INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2026IPPAX057

Thèse de doctorat

ÉCOLE POLYTECHNIQUE

IP PARIS

Capturing the Geometry of Self-Supervised Representations

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat: Informatique

Thèse présentée et soutenue à Palaiseau, le 21/09/2026, par

JULIE MORDACQ

Composition du Jury :

Julien Tierny Directeur de Recherche, CNRS, Sorbonne Université (LIP6, APR)	Président
Yuki M. Asano Full Professor, Université de Technologie de Nuremberg (FunAi Lab)	Rapporteur (Absent)
Bastian Grossenbacher Rieck Full Professor, Université de Fribourg (AIDOS Lab)	Rapporteur
Oriane Siméoni Research Scientist, Meta	Examinatrice
Raoul de Charette Directeur de Recherche, Inria Paris (Astra)	Examinateur
Steve Oudot Directeur de Recherche, Inria Saclay (GeomeriX)	Directeur de thèse
Vicky Kalogeiton Full Professor, École Polytechnique (Vista)	Co-encadrante de thèse
Guillaume Vimont Docteur, Ministère des Armées	Invité

‘Outrenoir’: noir qui, cessant de l’être,
devient émetteur de clarté, de lumière
secrète.

Pierre Soulages

Acknowledgements

A PhD may be a tortuous journey. Yet, I carry with me a plethora of vibrant and joyful moments, shared with wonderful people, to look back on.

I would like to thank my advisors, Steve and Vicky. Thank you for letting me embark upon this adventure, for both trusting and challenging my ideas. Thank you for teaching me that I should always believe in myself, no matter the deadline. Vicky, thank you for reaching out when it mattered. What I admire most is your determination. Steve, thank you for every discussion we shared, whether it be about Wasserstein distances, dimensions or T-Rex.

Nothing would have been possible without Guillaume Vimont. Thank you for this opportunity, and for your support throughout.

Thank you to the members of the jury, Bastian Grossenbacher-Rieck, Yuki Asano, Oriane Siméoni, Julien Tierny and Raoul de Charette, for agreeing to review our work and for their valuable feedback.

Thank you to all my colleagues at GeomeriX and VISTA.
I want to extend special thanks to David. Thank you for your *joie de vivre*, your support, both technical and emotional, and, above all, your love of yoghurt breaks.
I also extend special thanks to Arijit, Emery, Lefteris, Lucas, Nicolas, Nissim, Paul, Ridouane, Robin, Xi and Yuanzhi. Thank you for all the happy moments we shared both during and after work: conferences with nice dinners and karaoke, the Grand Breguet, PMUs and bars, and simple coffee breaks in the atrium. Thank you for your support and for always showing interest in my work (and my colour palettes). I am thankful to have shared this journey with you.

Merci à mes ami·es d'avoir toujours été là, pendant les heures les plus heureuses comme les plus dures. Merci pour les fleurs, les cartes, les messages et tous les moments qu'on a pu partager, quels qu'ils soient, où qu'ils soient.
Julia, Mélissa, grandir et évoluer à vos côtés a été et continue d'être un privilège.
Chloé, ta bienveillance, ton souci d'autrui, ton courage et tes convictions m'inspirent.
Alix, Princina, que ce soit d'un côté ou de l'autre de l'Atlantique, merci pour votre

soutien. J'admire votre force de caractère et votre gentillesse. À tous les verres au coin de la rue qu'il nous reste !
Hélène, merci de m'avoir appris à écouter l'indicible.
Merci également à Aurélien, Estelle, Loan, Antoine, Max, Margaux, Aloïs et le D.

Merci à ma famille.
Merci à mon grand-père, pour qui le savoir était le plus précieux des biens.
Merci à ma grand-mère, qui, par son abnégation et son amour, est une source d'inspiration au quotidien.
Merci à mon frère, pour son soutien, toujours ancré dans la réalité.
Enfin, merci à mes parents, pour leur soutien et leur amour, inconditionnels, pendant les éclaircies, les tempêtes et tout ce qu'il y a entre les deux.

Contents

Notations vii

1 Introduction 1
- 1.1 Visual Representations: Towards Self-Supervised Learning . . . 2
  - 1.1.1 Handcrafted representations . . . 2
  - 1.1.2 Deep supervised representations . . . 2
  - 1.1.3 Self-supervised representations . . . 5
- 1.2 Geometric Perspectives on Self-Supervised Learning . . . 6
  - 1.2.1 The geometry of representation spaces . . . 6
  - 1.2.2 Why capture the geometry of self-supervised representations? 8
- 1.3 Positioning . . . 9
  - 1.3.1 Shaping representations against dimensional collapse . . . 9
  - 1.3.2 Evaluating representations beyond downstream performance 10
- 1.4 Contributions and Outline . . . 10

2 Background 12
- 2.1 The Early Era of Self-Supervised Learning . . . 12
  - 2.1.1 Handcrafted pretext tasks . . . 13
  - 2.1.2 Generating images . . . 13
  - 2.1.3 Multi-view invariance . . . 14
- 2.2 Joint-Embedding Self-Supervised Approaches . . . 16
  - 2.2.1 Contrastive methods . . . 19
  - 2.2.2 Clustering methods . . . 20
  - 2.2.3 Self-distillation methods . . . 21
  - 2.2.4 Redundancy reduction methods . . . 22
- 2.3 Beyond Joint-Embedding Approaches . . . 25
  - 2.3.1 Masked-Image Modeling . . . 25
  - 2.3.2 Multimodal joint-embedding methods . . . 26
  - 2.3.3 Scaling SSL . . . 27
- 2.4 Minimum Spanning Trees . . . 28
  - 2.4.1 MST and Dimension Estimation . . . 29

2.4.2 Connections to Topological Data Analysis and Differentiability 31
2.4.3 MST Computation . . . 34

3 Minimum Spanning Tree Regularization for Self-Supervised Learning 35
3.1 Introduction . . . 36
3.2 T-REG: Minimum Spanning Tree based Regularization . . . 38
3.2.1 Theoretical analysis . . . 40
3.2.2 Uniformity properties . . . 45
3.2.3 Empirical evaluation . . . 46
3.3 T-REGS: T-REG for Self-supervised learning . . . 48
3.3.1 Evaluation on standard SSL benchmark . . . 49
3.4 Ablation Studies and Analyses of T-REGS . . . 50
3.4.1 Loss coefficients . . . 51
3.4.2 Computational overhead of T-REGS . . . 52
3.4.3 Projector architecture . . . 53
3.4.4 Batch size . . . 54
3.4.5 Normalization . . . 54
3.4.6 Study of the embeddings . . . 55
3.4.7 Study of redundancy-reduction methods . . . 55
3.5 Implementation Details . . . 57
3.6 Computational Cost . . . 57
3.7 Discussion . . . 58
3.8 Conclusion . . . 59

4 Anchored Multimodal Representations for Physiological Change Detection under Missing Modalities 60
4.1 Introduction . . . 61
4.2 Background . . . 62
4.2.1 Handling missing modalities . . . 62
4.2.2 Multimodal transformers . . . 63
4.3 Motivation: T-REGS for Multimodal Alignment . . . 63
4.3.1 The modality gap . . . 64
4.3.2 T-REGS for CLIP fine-tuning . . . 65
4.4 Method . . . 66
4.4.1 Anchoring modality-specific encoders . . . 67
4.4.2 Masked multimodal transformer . . . 68
4.5 Experiments . . . 69
4.5.1 Dataset . . . 69
4.5.2 Metrics . . . 70
4.5.3 Comparison to the state of the art . . . 70

4.5.4 Ablations . . . 70
4.5.5 Implementation details . . . 73
4.6 Computational Cost . . . 73
4.7 Discussion . . . 73
4.8 Conclusion . . . 74

5 Assessing Self-Supervised Learning Representations via Intrinsic Dimension 76
5.1 Introduction . . . 77
5.2 Background: Label-Free Geometric Analysis of Representations . . . 79
5.2.1 Representation spectrum . . . 79
5.2.2 Intrinsic dimension . . . 80
5.3 Generalization and Dimension Estimation . . . 80
5.3.1 Theoretical connection to generalization . . . 80
5.3.2 Intrinsic dimension estimators . . . 81
5.4 IdEst: $\dim_{\text{MST}}$ for Unsupervised Assessment of Self-Supervised methods . . . 83
5.4.1 Implementation of IdEst . . . 84
5.4.2 IdEst and accuracy of foundation models . . . 84
5.4.3 Training dynamics: offline and online probing . . . 87
5.4.4 Label-free metric for model selection . . . 90
5.4.5 Computational complexity and scalability . . . 92
5.4.6 Situating IdEst among collapse-tailored JE-SSL metrics . . . 93
5.5 Implementation Details . . . 97
5.5.1 Image classification training details . . . 97
5.5.2 Models studied in Section 5.4.2 . . . 98
5.5.3 Models trained in Section 5.4.4 . . . 98
5.6 Computational Cost . . . 100
5.7 Discussion . . . 100
5.8 Conclusion . . . 101

6 Conclusion 102
6.1 Summary . . . 102
6.2 Discussion . . . 103
6.2.1 On the role of minimum spanning trees . . . 103
6.2.2 On the interplay between identifying and enforcing geometric properties . . . 104
6.3 Perspectives . . . 105

Bibliography 107

List of Figures 121

List of Tables 126

A Résumé étendu 131

Abstract/Résumé 133

Notations

Common Notations

Dataset

x	sample point (e.g., an image)
$\mathcal{D}$	dataset
x_i	i-th sample of $\mathcal{D}$
x_i^m	i-th sample of modality m in $\mathcal{D}$
N	number of samples in $\mathcal{D}$
B	batch size
M	number of modalities

Encoders & Projections

$f(\cdot)$	encoder (maps high-dimensional input data into low-dimensional latent representation)
$f_m(\cdot)$	encoder for modality m
$h(\cdot)$	projection head
$h_m(\cdot)$	projection head for modality m
d	embedding dimension
z, z_i	embedding (encoder and projection head output, $g \circ f$)
$Z = \{z_i\}_{i=1}^n$	set of embeddings
z_i^m	projected representations of anchor and modality m
$Z^m = \{z_i^m\}_{i=1}^n$	set of projected representations for modality m
$\mathbb{S}^{d-1}$	unit hypersphere in $\mathbb{R}^d$

Minimum Spanning Tree

$\text{MST}(X)$	Minimum Spanning Tree of X
$L(\text{MST}(X))$	length of the MST of X

Chapter 1

Introduction

What do computers see? While the visual world unfolds along three spatial dimensions, encoding it digitally requires discretizing continuous visual signals into numerical values. Images are digitized by combining individual color picture elements, *pixels*, each formed by combining red, green, and blue intensities. Drawing on these pixels, the field of *computer vision* aims to enable machines to understand and interpret visual information, e.g., recognizing patterns, tracking motion, or locating objects. However, the resulting pixel space vastly exceeds three dimensions. The complexity of the pixel space scales sharply with resolution: while a low-resolution image spans thousands of pixels, a high-resolution counterpart can exceed $1,000,000$ dimensions. This high dimensionality renders downstream tasks such as classification computationally intractable without first extracting meaningful structure.

