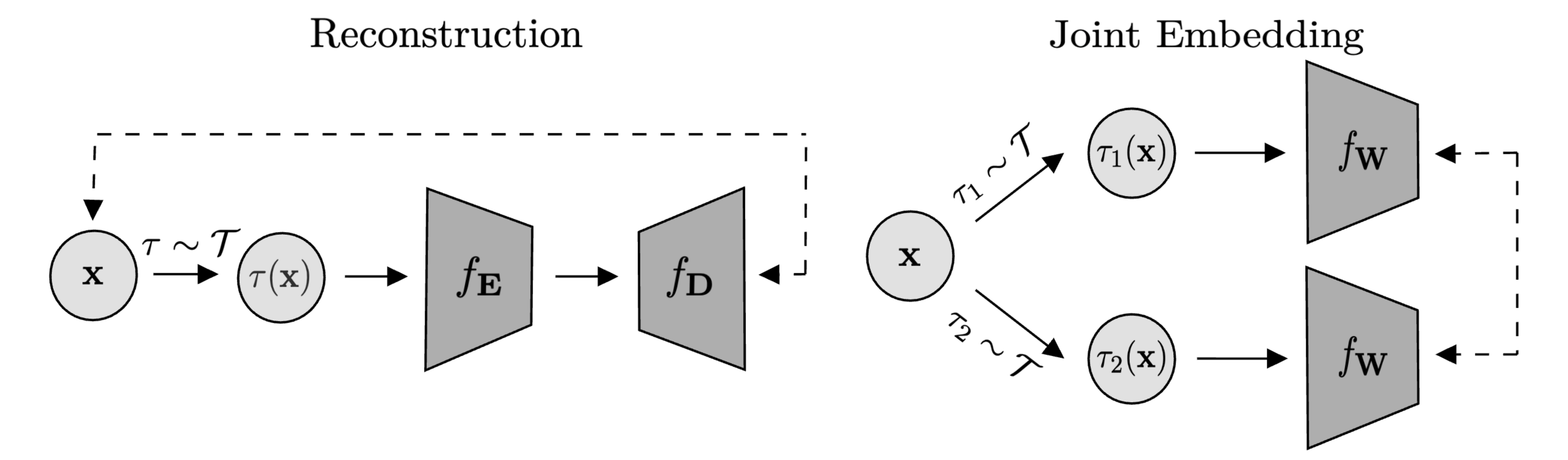
Joint-Embedding vs Reconstruction
Provable Benefits of Latent Space Prediction for Self-Supervised Learning

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

Paper Blog

Two Paradigms of SSL



Reconstruction-Based SSL (RC)

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

Examples: Large Language Models; Masked Autoencoder

Joint-Embedding SSL (JE)

$$\min_{\mathbf{W}} \frac{1}{n}\sum_{i\in[n]} \mathbb{E}_{\tau_1,\tau_2\sim\mathcal{T}}\left[\left\|f_{\mathbf{W}}(\tau_1(\mathbf{x}_i)) - f_{\mathbf{W}}(\tau_2(\mathbf{x}_i))\right\|_2^2\right]$$

$$\text{s.t. } \frac{1}{n}\sum_{i} \mathbb{E}_{\tau\sim\mathcal{T}}\left[f_{\mathbf{W}}(\tau(\mathbf{x}_i)) f_{\mathbf{W}}(\tau(\mathbf{x}_i))^\top\right] = \mathbf{I}_k$$

Examples: SimCLR; BYOL; DINO; VICReg; JEPA

Setup: Noise, Augmentations, and Alignment

Controlled setup: we study regimes of alignment between augmentations and the true noise.

Data model (corrupted inputs): $\forall\, i \in [\![n]\!],\ \widetilde{\mathbf{x}}_i = \mathbf{x}_i + \boldsymbol{\gamma}_i,\ \boldsymbol{\gamma}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$

$$\mathcal{T}(\alpha) := \left\{\tau \mid \tau(\mathbf{x}) = \mathrm{x} + \boldsymbol{\theta} + \alpha\boldsymbol{\gamma},\ \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}),\ \boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})\right\}$$

α controls augmentation–noise alignment: increasing α adds augmentation along the directions of the *irrelevant features* (data noise $\boldsymbol{\gamma}$).

Supervised Works Regardless of Augmentations

Proposition (Supervised Learning)

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

Let $\mathbf{V}^\star$ and $\widetilde{\mathbf{V}}^\star$ solve the above on clean and corrupted data. Then

$$\widetilde{\mathbf{V}}^\star \xrightarrow[\text{a.s.}]{} \mathbf{V}^\star$$

(i) $\alpha \to \infty$ for any n (Perfect alignment).
(ii) $n \to \infty$ for any α (Large sample size **for any alignment**).

