Joint Embedding vs Reconstruction

Provable Benefits of Latent Space Prediction for Self-Supervised Learning

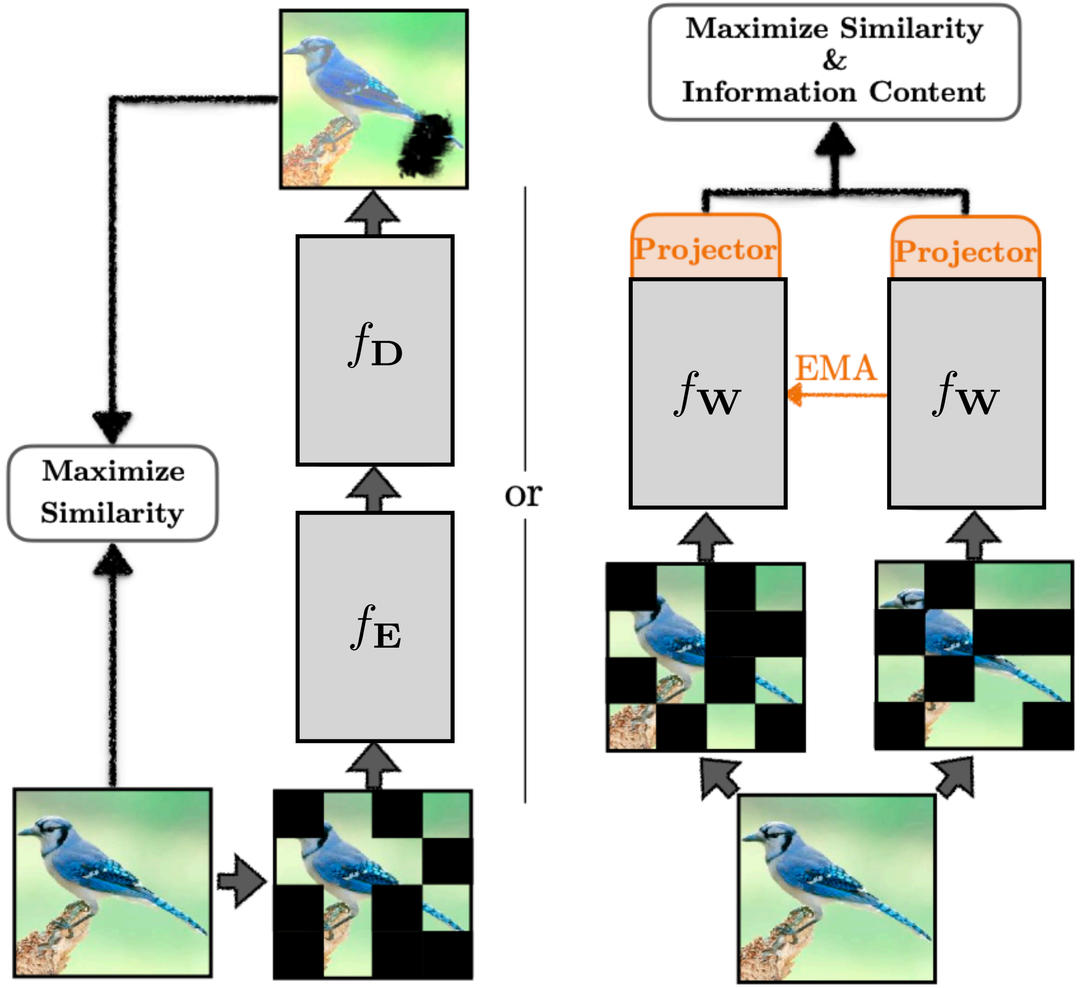
NeurIPS 2025

M. Ibrahim

T. Biancalani

A. Regev

R. Balestriero



In modern ML, most modalities follow a common pattern

Step 1 : Pretraining

- Learn data representations via a pretext task
- Train on very large diverse dataset
- Learn via **view generation / invariances**

Step 2 : Fine-tuning

- Adapt the model to the target task
- Train with labeled data from target domain

In modern ML, most modalities follow a common pattern

Step 1 : Pretraining

- Learn data representations via a pretext task
- Train on very large diverse dataset
- Learn via **view generation / invariances**

Our focus today !