Consequently, since its early days, computer vision has revolved around a fundamental question: *How should images be represented?* Rooted in the Latin *repraesentare*, *to make present*, representation is the act of surfacing what is latent; bringing forward only what matters. In computer vision, this question of *what matters* has driven decades of research. Visual data must be distilled into compact, structured forms, known as representations, or features; each reflecting a different answer to what should be extracted from an image, and why. From hand-crafted algorithms, designed to capture edges, textures, and local geometry, to general-purpose features learned through deep neural networks, the evolution of visual representations has shaped every milestone in the field.

Recently, *self-supervised learning*, a paradigm in which deep neural networks are

trained without labelled data, has demonstrated remarkable success across a wide range of downstream tasks, opening new opportunities for representation learning. Yet understanding what makes these representations effective remains an open challenge. Pivotal questions, such as what geometric properties are desirable or how to characterize them, are still unanswered. In this manuscript, we aim to advance the understanding of self-supervised representations by effectively capturing geometric properties.

To motivate our work, we begin with a historical overview of visual representations (Section 1.1), before situating our work within this landscape (Sections 1.2 and 1.3), and finally introducing our contributions (Section 1.4).

1.1 Visual Representations: Towards Self-Supervised Learning

In this section, we trace the evolution of visual representations, from manually engineered features through deep supervised learning, to self-supervised learning (SSL), a paradigm that learns without labels and forms the foundation of our work. Figure 1.1 presents an overview of the three main paradigms.

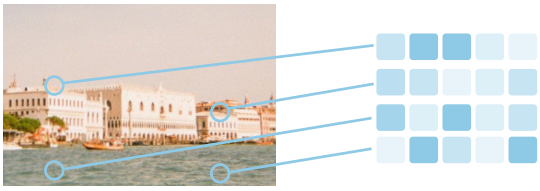
1.1.1 Handcrafted representations

Until the mid to late 2000s, approaches sought to surface distinctive attributes of images through hand-engineered algorithms tailored to specific tasks. Researchers meticulously designed these algorithms to capture particular visual characteristics, edges, corners, gradients, and beyond [22, 126, 44]. For instance, SIFT [126] describes local gradient distributions around keypoints, while HOG [44] encodes the distribution of edge orientations across image regions. While such hand-crafted representations proved effective for constrained settings, they lacked the flexibility to scale to the diversity and complexity of real-world visual data.

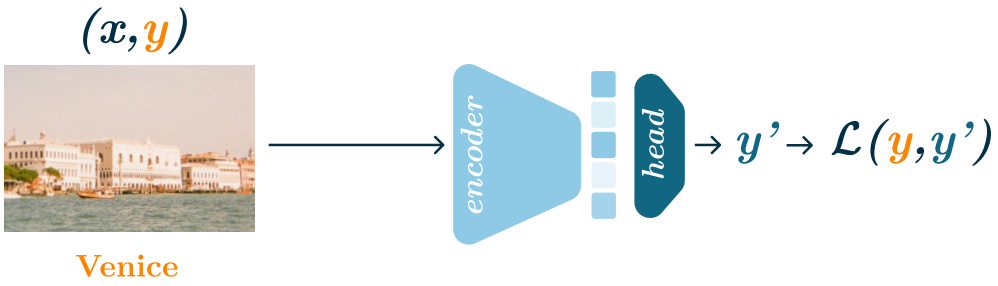
1.1.2 Deep supervised representations

The rise of deep learning marked a decisive paradigm shift. Since the mid-2010s, deep neural networks have become the cornerstone of most computer vision applications.

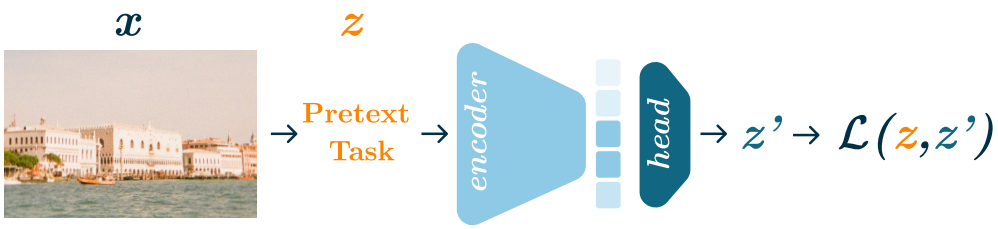
Deep learning. Neural networks lie at the core of deep learning. The perceptron [132, 155] is a single computational unit that maps a learned linear combination of inputs through a nonlinearity. Despite its simplicity, this architecture



(a) Handcrafted local descriptors
(Section 1.1.1)



(b) Deep supervised representations
(Section 1.1.2)



(c) Self-supervised representations
(Section 1.1.3)

Figure 1.1: **Evolution of Visual Representation Learning.** Three paradigms for extracting representations from images: *(a)* handcrafted descriptors encode local gradient statistics around salient keypoints as fixed-length vectors (e.g., SIFT [126]); with *Deep Learning*, representations are learned end-to-end via backpropagation, where a head is trained on top of the encoder: *(b) deep supervised representations*: the head predicts the class label; *(c) self-supervised representations*: a pretext task provides a pseudo-label, enabling training without manual annotations. In both cases the head is discarded and the encoder and its representations are retained.

forms the fundamental building block of modern deep learning. Stacking several layers yields a *multilayer perceptron* (MLP), and modern architectures can largely be understood as MLPs equipped with different structural priors. *Convolutional Neural Networks* (CNNs) [65, 87, 111] impose a translation-invariant prior, convolving spatially shared filters across the input rather than applying weights globally. Transformers [196] impose no such locality constraint, instead computing contextual representations via scaled dot-product *attention*.

Ultimately, the objective is to find parameters θ for $f_\theta : \mathcal{X} \to \mathcal{Y}$ that generalize to unseen data, mapping an input space $\mathcal{X}$ (e.g., images) to an output space $\mathcal{Y}$ (e.g., labels). Parameters are optimized via backpropagation [157] using a loss function $\mathcal{L}$ to minimize prediction error. Crucially, in 1989, Hornik, Stinchcombe, and White [88] and Cybenko [42] proved a fundamental result: the *universal approximation theorem*. It guarantees that a sufficiently wide network can approximate any real-valued continuous function with compact support. At the time, compute resources were insufficient to realize this potential; the theorem nonetheless provided early theoretical foundation for the deep learning revolution that followed.

Supervised representation learning. With deep learning, the field moved away from handcrafted representations toward hierarchical ones learned end-to-end from data [107, 204, 154, 27]. Networks can be trained on large amounts of data via *supervised learning*, which requires both samples and their corresponding labels. Representations learned on a large supervised dataset can be reused to solve other tasks on other datasets. This paradigm, known as *transfer learning*, decomposes into two stages: *(i) pre-training*, in which the network is trained on a large source dataset; and *(ii) fine-tuning*, in which the learned representations are adapted to solve downstream tasks. For instance, pre-training on ImageNet [47] has been shown to consistently outperform random initialization on downstream tasks [163, 142]. This transferability catalyzed the early development of vision foundation models. Early research focused on refining convolutional architectures [170, 179, 84], and their modern variants [210, 182, 102], before attention-based architectures ultimately emerged as the dominant paradigm [52, 46].

Limitations of using labels. *Labels are expensive.* Supervised pre-training hinges on large-scale labeled data, a requirement that quickly becomes prohibitively expensive. ImageNet [47], the most prominent vision benchmark, comprises 1.2 million images across 1,000 categories, yet assembling it required 49,000 workers to sift through 160 million candidates, an effort estimated at 19 human-years of labor [118]. Yet this dataset represents only a negligible fraction of the images produced and uploaded online daily; labeling even a small portion of that stream at the same quality would be entirely intractable.

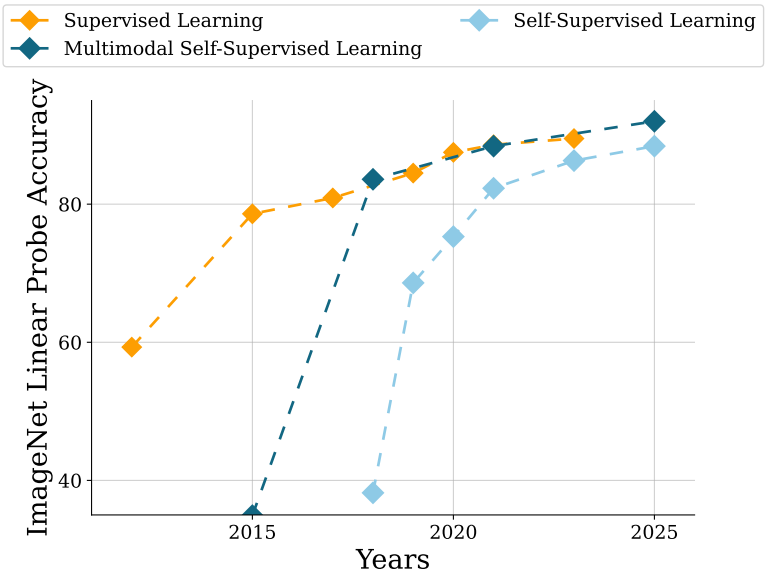


Figure 1.2: **Evolution of linear probing accuracy on ImageNet over time.** We compare fully Supervised Learning, Self-Supervised Learning (SSL) and Multimodal SSL. Although SSL emerged later, it improved rapidly and now approaches the saturated top-1 accuracy regime typical of recent Supervised Learning and Multimodal SSL architectures. Figure adapted from Siméoni et al. [168].

Labels are ambiguous. Annotation is inherently ill-defined. Consider segmenting objects in an image, should a building's windows be delineated individually? Its individual panes of glass? There is no universal answer, and every such decision ties the resulting model to a specific task. A model trained to segment buildings cannot straightforwardly generalize to segmenting windows without an entirely new dataset.

Self-supervised learning (SSL) aims to circumvent these bottlenecks by deriving supervisory signals directly from the data itself, bypassing the need for manual annotation.

1.1.3 Self-supervised representations

The goal of self-supervised representation learning is to automatically uncover some structure and understanding from images without relying on labelled data. Driven by the promise of transfer learning, recent advances in self-supervised learning (SSL), spanning both unimodal and multimodal training, have yielded representations that rival or even surpass their supervised counterparts (Figure 1.2).