SSL Requires Aligned Augmentations

Proposition (Self-Supervised Learning)

Let $\mathbf{W}^\star$, $\widetilde{\mathbf{W}}^\star$ (resp. $\mathbf{E}^\star$, $\widetilde{\mathbf{E}}^\star$) solve JE (resp. RC) on clean and corrupted data. Then

$$\widetilde{\mathbf{W}}^\star \xrightarrow[\text{a.s.}]{} \mathbf{W}^\star \quad \text{and} \quad \widetilde{\mathbf{E}}^\star \xrightarrow[\text{a.s.}]{} \mathbf{E}^\star$$

(i) $\alpha \to \infty$ for any n (Perfect alignment).
(ii) $n \to \infty$ iff $\alpha \geq \alpha_{\mathrm{JE}}$ (resp. $\alpha \geq \alpha_{\mathrm{RC}}$).

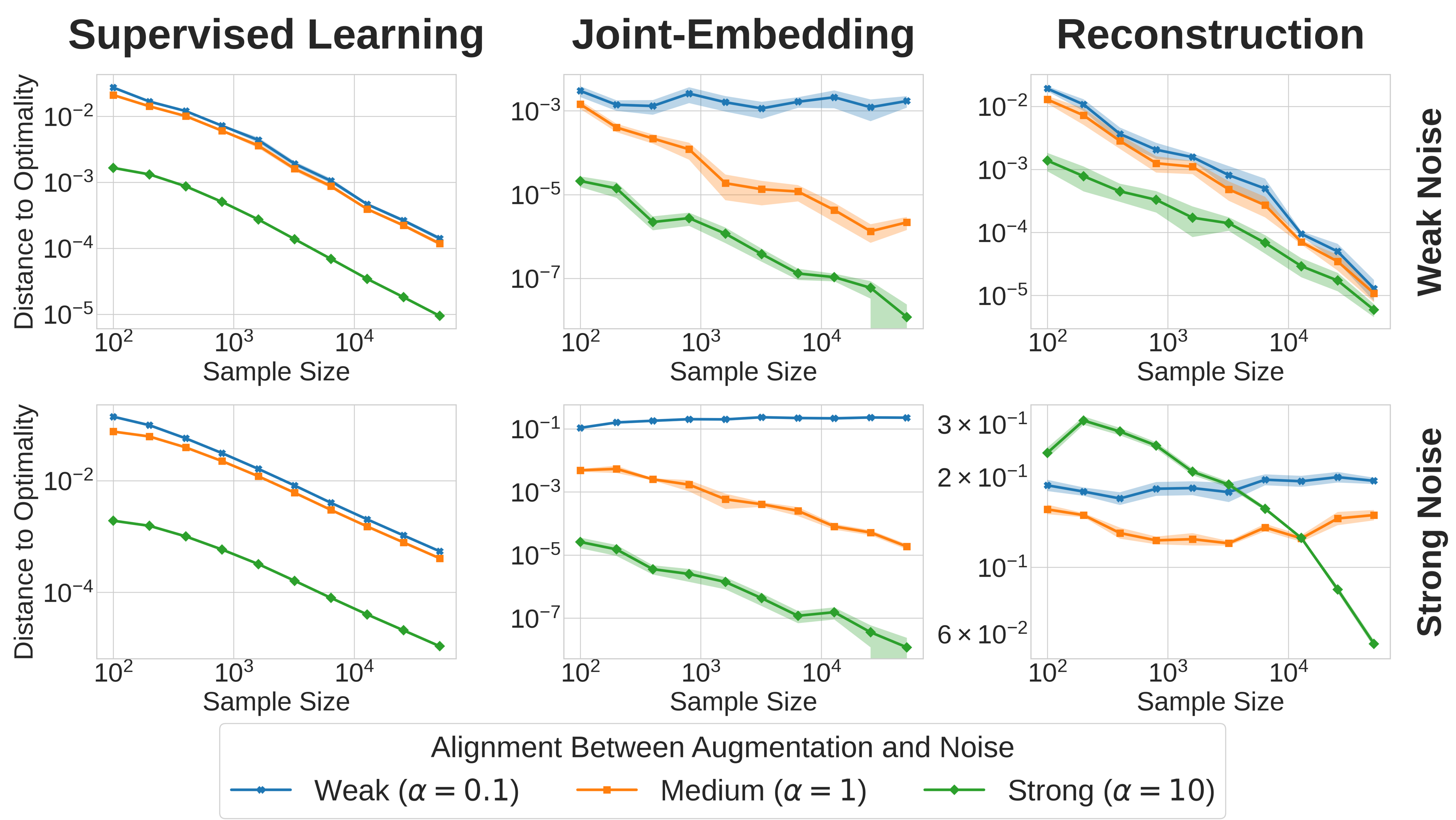
Key difference: Unlike supervised learning, SSL **cannot** overcome misalignment by increasing sample size alone. For SSL to benefit from more samples, augmentations must first be sufficiently aligned with irrelevant features.