Step 2 : Fine-tuning

- Adapt the model to the target task
- Train with labeled data from target domain

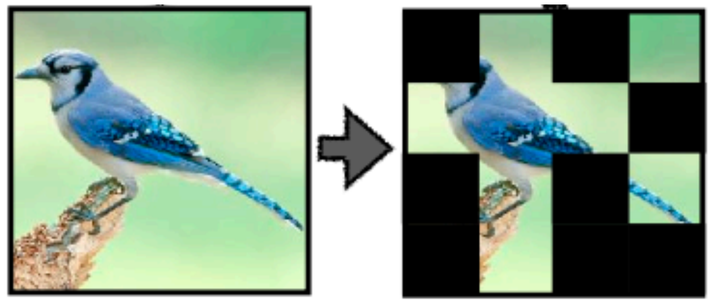
Step 1 : Pretraining

- Learn data representations via a pretext task
- Train on very large diverse dataset
- Learn via **view generation / invariances**

Our focus today !

View generation / Invariance is encoded via a data augmentation function:

$$\tau \sim \mathcal{T}$$



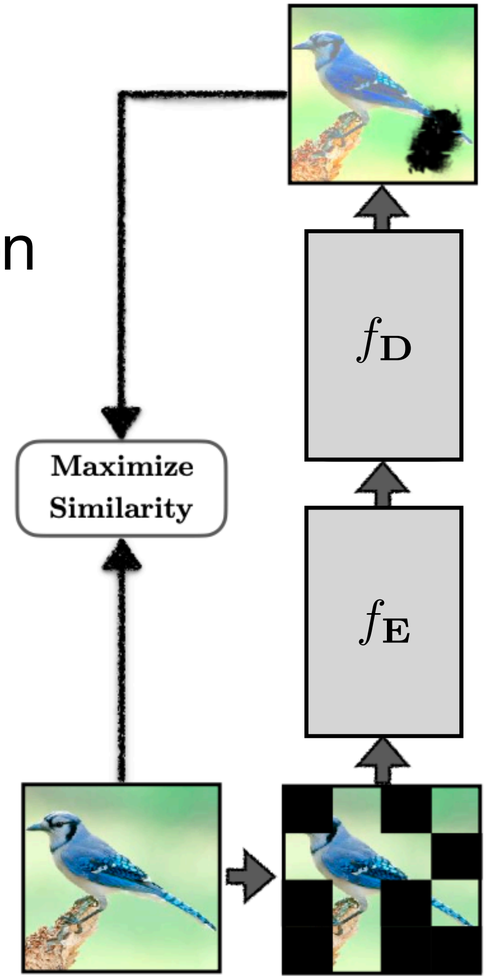
How to go from view generation to pre-training ?

Option A : Reconstruction

Input-space prediction

Ex : Auto-Encoder, LLMs

Popular in NLP.