The central challenge of self-supervised learning lies in designing 'pretext tasks'

that force a network to capture deep structural semantics rather than low-level shortcuts. Historically, this problem has been approached through two parallel strategies. The first relies on generative reconstruction or inpainting, tasking the network with predicting masked or corrupted image regions, a visual analogue to the masked token prediction that revolutionized natural language processing [82]. The second enforces invariance to transformations, relying on the intuition that an image's underlying semantic meaning remains unchanged despite variations in crop, color, or blur (Figure 2.2). This latter paradigm, known as Joint-Embedding SSL (JE-SSL), forms the backbone of modern representation learning [26, 12] where two networks are trained to produce similar embeddings for different views of the same input. An illustration is provided in Figure 2.3. Increasingly, however, the boundary between these paradigms is blurring. Recent hybrid methods bridge the two by combining the fine-grained local reconstructions of generative objectives with the global semantic consistency of joint-embedding architectures, while scaling model capacity and refining dataset curation [143, 168, 197].

Overall, these methods demonstrate the ability to learn strong features and achieve high performance on multiple benchmarks, marking a shift from task-specific models to general-purpose visual backbones.

1.2 Geometric Perspectives on Self-Supervised Learning

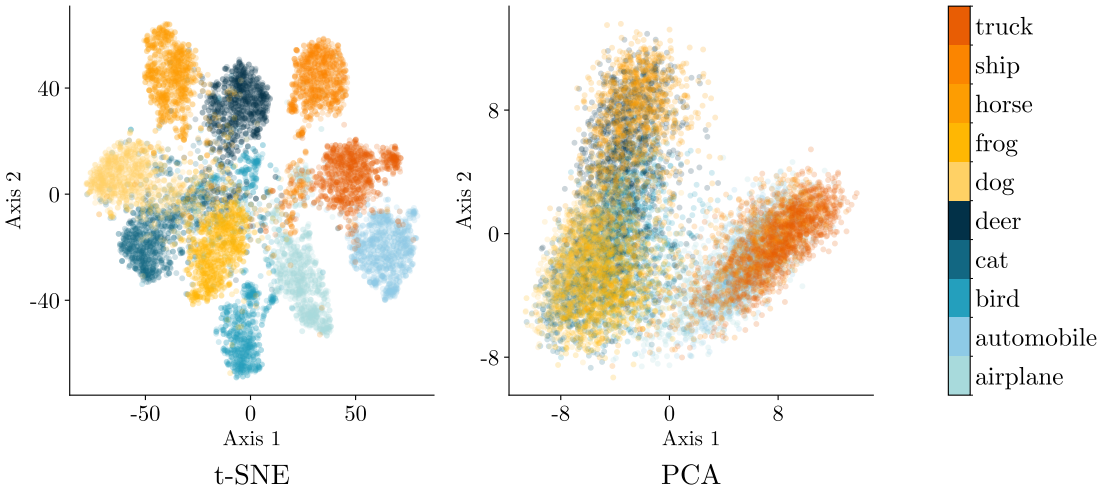
The success of self-supervised learning raises a deeper question: *what makes* these representations transfer so well across tasks and domains?

1.2.1 The geometry of representation spaces

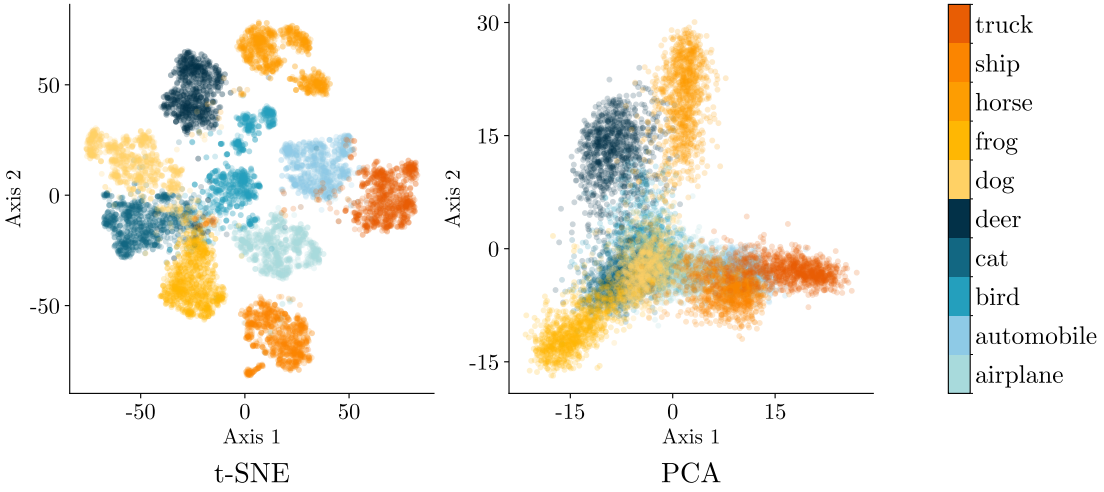
Scaling and improved pretext tasks have driven empirical progress, but offer limited insight into the internal structure underlying their effectiveness. A key insight is that the answer lies, at least partly, in the *geometry* of the representation space [5, 202, 67, 68], that is, how learned embeddings are distributed and structured in the high-dimensional space they inhabit.

Intuitively, a representation can be considered *'useful'* insofar as it captures the structure of the data, i.e., if the geometry of the embedding space reflects the underlying semantic organization of the data. Yet different models can embed the same data in very different ways.

To illustrate this, Figure 1.3 visualizes the representations produced by two SSL models on the standard benchmark CIFAR-10. We project them using two com-



(a) CLIP (ViT-B)



(b) DINOv3 (ViT-B)

Figure 1.3: **SSL models trained with different objectives produce geometrically distinct representation spaces.** CIFAR-10 representations from *(a)* CLIP ViT-B and *(b)* DINOv3 ViT-B, projected via t-SNE (left) and PCA (right). The two models structure the same frozen representations very differently: DINOv3 yields tight, well-separated clusters, while CLIP produces a more diffuse, overlapping arrangement.

plementary dimensionality-reduction techniques that each highlight distinct structural properties: t-SNE [193], which preserves local neighbourhood structure and exposes cluster organization, and PCA, which retains the directions of greatest global variance. The comparison reveals that these two models organize the representation space in markedly different ways, and that the two projection techniques surface complementary aspects of that structure.

Representation geometry is therefore not uniform, and, as we argue next, not arbitrary either.

1.2.2 Why capture the geometry of self-supervised representations?

Crucially, capturing the geometry of representations goes beyond description. It carries direct theoretical and practical consequences.

Understanding what networks learn. Training loss measures how well the network optimizes its objective rather than representational quality. The two can diverge in ways that are difficult to diagnose. This gap is especially acute in SSL, where the pretext objective bears no direct relation to downstream performance. Consider, for instance, objectives that enforce invariance between representations of two augmented views of the same image: a low loss indicates that the model has learned to align these views, but says nothing about how discriminative or transferable the resulting representations are. Geometric analysis can provide a complementary lens: it characterizes the internal structure a network is learning, independently of any training objective.

Diagnosing failure modes. Some failure modes are invisible to downstream metrics, making them difficult to detect and address. A canonical example is representational collapse [95], where the encoder maps all inputs to the same point or, in a more subtle variant, confines representations to a low-dimensional subspace, both cases yielding embeddings that carry no discriminative information. Such collapse is a trivial solution to joint-embedding objectives, which enforce agreement between representations of two augmented views of the same image, since a constant encoder satisfies this constraint without any additional mechanism. Yet it can evade detection by downstream metrics: accuracy reveals *how well* a model performs, but not *why* it falls short.

Informing the design of training objectives. Identifying which geometric properties signal failure modes, enables their explicit enforcement as regularizers or auxiliary losses that support the primary objective [202, 63].

Enabling unsupervised model evaluation. In practice, evaluating SSL models typically requires labelled data: hyperparameters, architectures, and training choices are selected by monitoring supervised downstream performance. This dependence on labels undermines one of the core promises of SSL and makes deployment on new, unlabelled datasets costly. If geometric properties of the representation space correlate with downstream performance, they can serve as cheap, label-free proxies for representation quality, enabling model selection and hyperparameter tuning without any annotation.

A rich and modular space of properties to study. Geometric analysis is not monolithic: it spans a broad spectrum of complementary perspectives. Intrinsic dimensionality, isotropy, uniformity, and manifold structure each illuminate distinct facets of what a representation encodes, and a rich variety of methods exists to *capture* them.

Yet, *which* property to leverage, for *what* purpose, and *how* to reliably measure it, all remain open questions.

1.3 Positioning

Geometry provides a principled lens for deepening our understanding of self-supervised representations. This thesis leverages it along two complementary directions: *shaping* representations by enforcing geometric constraints during training to prevent known failure modes, and *evaluating* them through geometric measures to characterize their quality without labels or downstream tasks.

1.3.1 Shaping representations against dimensional collapse

A dominant approach in self-supervised learning is *joint-embedding self-supervised learning* (JE-SSL), in which two networks are trained to produce similar embeddings for different augmented views of the same image, known as positive pairs. Its central challenge is *dimensional collapse*: embeddings tend to span only a low-dimensional subspace, thereby reducing representational diversity.

Several works have analyzed desirable properties of JE-SSL representations to better understand and prevent this collapse [202, 89, 67, 64]. Crucially, Wang and Isola [202] characterized JE-SSL objectives through two complementary properties: *alignment*, which pulls positive pairs together, and *uniformity*, which spreads features over the unit hypersphere, the latter providing a principled mechanism against collapse. Yet Fang et al. [63] showed, leveraging tools from optimal transport, that standard uniformity formulations fail to capture the full geometric spread of embeddings, leaving a gap between the property as defined and the

property as desired. Addressing this gap is a primary objective of this thesis: we investigate a principled regularization approach to enforce a geometrically faithful notion of uniformity and prevent dimensional collapse.

1.3.2 Evaluating representations beyond downstream performance

Beyond designing better objectives, a complementary avenue is to *understand* the underlying geometric structure of SSL representations, independently of the training objective or downstream labels.

Such understanding is especially valuable in the SSL setting, where no supervision signal is available to directly assess representation quality during training. Self-supervised loss values give little indication of downstream transferability: a model can minimize its training objective while learning representations that transfer poorly to unseen tasks. Additionally, standard evaluation protocols, i.e., linear probing or fine-tuning, are computationally expensive, sensitive to hyperparameters, require labelled data, and measure task performance rather than the structure of the representation space. A natural answer is to seek label-free criteria grounded in the intrinsic geometry of the representation manifold. However, existing work has focused primarily on diagnosing known failure modes [5, 67], rather than providing a general account of what makes representations transferable; a geometry-based characterization of representation quality remains to be investigated. Thus, a second objective of this thesis is to provide such a characterization, one that is unsupervised and indicative of downstream transfer performance.

1.4 Contributions and Outline

This thesis advances both of the aforementioned directions, *shaping* and *evaluating* self-supervised representations, through a common geometric tool: *Minimum Spanning Trees* (MST), whose length encodes the intrinsic geometry of a data distribution. This dissertation is organized into four chapters following this introduction.

Chapter 2 reviews the concepts underpinning the thesis. It covers self-supervised learning from its early formulations (Section 2.1) to more recent developments (Sections 2.2 and 2.3), and introduces Minimum Spanning Trees (Section 2.4).

