaussianblur, spatter, brightness, pixelate, fog, specklenoise, shotnoise and zoomblur. These transformations are sourced from Imagenet-C [35]. The bottom two rows display results for BYOL and VICReg on a subset of 4 corruptions: fog, specklenoise, shotnoise and zoomblur.