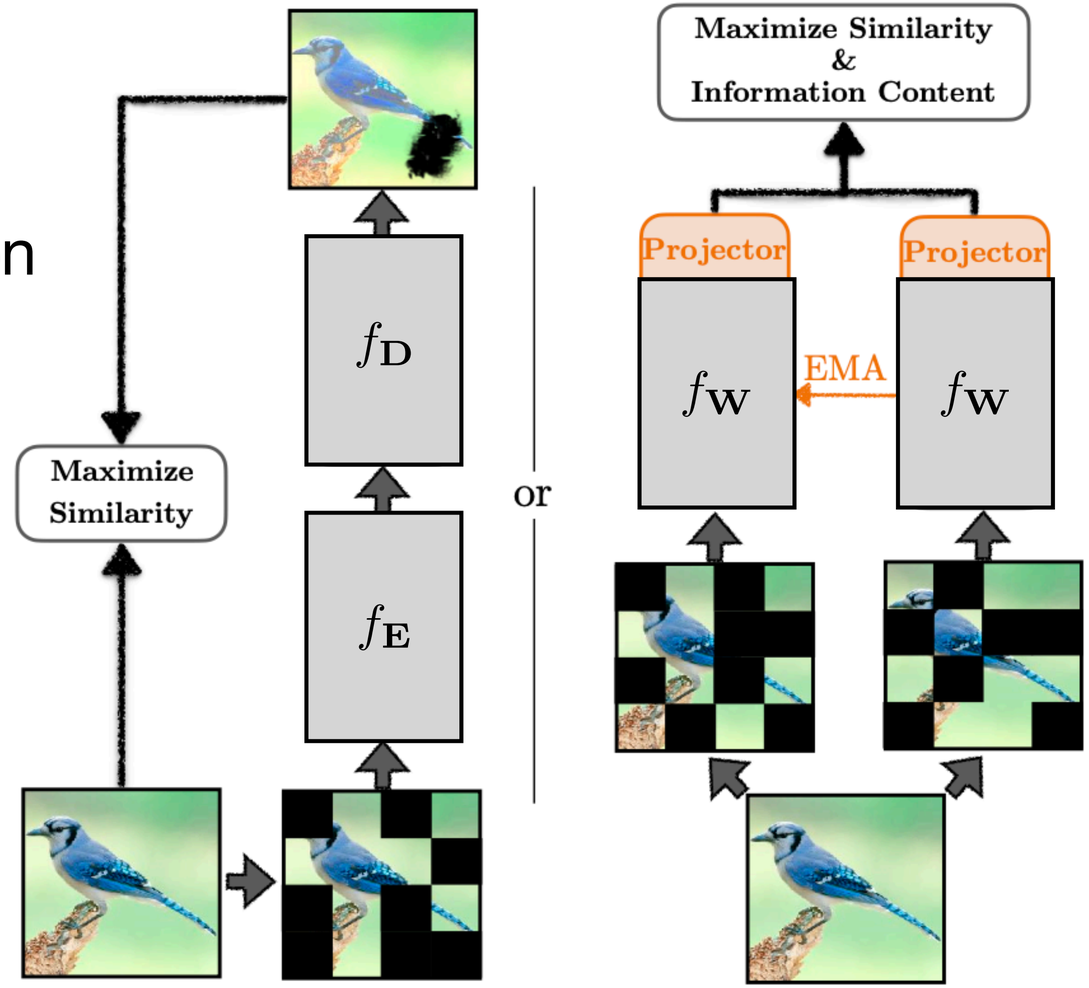
How to go from view generation to pre-training ?

Option A : Reconstruction

Input-space prediction

Ex : Auto-Encoder, LLMs

Popular in NLP.



Option B : Joint Embedding

Latent-space prediction

Ex : SimCLR, DINO, CLIP

Popular in computer vision.

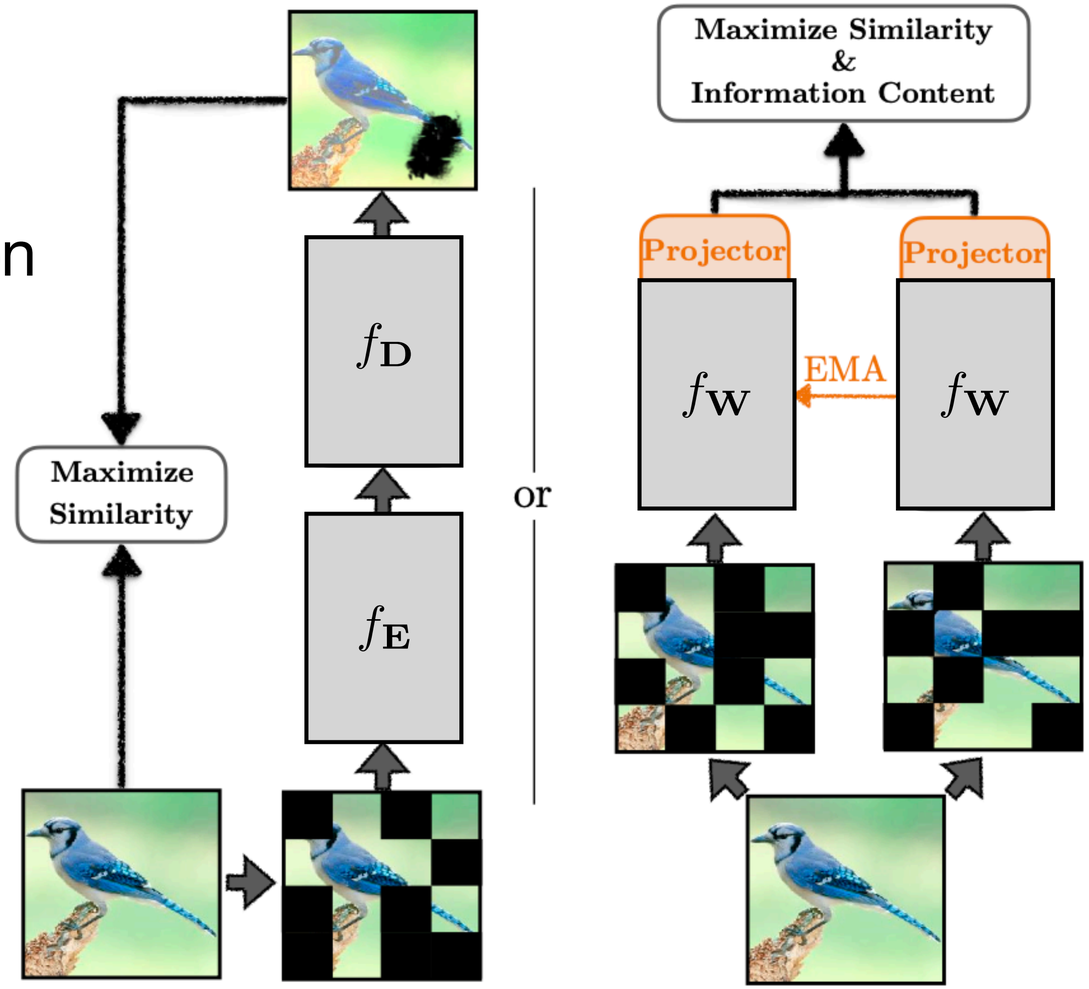
How to go from view generation to pre-training ?