JE vs RC: Different Alignment Thresholds

Smaller threshold is better (less alignment needed)
High-magnitude noise: $\alpha_{\mathrm{JE}} < \alpha_{\mathrm{RC}}$ (JE better)
Low-magnitude noise: $\alpha_{\mathrm{RC}} < \alpha_{\mathrm{JE}}$ (RC better)

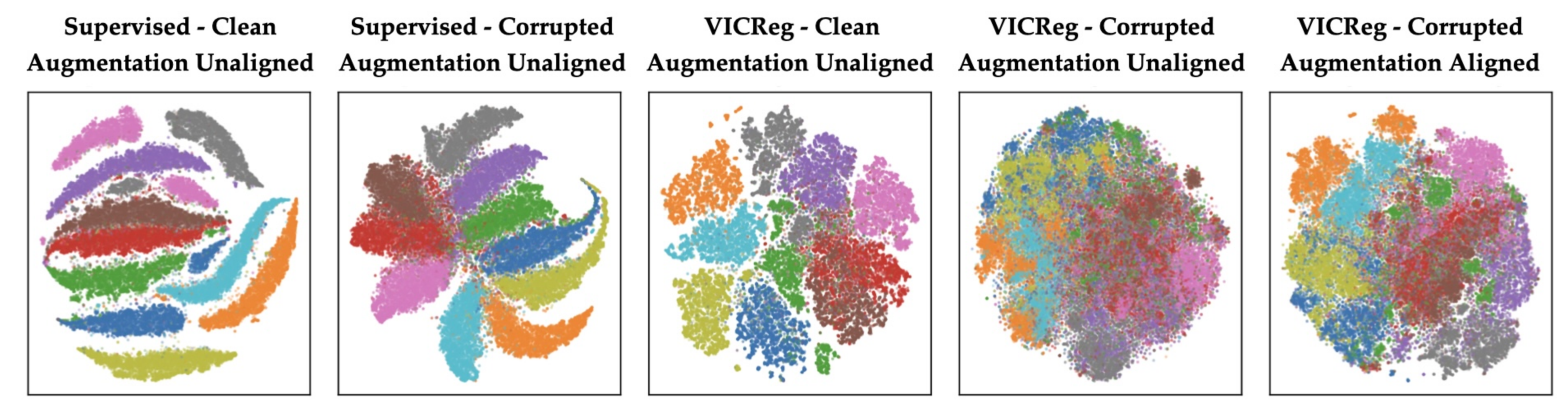
Magnitude refers to eigenvalues of $\boldsymbol{\Gamma}$: low = small max eigenvalue; high = large min eigenvalue.

Experimental Validation (MNIST)



Y-axis: distance to optimality vs. sample size n. Supervised is consistent for any α. Under strong noise, RC fails unless α is very large, while JE succeeds for a wider range of α. Under weak noise, RC requires less alignment to recover optimal performance.

Supervised vs SSL Under Corruption



t-SNE visualizations on CIFAR-10 (left to right): supervised (clean), supervised (fog-corrupted), VICReg (clean), VICReg (fog-corrupted), VICReg (fog-corrupted + aligned augmentation). Unlike supervised, SSL degrades under corruption. Aligning augmentations with noise recovers class separability.

ImageNet-C Corruptions: JE vs RC

	Pixelate				Gaussian Noise				Zoomblur			
Method	S1	S3	S5	Drop	S1	S3	S5	Drop	S1	S3	S5	Drop
MAE	64.9	52.3	46.8	28%	61.6	46.7	44.8	27%	64.1	58.4	51.3	20%
DINO	68.7	64.9	60.2	12%	67.6	62.4	59.0	13%	69.4	67.2	64.9	7%
BYOL	66.7	61.3	58.7	12%	67.2	63.1	56.4	16%	70.1	67.0	63.8	9%

Linear probing accuracy (S1/S3/S5 = Severity). RC (MAE) drops $\sim 2\times$ more than JE methods!

When noise is added to data, creating misalignment between augmentations and the corrupted inputs, RC methods (MAE) degrade significantly faster than JE methods (DINO, BYOL).

Interpretations

The choice between JE and RC depends on whether statistically dominant features are semantically meaningful.

Language: Tokens are semantically compressed. Predicting masked tokens operates directly in semantic space. High-variance IS high-semantics, so RC works well.

Vision & sensors: Pixels and physical measurements contain high-variance features (e.g., textures, edges, noise) that are statistically dominant but semantically shallow. RC learns what's dominant, not what's useful. JE filters noise by focusing on shared semantic content across views.

Recommendations

Use RC **(input-space prediction):** Low-magnitude irrelevant features. Biased toward high-variance components.

Use JE **(latent-space prediction):** High-magnitude irrelevant features. Avoids reconstructing noise.