Shaping representations against dimensional collapse (Chapters 3 and 4). Chapter 3 introduces T-REGS, a regularizer that counters dimensional collapse by promoting uniform coverage of the unit hypersphere, using the MST length as a

tractable geometric proxy. We establish theoretical guarantees connecting MST length maximization to entropy maximization and uniformity on compact Riemannian manifolds, and validate T-REGS empirically on standard SSL benchmarks.

Associated publication: T-REGS: Minimum Spanning Tree Regularization for Self-Supervised Learning., *J. Mordacq, D. Loiseaux, V. Kalogeiton, S. Oudot.* In The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS) 2025.

Chapter 4 extends T-REGS to the multimodal joint-embedding setting, where an analogous failure mode arises: *the modality gap*, in which embeddings from different modalities cluster in separate regions of the space. As a concrete application, we address stress detection from wearable signals, where physiological data spans multiple modalities and missing modalities are frequent at test time. We introduce ADAPT, which aligns modality-specific embeddings to a shared space via T-REGS and fuses them through a masked transformer robust to missing modalities.

Associated publication: ADAPT: Multimodal Learning for Detecting Physiological Changes under Missing Modalities, *J. Mordacq, L. Milecki, M. Vakalopoulou, S. Oudot, V. Kalogeiton.* In Medical Imaging with Deep Learning (MIDL) 2024 & Conference on Computer Vision and Pattern Recognition - Workshop (CVPR-W) 2024.

Evaluating representations beyond downstream performances (Chapter 5). Chapter 5 complements the shaping contributions by studying the geometric structure of SSL representations in a label-free setting. In particular, we study the *intrinsic dimension* (ID) of SSL representations, the effective number of degrees of freedom required to represent data in an embedding space. We introduce `IdEst`, an unsupervised criterion that assesses representation quality via the MST-based ID estimator ($\dim_{\text{MST}}$). We show that `IdEst` correlates with downstream performance across a wide range of SSL objectives and architectures, provides insight into the evolution of representations during pre-training, and enables unsupervised hyperparameter selection.

Associated publication: `IdEst`: Assessing Self-Supervised Learning Representations via Intrinsic Dimension, *J. Mordacq, V. Kalogeiton, S. Oudot.* In International Conference on Machine Learning (ICML) 2026.

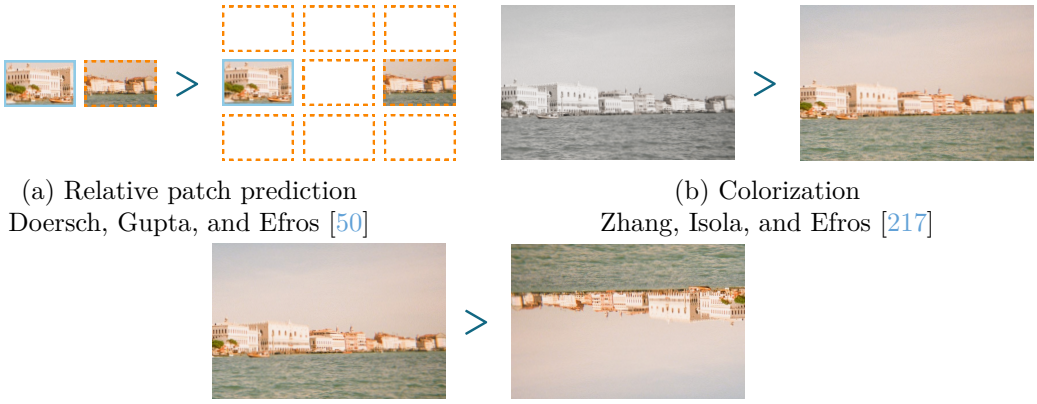
Chapter 2

Background

This chapter introduces the background concepts underpinning the contributions of this thesis. First, it provides an overview of the core concepts in SSL before 2020 (Section 2.1). Subsequently, it transitions into recent deep learning-based SSL methods, focusing on Joint-Embedding Self-supervised learning (JE-SSL) (Section 2.2). It then looks beyond JE-SSL, covering masked-image modeling, multimodal learning, and large-scale pretraining (Section 2.3). Finally, Section 2.4 provides relevant background on a core tool of this thesis: Minimum Spanning Trees (MSTs).

2.1 The Early Era of Self-Supervised Learning

Contemporary SSL methods are built upon insights gained from early research. This section provides a brief overview of the core paradigms that defined self-supervised learning prior to 2020. While many of these specific methods have been superseded by modern architectures, their underlying principles remain highly relevant. For instance, the concept of enforcing consistency between representations of an image under small transformations (ensuring they 'agree'), forms the bedrock of current SSL frameworks and dates back to the early 1990s [14]. In this section, we will focus on the early era of SSL through three main lenses: *the design of pretext tasks from data*, *generative modelling* and *multi-view invariance*.



(a) Relative patch prediction
Doersch, Gupta, and Efros [50]

(b) Colorization
Zhang, Isola, and Efros [217]

(c) Rotation
Gidaris, Singh, and Komodakis [71]

Figure 2.1: **Self-supervised pretext tasks: Learning from self-generated signals by predicting withheld or distorted input properties.** Surrogate tasks are constructed by extracting supervisory signals directly from the unlabeled data itself, such as: *(a)* predicting the position of a target patch relative to a reference patch, *(b)* predicting pixel colors from a grayscale image and *(c)* rotating an image and predicting the corresponding angle.

2.1.1 Handcrafted pretext tasks

Initial explorations into self-supervised learning often involved designing pretext tasks centered around predicting properties of the input data that were intentionally withheld or distorted. For instance, context prediction tasks trained models to predict the relative spatial location of image patches (Figure 2.1a) [50, 139]. Given a central patch, the model predicts which of several candidate patches correctly belongs to its right, left, top, or bottom. Another popular approach was colorization (Figure 2.1b) [217, 110], where the model was given a grayscale version of an image and tasked with predicting its original colors (often represented in a quantized color space). This forced the model to understand object semantics to infer plausible colors (e.g., the sea is typically blue). Similarly, predicting the angle of a randomly applied image rotation (Figure 2.1c) [71] encouraged the model to learn about object orientation and canonical views.

2.1.2 Generating images

A straightforward unsupervised task involves reconstructing input data or generating new, probable samples. We refer to this class of methods as generative approaches, as they operate within the image space and directly model the data-generating distribution. A quintessential example is the *autoencoder*. Originally



Figure 2.2: **Semantic invariance under data augmentations.** Various random transforms applied to the original image *(left)* create distorted views that preserve the underlying semantic content [54].

introduced by Hinton and Salakhutdinov [85], autoencoders aim to learn a mapping that encodes an image into a latent feature space and subsequently decodes it back to the original input. Ranzato et al. [152] extended this by forcing the learned representations to be sparse, while Vincent et al. [198] introduced Denoising Autoencoders, which train the network on corrupted data to improve robustness against altered inputs. Alternatively, *Generative Adversarial Networks* (GANs) [74, 174] simultaneously train two networks on competing objectives. The generator learns to synthesize images capable of fooling a discriminator, which itself learns to distinguish between 'fake' and 'real' samples. Finally, these two paradigms, GANs and autoencoder networks, can be integrated into unified frameworks [51, 56].

2.1.3 Multi-view invariance

> "These signals are generated by using the assumption that different parts of the perceptual input have common causes in the external world. Small modules that look at separate but related parts of the perceptual input discover these common causes by striving to produce outputs that agree with each other."
>
> — Becker and Hinton [14]

In the early 1990s emerged the concept of enforcing consistency between image representations under different views [167, 14]. Several years later, the use of diverse data augmentations to generate semantic-preserving views, concurrently with improvements in how agreement is enforced, has enabled rapid progress in the quality of representations for downstream tasks.

Exemplar-CNN [54] introduced a first pivotal step forward by leveraging these data augmentations to learn representations that are invariant to stochastic variations of the input sample. They relied on the observation that the perception of a scene does not usually affect its semantic content; for instance, the meaning or purpose of an image of Pizza San Marco in Venice is typically preserved regardless of its colorization, rotation, or angle (Figure 2.2). However, invariance alone can be

trivially achieved by collapsing all representations to a single constant vector. To prevent this, the problem was formulated as an 'N-way' classification problem, where N is the size of the dataset. In other words, each image acts as a distinct class, and a network is trained to discriminate between them regardless of applied augmentations. However, explicitly learning a classifier to discriminate between all images scales poorly (e.g., ImageNet has $N = 1,281,167$ images).

To overcome this limitation, subsequent work focused on comparing instances without explicit classification. Several approaches proposed using Noise-Contrastive Estimation (NCE) [79] to directly compare instances. Wu et al. [207] introduced an NCE-based loss that avoided the need for N explicit class labels. Denoting by $z_i \in \mathbb{R}^d$ the representation of sample i, $\cos(\cdot,\cdot)$ the cosine similarity, $\tau > 0$ a temperature, and $Z^{(t-1)}$ the memory bank from the previous epoch:

$$\mathcal{L}_{\text{NCE}} = -\sum_{n=1}^{N} \log\left(\frac{e^{\cos(z_i, z_i^{t-1})/\tau}}{\sum_{k=1}^{N} e^{\cos(z_i, z_k)/\tau}}\right) + \beta||Z - Z^{(t-1)}||_F^2, \tag{2.1}$$

This approach, also termed *non-parametric softmax*, made three key contributions: it introduced the temperature parameter τ to control the sharpness of the distribution, stored representations in a memory bank $Z^{(t-1)}$ to avoid recomputing features at each step, and used the proximal regularization term $\beta\|Z - Z^{(t-1)}\|_F^2$ to prevent representations from drifting too far between epochs. Furthermore, since computing the normalization constant in the softmax requires summing over all N samples at each step, intractable for large datasets, NCE approximates it by contrasting each sample against a small set of $m \ll N$ noise samples drawn from a uniform distribution. Subsequently, Oord, Li, and Vinyals [141] introduced *InfoNCE*, removing the proximal constraint and explicitly using positive pairs from different views, denoting by $\mathbb{P}$ the set of such pairs (i, j):

$$\mathcal{L}_{\text{InfoNCE}} = -\sum_{(i,j)\in\mathbb{P}} \log\left(\frac{e^{\cos(z_i, z_j)/\tau}}{\sum_{k=1}^{N} e^{\cos(z_i, z_k)/\tau}}\right) \tag{2.2}$$

All in all, while these early methods demonstrated the potential of SSL, the learned representations often lagged behind supervised pre-training on complex downstream tasks. Nevertheless, they paved the way for modern SSL, and namely joint-embedding self-supervised learning which arose from a few key refinements: the use of heavy data augmentation policies to define positive and negative pairs, the scaling of architectures and datasets, and the development of objectives that scale with the size of available data. These early methods form a bridge between foundational consistency principles and the state-of-the-art joint-embedding self-supervised learning methods discussed in the next section.