Option A : Reconstruction

Input-space prediction

Ex : Auto-Encoder, LLMs

Popular in NLP.



Option B : Joint Embedding

Latent-space prediction

Ex : SimCLR, DINO, CLIP

Popular in computer vision.

Which approach to choose for pretraining : Reconstruction or Joint Embedding ?

Controlled Setting

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

Core Features (k first components)

Data Noise (d-k last components)

Data Augmentation.

$$\tau(\mathbf{x}) = \mathbf{x} + \boldsymbol{\theta} + \alpha\,\boldsymbol{\gamma},\ \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}),\ \boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$$

Other Noise

Data Noise

Aug. / Data Noise Alignment

Linear Models. For tractability $f_{\mathbf{V}} : \mathbf{x} \mapsto \mathbf{V}\mathbf{x}$

Weight Matrix

We consider these 3 problems:

$\mathcal{T}$: data augmentation distribution

Supervised Learning

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

Joint-Embedding (SSL)

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

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

Reconstruction (SSL)

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

We consider these 3 problems:

$\mathcal{T}$: data augmentation distribution

Supervised Learning

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

Joint-Embedding (SSL)

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

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

Reconstruction (SSL)

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

What data augmentation is needed to recover optimal performances ?

1 - Supervised Learning : Consistency Regardless of Augmentations

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

Labels

1 - Supervised Learning : Consistency Regardless of Augmentations

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

Labels

Equivalence with Ridge regularization

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

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

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

1 - Supervised Learning : Consistency Regardless of Augmentations

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

Labels

Proposition

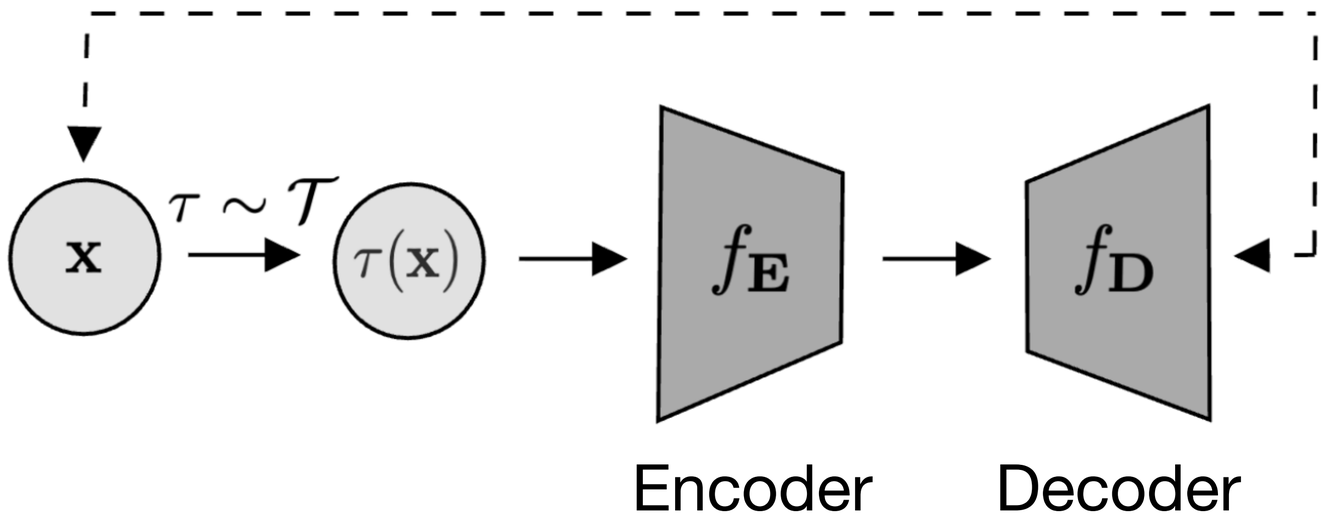
Let $\mathbf{V}^\star$ (resp. $\widetilde{\mathbf{V}}^\star$) solve the Supervised Learning problem with data augmentation $\mathcal{T}(\alpha)$ for the clean data $\mathbf{X}$ (resp. the corrupted data $\widetilde{\mathbf{X}}$). Then:

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

holds almost surely as $n \to +\infty$ (infinite samples) for any alignment α.

2 - Reconstruction

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



2 - Reconstruction

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

Let $\overline{\mathbf{x}}_i = \mathbb{E}_{\tau \sim \mathcal{T}}[\tau(\mathbf{x}_i)]$, $\overline{\mathbf{X}} = (\overline{\mathbf{x}}_1, \ldots, \overline{\mathbf{x}}_n)^\top$. Consider the SVD

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

where $\mathbf{R}, \mathbf{P} \in \mathbb{R}^{d \times d}$ are orthogonal and $\boldsymbol{\Phi} = \operatorname{diag}(\phi_1, \ldots, \phi_d)$ with $\phi_1 \geq \cdots \geq \phi_d \geq 0$. Then solutions take the form

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

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

2 - Reconstruction : Requires Minimal Alignment

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

Proposition

Let $\mathbf{E}^\star$ (resp. $\widetilde{\mathbf{E}}^\star$) be the optimal encoder of the Reconstruction Problem for the clean data $\mathbf{X}$ (resp. the corrupted data $\widetilde{\mathbf{X}}$). The following limit:

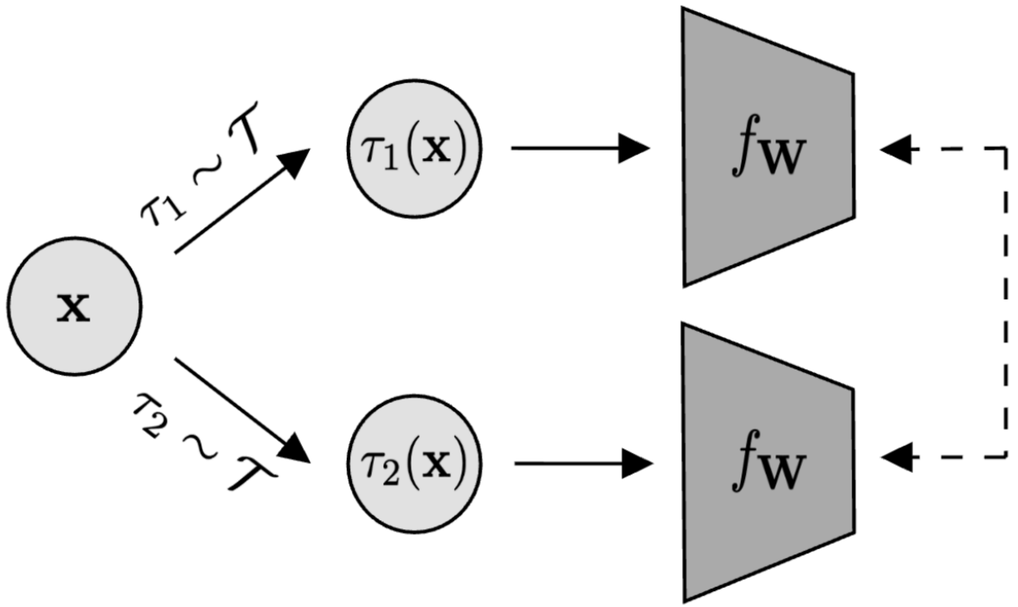
$$\widetilde{\mathbf{E}}^\star \xrightarrow[\text{a.s.}]{} \mathbf{E}^\star$$

holds as $n \to +\infty$ (infinite samples), if and only if the alignment satisfies $\alpha > \alpha^{\mathrm{RC}}$.