2.2 Joint-Embedding Self-Supervised Approaches

Recent and highly successful Self-Supervised Learning (SSL) approaches have largely converged toward a few dominant families, significantly closing the gap with supervised methods and opening broad transfer opportunities: the learned representations can be adapted to a wide range of downstream tasks [55, 9, 169, 70, 45, 131, 41, 190]. Among these, Joint-Embedding Self-Supervised Learning (JE-SSL) has emerged as a pivotal paradigm.

JE-SSL learns representations without labeled data by enforcing invariance to input transformations, such as rotation or blurring. An overview of the JE-SSL architecture is presented in Figure 2.3.

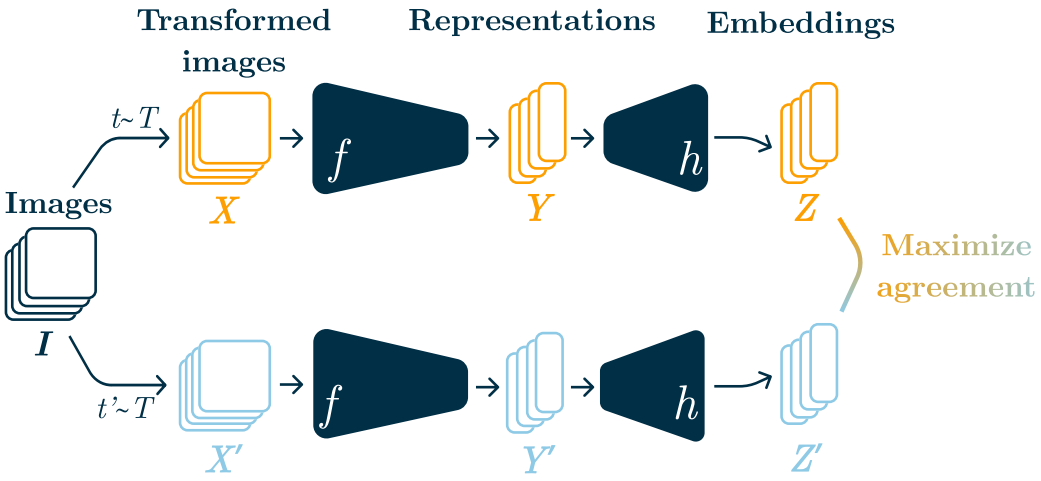


Figure 2.3: **Overview of joint-embedding architectures.** Two separate data augmentation operators are sampled from the same distribution ($t, t' \sim \mathcal{T}$) and applied to each image in a batch I to obtain two views, X and X'. An encoder network f and a projection head g are trained to maximize agreement between the resulting embeddings. After training, the encoder f and its representations Y are retained for downstream tasks.

More precisely, an encoder (f) is trained to produce similar embeddings for different 'views' of the same image. The views are typically generated by applying various combinations of stochastic augmentations, including random resizing, cropping, color jitter, and Gaussian blur. After encoding each view, JE-SSL methods typically employ a projector (h), often a Multi-Layer Perceptron (MLP) with Rectified Linear Unit (ReLU) activations, to map the initial embeddings into a latent space. An invariance loss is then applied in the embedding space to maximize agreement between the views. However, an invariance loss function (such as Euclidean distance) alone can lead to a trivial solution where the network outputs

identical, non-informative vectors regardless of the input; a phenomenon known as representation collapse. To mitigate this, researchers have developed various strategies.

The remainder of this section is organized as follows. We review methods in four families, grouped by the mechanism that prevents collapse. Section 2.2.1 covers contrastive methods, which explicitly contrast embeddings of different images as negative pairs. Section 2.2.2 covers clustering-based methods, which balance assignments across prototypes. Section 2.2.3 covers self-distillation methods, which rely on an asymmetric student–teacher architecture. Section 2.2.4 covers redundancy-reduction methods, which enforce decorrelation across embedding dimensions. Finally, Section 2.3 goes beyond these four families, covering both extensions of the joint-embedding principles (multimodal architectures and large-scale SSL methods) and alternative paradigms such as masked image modeling. An overview of methods is presented in Table 2.1.

Table 2.1: **Overview of self-supervised learning methods.** Joint-Embedding SSL (JE-SSL) methods are first grouped by the mechanism that prevents collapse: *contrastive* methods explicitly treat embeddings of different images as negative pairs; *clustering-based* methods balance assignments across prototypes; *self-distillation* methods rely on an asymmetric student–teacher architecture; and *redundancy-reduction* methods enforce decorrelation across embedding dimensions through soft or hard whitening. Finally, approaches *beyond JE-SSL* cover masked-image modeling, multimodal joint-embedding, and large-scale pretraining, which build on or complement these principles.

Method	Description	Objective
Joint-Embedding Self-Supervised Approaches (Section 2.2)		
Contrastive Methods (Section 2.2.1)		
MoCo [83, 35]	Uses a momentum encoder and queue-based dictionary to enable contrastive learning without large batch sizes	InfoNCE
SimCLR [34]	Introduces in-batch negatives and a non-linear projector, pioneering end-to-end contrastive SSL	InfoNCE
Clustering-based Methods (Section 2.2.2)		
SwAV [25]	Performs online clustering via Sinkhorn-Knopp, which enforces balanced prototype assignments, and predicts them across views	Swapped prediction (cross-entropy)

Continued on next page...

Table 2.1 *(continued)*

Method	Description	Objective
DINO [26]	Distills soft prototype assignments from an EMA teacher, preventing collapse through output centering & sharpening	Cross-entropy
Self-Distillation Methods (Section 2.2.3)		
BYOL [78]	Eliminates negative samples by training an online network to predict the representations of an EMA-updated target network	ℓ_2-normalized MSE
SimSiam [36]	Prevents collapse without negative pairs or momentum encoders by utilizing a simple stop-gradient operation	Cosine similarity
Redundancy Reduction Methods (Section 2.2.4)		
BarlowTwins [213]	Enforces invariance and prevents collapse by optimizing the empirical cross-correlation matrix between views	Cross-correlation
VICReg [12]	Prevents collapse by regularizing variance and covariance independently within each view, combined with an explicit invariance objective	Var. + Inv. + Cov.
FroSSL [172]	Extends VICReg-style soft whitening by minimizing covariance Frobenius norms to avoid collapse	Frob. norm + Inv.
W-MSE [60]	Bypasses negative samples by explicitly decorrelating batch representations through a hard whitening transform before aligning positive views	Whitened MSE
INTL [206]	Modulates the embedding spectrum beyond whitening to prevent collapse	Spectral modulation
Beyond Joint-Embedding Approaches (Section 2.3)		
Masked-Image Modeling (Section 2.3.1)		
MAE [82]	Reconstructs masked patches from raw pixels	Pixel reconstruction (MSE)

Continued on next page...

Table 2.1 *(continued)*

Method	Description	Objective
iBOT [221]	Performs MIM via self-distillation using an online tokenizer; combines patch-level prediction with a cross-view objective inspired by DINO	Cross-entropy×2 (MIM + cross-view)
I-JEPA [8]	Predicts masked target representations from a context region in latent space using an EMA encoder, eliminating the need for hand-crafted augmentations	Latent prediction (MSE)
Multimodal Joint-Embedding (Section 2.3.2)		
CLIP [151]	Aligns image and text embeddings at scale via cross-modal contrastive learning	InfoNCE (symmetric)
SigLIP [214]	Replaces softmax contrastive loss with a pairwise sigmoid objective	Sigmoid loss
ImageBind [72]	Binds six modalities using image-paired data as a common anchor	InfoNCE (image-anchored)
Scaling SSL (Section 2.3.3)		
DINOv2 [143]	Combines improved training recipes on the curated LVD-142M dataset, becoming the first SSL model to match CLIP's general-purpose features	iBOT + SwAV + KoLeo
DINOv3 [168]	Extends DINOv2 with Gram anchoring to preserve dense features over long training	DINOv2 + Gram
Franca [197]	Refines DINOv2 via a nested Matryoshka clustering projector, combining CyclicMask and RASA to decouple semantic representations from absolute positional biases	DINOv2 + Multi-head clustering + RASA

2.2.1 Contrastive methods

Contrastive methods [86, 135, 34, 83, 35] encourage the embeddings of augmented views of the same image to be similar, while ensuring that embeddings from different images remain distinct. Formally, for a given sample x_i and its embedding $z_i = g(f(t(x_i)))$, a positive pair is formed by computing the embedding z_i' of a second, independently sampled view of the same instance x_i. Conversely, negative

pairs are constructed by pairing z_i with the embeddings $\{z_j\}_{j \neq i}$ of $N-1$ distinct instances within the same batch or from a memory bank. The objective is typically InfoNCE Equation (2.2), which maximizes agreement between positive pairs while minimizing the similarity between negative pairs.

SimCLR [34] was a landmark method that paved the way for modern JE-SSL. Building upon earlier work in multi-view invariance [54, 141], it introduced a simplified framework for learning visual representations by relying on sampling negatives from within the same batch and employing the InfoNCE loss. Furthermore, it established several pivotal concepts: *(i)* the composition of multiple data augmentations is critical for defining effective pretext tasks; *(ii)* the inclusion of a non-linear projector (h in Figure 2.3) significantly improves representation quality; and *(iii)* performance scales consistently with larger batch sizes and extended training schedules.

One limitation of in-batch contrastive methods is their reliance on large batches, which are needed to provide enough negatives. To bypass this issue, negatives can instead be drawn from a memory bank, as in MoCo [83, 35]. More precisely, MoCo maintains a queue of keys from previous mini-batches, encoded by a momentum encoder whose weights are an exponential moving average (EMA) of the main encoder's weights, which keeps the keys consistent over time. Collapse is still prevented by the InfoNCE negatives: the queue only makes it possible to use many of them without large batches.

However, despite their effectiveness, instance-level contrastive approaches require numerous negative comparisons to work well, which can lead to high memory consumption. This has spurred the exploration of alternative approaches.

2.2.2 Clustering methods

Clustering methods [209, 23, 7, 25, 69] are a first such alternative. Contrastive methods operate at the level of instances: each sample is treated as its own class, and every other sample acts as a negative. Clustering-based methods relax this view by grouping samples into clusters and using the cluster assignments as coarser pseudo-labels. Two views of an image are then encouraged to agree on their assignment to a set of prototypes, rather than on their individual embeddings. Note, however, that some of these methods, such as SwAV, still describe themselves as contrastive, since they contrast cluster assignments across views. They nonetheless use no explicit negative pairs: collapse is instead prevented by balancing the assignments across prototypes, as detailed below.

DeepCluster [23] is an early precursor of joint-embedding SSL: it does not compare

two views of an image directly, but alternates between clustering the features of the whole dataset with k-means and training the network to predict the resulting assignments as pseudo-labels. Since this requires a full pass over the dataset at each clustering step, it is an offline approach, and collapse is avoided with heuristics such as reassigning empty clusters and sampling uniformly over clusters. SwAV [25] later made this idea online and multi-view: it learns prototypes and uses the Sinkhorn-Knopp algorithm to compute balanced cluster assignments, which are used in a swapped-prediction loss between views. Collapse is thus prevented by the equipartition constraint imposed by Sinkhorn-Knopp, which plays the role that negatives play in contrastive methods.

Clustering methods were further simplified by DINO [26], in which the target pseudo-labels are produced by a teacher network whose weights are an EMA of the student's (see Section 2.2.3), and the student is trained to match them with a simple cross-entropy loss. The clustering step is fully absorbed into the network: the prototypes are the weights of the last layer of the projector, and the loss compares the teacher and student distributions over this large set of prototypes ($K = 65{,}536$ in the original work, far more than in SwAV). Although DINO is often presented as a self-distillation method, its non-collapse mechanism is thus that of clustering-based methods. Works building on DINO, such as iBOT [221] and DINOv2/v3 [143, 168], have produced some of the strongest visual representations available; they are discussed further in Section 2.3.

Clustering methods nonetheless inherit some issues of contrastive ones: samples are contrasted with prototypes rather than with each other, and performance depends on having many of them. DINOv2 thus uses 128,000 prototypes [143], and Sansone, Lebailly, and Tuytelaars [160] improves scaling with the number of prototypes by increasing their usage and diversity.

2.2.3 Self-distillation methods

Self-distillation methods such as BYOL [78] and SimSiam [36] incorporate architectural strategies inspired by knowledge distillation to avoid representation collapse.

BYOL (Bootstrap Your Own Latent) [78] first introduced self-distillation as a means to avoid representation collapse without requiring negative pairs. BYOL employs two neural networks: a student network and a teacher network. The student network includes an additional predictor to map its outputs to those of the target. Let x, x' denote two augmented views of an image. From the first view x, the student network outputs a representation $y = f_{\theta_s}(x)$ and an embedding $z = g_{\phi_s}(y)$. The target network outputs $y' = f_{\theta_t}(x')$ followed by an embedding $z' = g_{\phi_t}(y')$. BYOL then produces a prediction $q_\zeta(z)$ of z', where q_ζ is a predictor

network with parameters ζ. Both $q_\zeta(z)$ and z' are $l2$-normalized to:

$$\overline{q_\zeta(z)} = \frac{q_\zeta(z)}{\|q_\zeta(z)\|_2}, \quad \overline{z'} = \frac{z'}{\|z'\|_2}$$

Importantly, the predictor q_ζ is only applied to the online branch, making the architecture asymmetric. Finally, the overall objective is defined as the mean squared error between the normalized predictions and target embeddings:

$$\mathcal{L}_{\text{BYOL}} = \mathbb{E}\left[||\overline{q_\zeta(z)} - \overline{z'}||_2^2\right] \tag{2.3}$$

The teacher network f_{θ_t} has the same architecture as the student network f_{θ_s}. It is updated by EMA (exponential moving average), with ξ controlling to what degree the teacher network preserves history:

$$\theta_t = \xi\theta_t + (1-\xi)\theta_s \tag{2.4}$$

Subsequent methods build upon BYOL: SimSiam [36] replaces the moving average encoder with a simple stop-gradient operation.

Although these methods have shown strong results, some limitations remain. Since they rely only on an asymmetric architecture, without explicit negatives or prototypes, the mechanism preventing collapse is still only partially understood [206]. In particular, the objective itself does not formally rule out collapse: a constant output is a global minimum of the loss. Existing analyses suggest that the predictor, the stop-gradient, and the optimization dynamics play a central role in steering training away from this trivial solution, but a complete theoretical account is still lacking.

2.2.4 Redundancy reduction methods

Redundancy reduction methods attempt to avoid collapse by relying on *whitening*. Whitening is a preprocessing step in statistical analysis that decorrelates random variables so that their components are mutually orthogonal, equivalently yielding identity 'white' covariance: $\text{Cov}(Z) = I$ [100]. We group these methods in two families: *soft whitening* and *hard whitening*.

Soft-whitening methods. Soft-whitening methods encourages the embedding to be decorrelated through regularization [213, 12, 172].

A notable example is VICReg (Variance Invariance Covariance Regularization) [12]. VICReg builds on BarlowTwins [213]. It leverages the same decorrelation principle, penalizing off-diagonal covariance terms, while adding an explicit per-dimension

variance term for both embedding views, which stabilizes scales without the batch-normalization proposed in BarlowTwins. More precisely, VICReg combines three terms:

1. *Invariance*: Mean squared Euclidean distance between corresponding embedding vectors across views.
2. *Variance*: A hinge penalty that encourages the standard deviation of each embedding dimension (within the batch) to be at least one—equivalently, it targets unit diagonal variance—thereby preventing collapse to a constant representation.
3. *Covariance*: Off-diagonal regularization of the empirical covariance matrix, pushing cross-dimension correlations toward zero.

Formally, let $Z = \{z_1, ..., z_n\} \subseteq \mathbb{R}^d, Z' = \{z'_1, ..., z'_n\} \subseteq \mathbb{R}^d$ be two sets of n embeddings in d-dimensional space coming out of the two branches of the joint architecture (see Figure 2.3). For each dimension $j \in \{1, ..., d\}$, we denote z^j as the vector containing all values at dimension j across the embeddings. The variance term is defined as:

$$v(Z) = \frac{1}{d} \sum_{j=1}^{d} \max(0, 1 - S(z^j, \varepsilon)) \tag{2.5}$$

where $S(x, \varepsilon) = \sqrt{\text{Var}(x) + \varepsilon}$ is a stability-adjusted standard deviation, with $\varepsilon > 0$ being a small constant that prevents numerical instabilities.

The covariance term is defined as:

$$c(Z) = \frac{1}{d} \sum_{i \neq j} [C(Z)]^2_{i,j} \tag{2.6}$$

where $C(Z)$ is the sample covariance matrix: $C(Z) = \frac{1}{n-1}(Z - \overline{z})^T(Z - \overline{z})$, with $\overline{z} = \frac{1}{n} \sum_{i=1}^{n} z_i$ the mean value of the embeddings.

Finally, the invariance criterion between Z and Z' is defined as the mean-squared Euclidean distance between each pair of vectors, without any normalization:

$$s(Z, Z') = \frac{1}{n} \sum_{i} ||z_i - z'_i||^2_2 \tag{2.7}$$

The overall objective is a weighted average of the three objectives:

$$\mathcal{L}_{\text{VICReg}}(Z, Z') = \lambda s(Z, Z') + \nu \left[v(Z) + v(Z')\right] + \tau \left[c(Z) + c(Z')\right] \tag{2.8}$$

where λ, ν, τ are hyperparameters controlling the importance of each term in the loss.

More recent work moves beyond the VICReg-style decomposition while preserving the broad redundancy-reduction goal [172, 91]. For example, FroSSL [172] defines its variance term through the logarithm of the squared Frobenius norm of each view's normalized covariance matrix.

Hard Whitening Methods. On the other hand, hard whitening methods apply explicit whitening transforms, such as batch or channel whitening, to the representations.

W-MSE (Whitening MSE) [60] transforms embeddings with a batch whitening map so that each view has approximately identity sample covariance, then minimizes the mean squared error (MSE) between the two whitened views. Formally, let $Z = \{z_1, \dots, z_n\} \subseteq \mathbb{R}^d$ and $Z' = \{z'_1, \dots, z'_n\} \subseteq \mathbb{R}^d$ be the two batches of d-dimensional embeddings produced by the joint architecture (see Figure 2.3). The problem is formalized as follows:

$$\min_{\theta} \ \frac{1}{n} \sum_{i=1}^{n} \|z_i - z'_i\|_2^2 \tag{2.9}$$

$$\text{s.t.} \quad \mathrm{Cov}(Z) = \mathrm{Cov}(Z') = I, \tag{2.10}$$

where $\mathrm{Cov}(\cdot)$ denotes the (centered) batch sample covariance. Because Equation (2.10) is difficult to enforce directly on raw network outputs, W-MSE instantiates the constraint by applying a *whitening transform* to each batch separately—for example, using the inverse matrix square root or Cholesky factor of the estimated covariance [60]. Let $\tilde{z}_i = \mathcal{W}_Z(z_i)$ and $\tilde{z}'_i = \mathcal{W}_{Z'}(z'_i)$ denote the whitened vectors, so that $\mathrm{Cov}(\tilde{Z}) \approx I$ and $\mathrm{Cov}(\tilde{Z}') \approx I$ for the corresponding whitened batches $\tilde{Z}$ and $\tilde{Z}'$. The W-MSE objective is the empirical MSE between whitened pairs:

$$\mathcal{L}_{\text{W-MSE}}(Z, Z') = \frac{1}{n} \sum_{i=1}^{n} \left\| \tilde{z}_i - \tilde{z}'_i \right\|_2^2. \tag{2.11}$$

CW-REG [205] introduces channel whitening, while Zero-CL [218] combines both batch and channel whitening techniques.

Building upon these methods, INTL [206] proposed to modulate the embedding spectrum and explore functions beyond whitening to prevent dimensional collapse.

Overall, Joint-Embedding Self-Supervised Learning (JE-SSL) methods have been shown to learn strong visual representations and to achieve competitive results on standard ImageNet-based benchmarks [47, 43]. Their success has spurred their wide adoption across the SSL landscape.

2.3 Beyond Joint-Embedding Approaches

The previous section focused on unimodal Joint-Embedding SSL, where two augmented views of an image are aligned in a shared latent space. Beyond purely JE-SSL, several complementary families have emerged: reconstruction-based methods that recover masked regions of the input, multimodal methods that treat distinct data sources as views, and large-scale pretraining recipes that go well beyond the canonical ImageNet-centric setup [37, 143].

2.3.1 Masked-Image Modeling

Masked-Image Modeling (MIM) methods draw inspiration from masked language modeling in NLP [48] and train models to reconstruct held-out (masked) regions of the input. Although MIM optimizes a reconstruction objective that differs substantially from standard JE-SSL losses, recent work increasingly draws connections between the two: MIM losses are combined with joint-embedding, invariance, or self-distillation terms, improving training stability and downstream representation quality relative to reconstruction alone.

Early MIM methods include BEiT [11], which applies BERT-style masked pretraining to Image Transformers: a random fraction of patches is replaced with learnable mask tokens, and the network is trained to predict their contents as discrete visual tokens from a fixed tokenizer (a pretrained discrete VAE). With this recipe, BEiT attains strong accuracy after full fine-tuning, but lags leading JE-SSL methods on frozen-backbone benchmarks (e.g., linear probing).

A few concurrent works expanded upon this idea, exploring in particular different choices of target encoder to replace the discrete VAE. iBOT (image BERT with an Online Tokenizer) [221] proposed replacing the pre-trained dVAE with an exponential moving average (EMA) of the online model. As this objective alone proved unstable, iBOT combined its MIM loss with a self-distillation JE-SSL objective, namely DINO [25]. This combination stabilizes training and yields state-of-the-art visual representations.