3 - Joint-Embedding

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

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



Encoder

3 - Joint-Embedding

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

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

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

where $\mathbf{\Omega} = \operatorname{diag}(\omega_1, \dots, \omega_d)$ with $\omega_1 \geq \cdots \geq \omega_d$. Solutions take the form:

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

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

3 - Joint-Embedding : Also Requires Minimal Alignment

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

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

Let $\mathbf{W}^\star$ (resp. $\widetilde{\mathbf{W}}^\star$) solve the Joint-Embedding Problem for the clean data $\mathbf{X}$ (resp. the corrupted data $\widetilde{\mathbf{X}}$). The following limit:

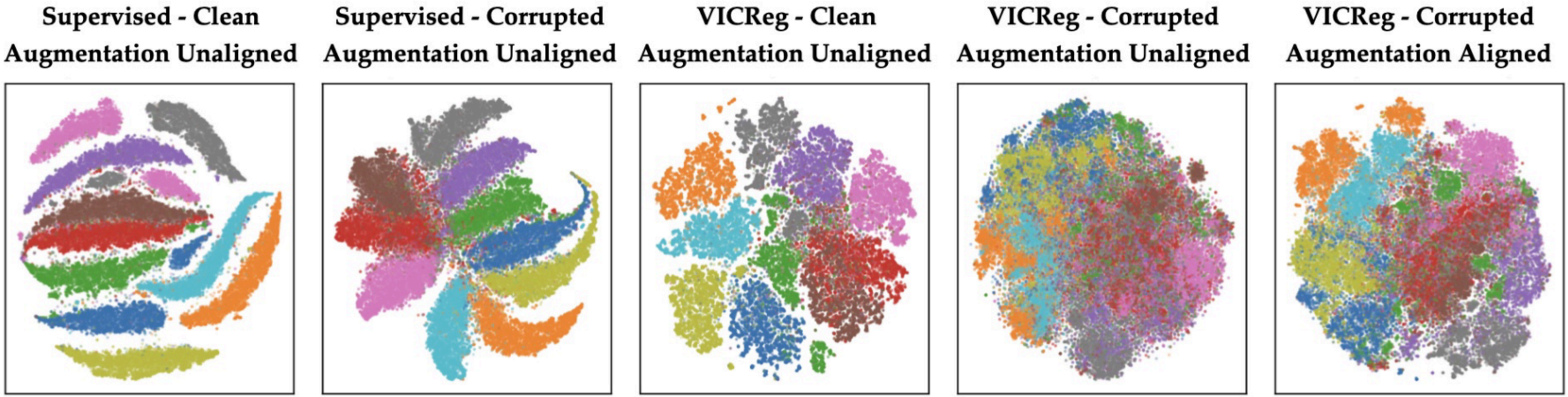
$$\widetilde{\mathbf{W}}^\star \xrightarrow[\text{a.s.}]{} \mathbf{W}^\star$$

holds as $n \to +\infty$ (infinite samples), if and only if the alignment satisfies $\alpha > \alpha^{\text{JE}}$.

Supervised vs SSL

Key Result

Unlike supervised learning, both SSL paradigms require a minimal alignment between augmentations and irrelevant features to achieve asymptotic optimality with increasing sample size.



Joint-Embedding or Reconstruction ?
It depends on the noise level !

Corollary

Let κ_i be the singular values of $\boldsymbol{X}$ (data features).

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

$$\alpha_{\mathrm{JE}}^2 \coloneqq \max_{i \in [\![k+1:d]\!]} \left(\frac{1-\delta}{\delta} - \frac{\lambda_i^{\boldsymbol{\Theta}}}{\lambda_i^{\boldsymbol{\Gamma}}} \right), \quad \text{where} \quad \delta = \min_{i \in [\![k]\!]} \frac{\frac{1}{n}\kappa_i^2}{\frac{1}{n}\kappa_i^2 + \lambda_i^{\boldsymbol{\Theta}}}.$$

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

Joint-Embedding or Reconstruction ?
It depends on the noise level !