Taking another approach, MAE (Masked Auto-Encoders) [82] proposed to reconstruct raw pixel values at a high masking ratio (e.g., 75%), relying on an asymmetric encoder–decoder architecture that processes only visible patches. However, a conceptual limitation shared by pixel-reconstruction methods is that the model must devote capacity to predicting low-level, potentially uninformative details. Instead *Joint-Embedding Predictive Architectures* such as I-JEPA [8] inspired by JE-SSL proposed to predict representations rather than reconstructing missing pixels. Instead, the model predicts the representation a target encoder would assign

to the masked region. More precisely, given a context block (the visible region), a predictor network is trained to output the representation that a momentum target encoder produces for the masked target blocks, at those specific spatial positions. Crucially, neither pixel reconstruction nor explicit data augmentation is required; invariance is instead induced implicitly by conditioning on positional information alone.

2.3.2 Multimodal joint-embedding methods

Beyond the unimodal setting, joint-embedding principles extend naturally to settings in which views come from different modalities, rather than from different augmentations of a single input. For instance, in Figure 2.4, the top image can be paired with the following text: 'Venice from the Punta della Dogana'. An overview of multimodal joint-embedding architecture is presented in Figure 2.4.

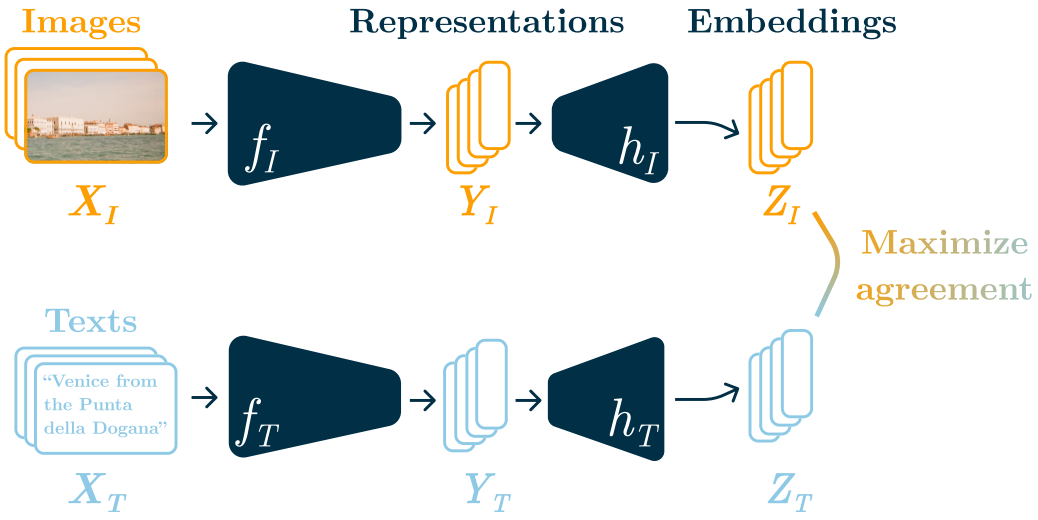


Figure 2.4: **Overview of Mmultimodal joint-embedding self-supervised learning.** Each modality (e.g., image, text) is processed by a dedicated encoder and projection head. An alignment objective (typically contrastive) pulls together embeddings of paired samples across modalities while pushing apart unpaired ones, yielding a shared latent space in which semantically corresponding inputs from different modalities are co-located.

Early work showed that training on large-scale image datasets with noisy captions yields strong visual features [98]. Building on this intuition, CLIP [151] and ALIGN [94] formalized cross-modal alignment at scale.

Given N image–text pairs $\{(x_i^{\mathrm{I}}, x_i^{\mathrm{T}})\}_{i=1}^{N}$, CLIP learns an image encoder $f_{\mathrm{I}}(\cdot)$ and a text encoder $f_{\mathrm{T}}(\cdot)$ such that the resulting ℓ_2-normalized embeddings $\{z_i^{\mathrm{I}}, z_i^{\mathrm{T}}\}_{i=1}^{N} =$

$\{f_{\mathrm{I}}(x_i^{\mathrm{I}}),\ f_{\mathrm{T}}(x_i^{\mathrm{T}}))\}_{i=1}^{N}$ are drawn closer for matched pairs and pushed apart otherwise. The training objective is an InfoNCE-style [141] loss:

$$\mathcal{L}_{\mathrm{I,T}} = \frac{1}{n}\sum_{i=1}^{N} -\log \frac{\exp\left(\cos(z_i^{\mathrm{I}}, z_i^{\mathrm{T}})/\tau\right)}{\sum_{j=1}^{N}\exp\left(\cos(z_i^{\mathrm{I}}, z_j^{\mathrm{T}})/\tau\right)}, \tag{2.12}$$

where τ is a learned temperature and $\cos(\cdot,\cdot)$ denotes cosine similarity. In practice, a symmetric loss is used: $\mathcal{L}_{\mathrm{I,\,T}} + \mathcal{L}_{\mathrm{T,\,I}}$.

Trained on hundreds of millions of image–text pairs, CLIP and ALIGN, along with subsequent methods (e.g., SigLIP [214], EVA-CLIP [178], SigLIP2 [189]), obtain representations that transfer remarkably well to downstream tasks. Similar bimodal frameworks have emerged beyond vision–language; for example, CLAP [59] aligns audio and text in a shared embedding space. Subsequent work scaled this to more than two modalities: ImageBind [72] binds six modalities by using image-paired data as a bridge, showing that exhaustive pairwise alignment is unnecessary. LanguageBind [222] further generalized this idea by using language as a unifying anchor to connect video, audio, depth, thermal, and inertial signals within a single joint embedding space.

Given the success of such large-scale multimodal frameworks and their transferable representations, a complementary line of research has focused on scaling self-supervised image pretraining.

2.3.3 Scaling SSL

Scaling model size and training data has proven a reliable lever for improving SSL representations. Early experiments [24, 77] used the YFCC dataset [185], with subsequent work showing that larger datasets and models yield further gains [75, 76] and that principled data curation offers a complementary avenue [186].

Successive improvements to training recipes, model scale, and dataset curation drove consistent gains in self-supervised learning, with DINOv2 [143] marks a significant milestone. DINOv2 builds on DINO [26] and iBOT [221]. It is pretrained on LVD-142M, an automatically curated corpus of 142 million images. Its objective combines four terms: a self-distillation loss on global `[CLS]` tokens (from DINO), a patch-level masked image modeling term for spatially grounded representations (from iBOT), a Sinkhorn–Knopp clustering objective (from SwAV), and a KoLeo regularizer guarding against collapse. DINOv2 was the first purely vision SSL model to match or surpass multimodal approaches such as CLIP across a range of tasks.

Several works build on DINOv2. DINOv3 [168] scales data curation and refines the

training recipe for improved dense prediction, addressing the observed tendency of global representations to dominate local ones over long training. Its key contribution, *Gram anchoring*, enforces patch-level consistency by training the student to match the Gram matrix of an early teacher checkpoint, where dense features are strongest, without directly constraining the features themselves. In a complementary direction, Franca [197] introduces Matryoshka multi-head clustering and a spatial-semantic disentanglement step applied after pretraining, achieving competitive results with fully open-source resources.

Taken together, these directions reflect a broader shift: Joint-Embedding principles, originally conceived for unimodal image representation, have progressively extended toward general-purpose visual encoders that transfer across tasks and modalities with little or no fine-tuning.

2.4 Minimum Spanning Trees

Meanwhile, from a different corner of computer science, come minimum spanning trees (MST). The MST is a classical combinatorial object [108, 149]: given a finite point cloud P, it is the connected, acyclic graph on vertex set P with minimal total edge length. While its origins lie in network design and graph theory, the MST length also encodes some intrinsic geometric properties of the underlying data distribution, scaling asymptotically with the Rényi entropy of distribution which provides a principled way to characterize a measure from its samples.

We start by defining the minimum spanning tree. An illustration is presented in Figure 2.5. Let $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be a finite point cloud, with pairwise distances given by the Euclidean distance $\|u - v\|_2$ for $u, v \in X$.

Definition 2.4.1. (Spanning Tree). A spanning tree of X is an undirected graph $G = (V, E)$ with vertex set $V = X$ and edge set $E \subseteq V \times V$ such that G is connected and acyclic.

Definition 2.4.2. (Minimum Spanning Tree). A minimum spanning tree (MST) of X is a spanning tree $G = (V, E)$ of minimum total length:

$$L\left(\mathrm{MST}(X)\right) \coloneqq \sum_{(u,v)\in E} \|u - v\|_2.$$

We refer to $L\left(\mathrm{MST}(X)\right)$ as the MST *length* of X.

The remainder of this section first introduces the connection between the MST length and the intrinsic dimension of the support of a distribution (Section 2.4.1), then covers the connection with topological data analysis and the differentiability

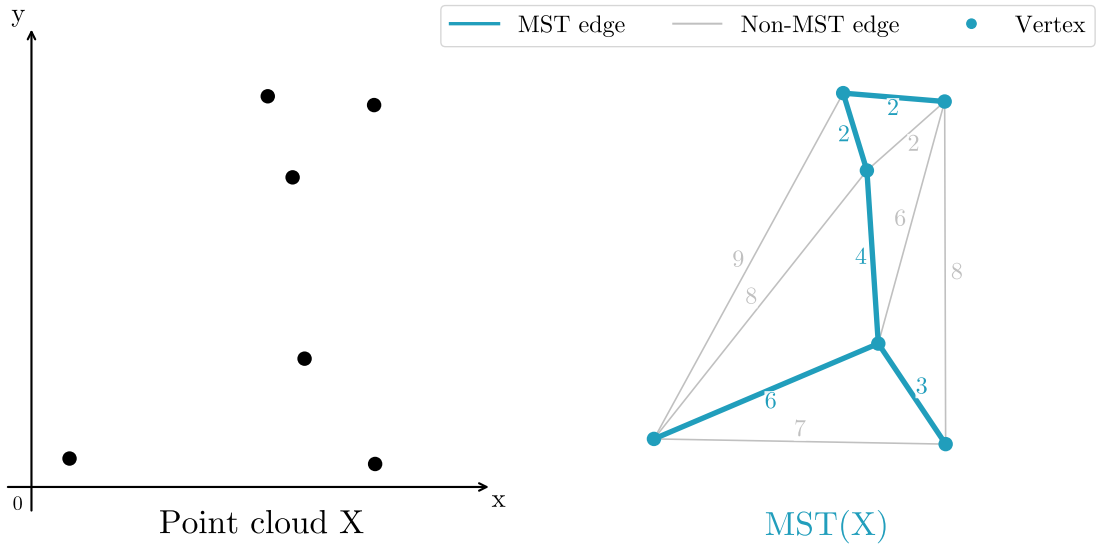


Figure 2.5: **Overview of a minimum spanning tree (MST).** *(Left)* a point cloud $X \subset \mathbb{R}^2$. *(Right)* The MST(X) connects all points without forming any cycle, while minimizing the total edge length. *Grey edges* indicate pairwise connections not retained in the MST: replacing a MST by one of them would either create a cycle or increase the total length.

properties of the MST (Section 2.4.2), before turning to its algorithmic construction (Section 2.4.3).