Key Result

- If the **data features dominate** in magnitude :

$$\alpha^{\mathrm{JE}} > \alpha^{\mathrm{RC}} \quad \rightarrow \quad \textbf{Reconstruction is preferable}$$

- If the **noise features dominate** in magnitude :

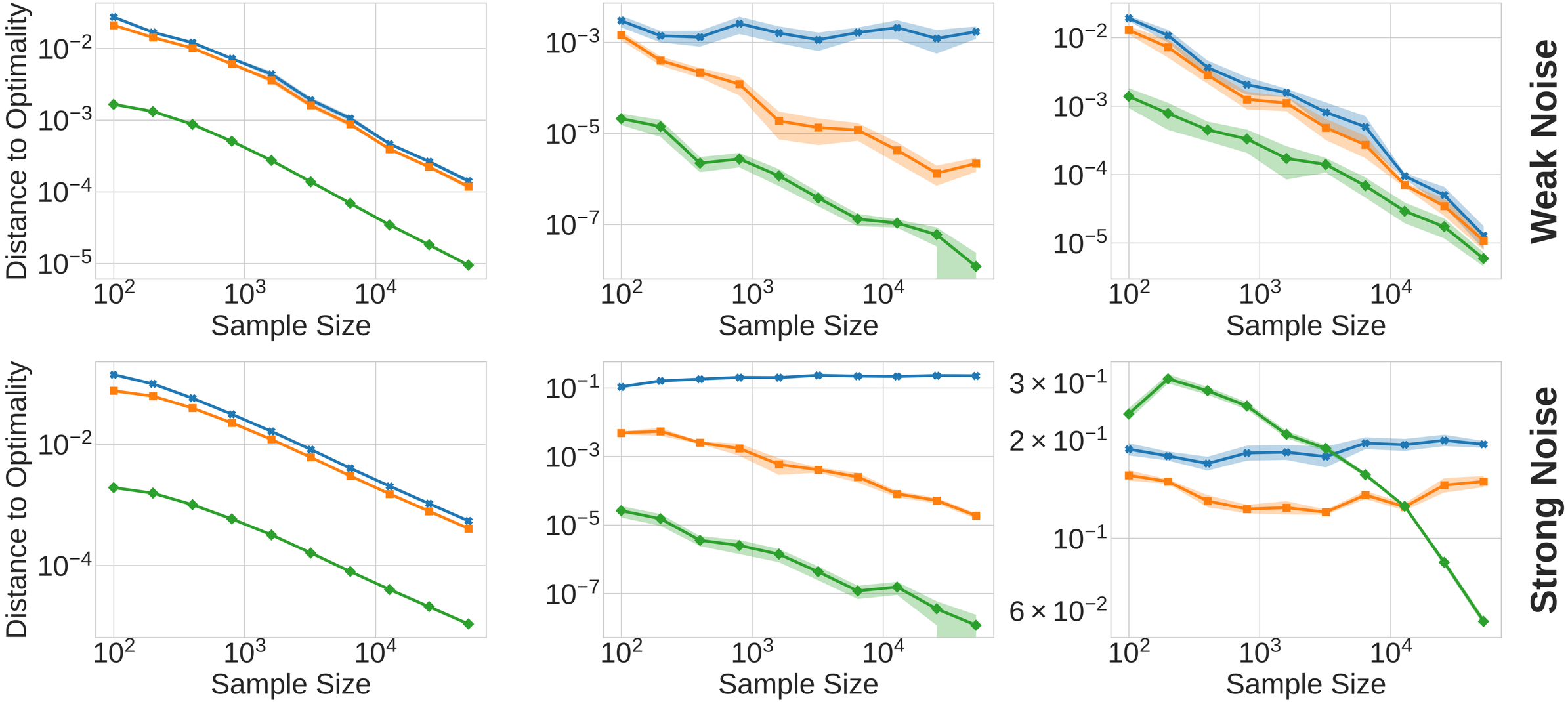
$$\alpha^{\mathrm{JE}} < \alpha^{\mathrm{RC}} \quad \rightarrow \quad \textbf{Joint-embedding is preferable}$$

Simulations with MNIST

Supervised Learning

Joint-Embedding

Reconstruction



Experiments on ImageNet

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 (%)
JE	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
JE	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
RC	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

As we introduce increasingly strong noise into the data, joint-embedding methods remain more robust : their performance declines less than that of reconstruction methods.

Conclusion

- **Favor reconstruction** (autoencoders, language modeling, etc.) when input features are semantically rich and the highest-magnitude components align well with downstream tasks.

  Example: NLP, pre-processed data

- **Favor joint-embedding** (SimCLR, CLIP, etc.) when input features are dominated by high-magnitude irrelevant components.

  Example : raw sensorial recordings of the physical world

Thank you !