2.4.1 MST and Dimension Estimation

Steele [176] established a foundational result on the total length of the minimum spanning tree (MST) for random subsets of Euclidean space.

Theorem 2.4.3 (Steele [176])**.** *Let X_n be an i.i.d. n-sample drawn from a probability measure P_X with compact support on $\mathbb{R}^d$. For $d \geq 2$, the growth rate of the length of* MST(X_n) *is controlled as follows:*

$$L\left(\mathrm{MST}(X_n)\right) \sim C n^{(d-1)/d} \text{ almost surely, as } n \to \infty, \tag{2.13}$$

where $\sim$ denotes asymptotic convergence, and where C is a constant depending only on P_X and d.

In particular, Eq. (2.13) shows that whenever the absolutely continuous part of P_X is non-trivial, the length of the MST grows as $n^{(d-1)/d}$, so the growth rate is governed by the dimension d of the support of P_X. Costa and Hero [40] subsequently extended this result to compact Riemannian manifolds, where the data lie

on a lower-dimensional support and the limit is expressed in terms of the density with respect to the manifold's volume measure.

Theorem 2.4.4 (Costa and Hero [40], Corollary 5)**.** *Given an i.i.d. n-sample X_n drawn from a fixed probability measure P_X supported on a compact Riemannian d-manifold M with density f_X w.r.t. the Hausdorff measure, there exists a universal constant C' independent of f_X and of M such that, almost surely:*

$$n^{-\frac{d-1}{d}} \cdot L\left(\mathrm{MST}(X_n)\right) \xrightarrow[n\to\infty]{} C' \int f_X^{\frac{d-1}{d}} \mathrm{d}\mathcal{H}, \tag{2.14}$$

The significance of this result lies in showing that the intrinsic dimension d of the support of P_X, an a priori unknown quantity, can be recovered solely from the scaling behavior of the MST length, without any prior knowledge of the manifold M or the density f_X. This is of particular practical relevance, as real-world data (e.g., images, text, or sensor measurements) is generally represented in a high-dimensional ambient space while effectively residing on a lower-dimensional support. The MST can thus provide a means of recovering this intrinsic dimension from finite samples, although the guarantees are asymptotic and the estimate may be biased in practice by noise, curvature, and limited sample size.

Intrinsic Dimension Estimators. Building on this connection, Theorem 2.4.4 motivates the definition of a concrete intrinsic dimension (ID) estimator, $\dim_{\mathrm{MST}}$, introduced by Costa and Hero [40]:

Definition 2.4.5. Given a bounded metric space M, the MST dimension of M, denoted by $\dim_{\mathrm{MST}}(M)$, is the infimal exponent $d \in \mathbb{N}$ such that $L\left(\mathrm{MST}(X)\right)/|X|^{\frac{d-1}{d}}$ is uniformly bounded for all finite subsets $X \subseteq M$:

$$\dim_{\mathrm{MST}}(M) := \inf\Big\{ d : \exists C \text{ s.t. } \frac{L\left(\mathrm{MST}(X)\right)}{|X|^{\frac{d-1}{d}}} \leq C \\ \text{for every finite subset } X \text{ of } M \Big\}.$$

$\dim_{\mathrm{MST}}$ theoretically coincides with the usual dimension in non-singular cases, and it empirically coincides with the real-valued Hausdorff dimension [62] in classical manifold examples or even fractal examples [130], such as the Cantor set or the Sierpiński triangle.

In practice, the ID is estimated via log-log linear regression [17, 16]. Specifically, given subsamples X_{n_i} with increasing sizes n_i, we fit:

$$\log(L(\mathrm{MST}(X_{n_i}))) \approx \frac{d-1}{d}\log(n_i) + \log(C).$$

The resulting slope m yields the intrinsic dimension via the relation $d = 1/(1-m)$. The complete algorithm for computing $\dim_{\mathrm{MST}}$ is given in Algorithm 1.

Algorithm 1 Computation of $\dim_{\mathrm{MST}}$

Input: The set of representations $Z = \{z_1, \ldots, z_N\}$, minimum sample size $n_{\min}$, skip size Δ
Output: The estimated dimension $\dim_{\mathrm{MST}}(Z)$
$n \leftarrow n_{\min}$, $E \leftarrow []$, $i \leftarrow 0$
while $n \leq N$ **do**
    $Z_n \leftarrow \mathrm{sample}(Z, n)$ ▷ random sample of n points
    $E[i] \leftarrow L\left(\mathrm{MST}(Z_n)\right)$ ▷ length of the minimum spanning tree
    $n \leftarrow n + \Delta$, $i \leftarrow i + 1$
end while
$m, b \leftarrow \mathrm{linear_regression}(\log(n_{\min}{:}\Delta{:}N), \log(E))$ ▷ log-log linear regression
return $\dim_{\mathrm{MST}}(Z) \leftarrow 1/(1-m)$

2.4.2 Connections to Topological Data Analysis and Differentiability

We recall the necessary background on Persistent Homology, restricting attention to the Vietoris-Rips (VR) filtration. Our interest in this framework is twofold: in degree 0, the total persistent homology of the VR filtration recovers exactly the MST length (see Eq. (2.16)), and TDA provides the mathematical tools needed for differentiability. We refer the interested reader to Edelsbrunner and Harer [58] and Oudot [145], or Loiseaux [124] for a thorough introduction to Topological Data Analysis.

The starting point of TDA is to extract quantitative topological information from structured objects, e.g., graphs, points sampled on a manifold, time series. Doing so relies on a piece of algebraic machinery called persistent homology (PH), which detects the presence of underlying topological features in a multiscale way. In this context, estimating topological properties should be understood as inferring the number of connected components (topology of dimension 0), of independent loops (topology of dimension 1), of enclosed voids (dimension 2), and so on in higher-dimensional settings.

Simplicial Complexes and Filtrations

A *simplicial complex* is a combinatorial object generalizing graphs and triangulations. Given a finite vertex set $X = \{x_0, \cdots, x_n\}$, a simplicial complex K is a downward-closed subset of the power set 2^X: each element $\sigma \in K$ is called a

simplex, and the closure condition requires $\sigma \in K \Rightarrow \tau \in K$ for all $\tau \subset \sigma$.

The elements of K are called its simplices. For any non-empty simplex σ, we call the number $|\sigma| - 1$ its *dimension*, denoted $\dim(\sigma)$. Given a simplicial complex K, a filtration of K is a sequence of sub-complexes that is increasing with respect to the inclusion $\emptyset \subset K^0 \subset \ldots \subset K^N = K$ such that every complex is obtained by adding one simplex to the previous one: $K^{i+1} = K^i \cup \{\sigma^{i+1}\}$.

The filtration is denoted by:

$$\emptyset \to K^0 \to \ldots \to K^N = K$$

and the corresponding simplices, in the order they are added to the filtration will be denoted $(\sigma_0, \cdots, \sigma_N)$.
Example. A classical example of simplicial filtration is the *Vietoris-Rips* (VR) filtration. An illustration is presented in Figure 2.6. Given a point cloud $X = \{x_0, \cdots, x_n\} \in \mathbb{R}^{n \times d}$, for any $t > 0$, the VR simplicial complex $\mathrm{Rips}(X, t)$ is defined as:

$$\{x_{i_1}, \cdots, x_{i_p}\} \in \mathrm{Rips}(X, t) \iff \forall j, j' \in \{1, \cdots, p\}, \quad \|x_{i_j} - x_{i_{j'}}\| \leq t.$$

Intuitively, a simplex enters the filtration at the scale equal to its diameter. Computationally, the VR filtration depends on X only through the pairwise distance matrix $(\|x_i - x_j\|)_{i,j}$.

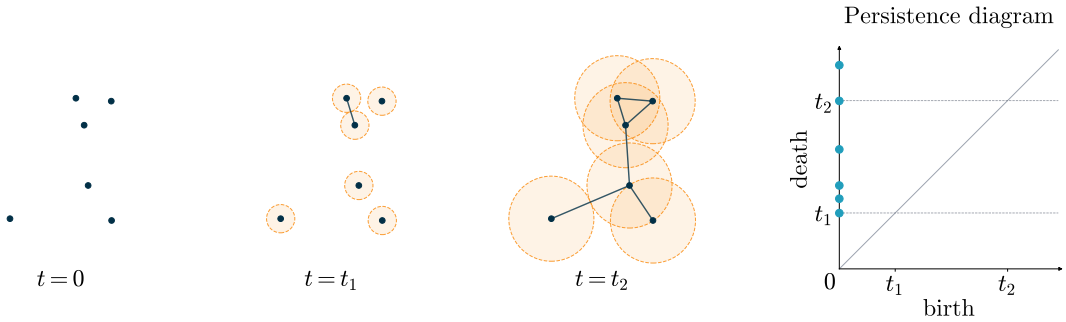


Figure 2.6: **Vietoris-Rips filtration on a point cloud (degree 0).** At $t = 0$ every point forms its own connected component, so all 0-dimensional classes are *born* simultaneously at 0. As the filtration parameter t increases, points within distance t are connected by an edge, progressively merging connected components (e.g. at t_1). Each merger is the *death* of a 0-dimensional homology class, recorded as a point $(0, t_{\text{death}})$ in the persistence diagram *(right)*.

Persistence Diagram and Total Persistence

Via a matrix reduction algorithm, Persistent Homology (PH) identifies, for each dimensional topological feature γ (connected component, hole, void and so on),

the *critical pair* $(\sigma_b^\gamma, \sigma_d^\gamma)$ of simplices that create and destroy γ. The *birth* and *death* times of γ are then:

$$b_\gamma := t(\sigma_b^\gamma), \qquad d_\gamma := t(\sigma_d^\gamma),$$

where $t(\sigma)$ denotes the filtration value at which simplex σ appears in the filtration. Its *lifetime interval* is $I(\gamma) := [b_\gamma, d_\gamma)$, with persistence $|I(\gamma)| := d_\gamma - b_\gamma \geq 0$.

For each dimension $i \geq 0$, the *persistence diagram* $\mathrm{Dgm}_i(X)$ is the multiset of all i-dimensional features:

$$\mathrm{Dgm}_i(X) := \Big\{(b_\gamma, d_\gamma) : \gamma \text{ is an } i\text{-dimensional feature of } (K_t)_t\Big\}.$$

The *total persistence* in dimension i then aggregates these lifetimes:

$$E^i(X) := \sum_{\gamma \in \mathrm{Dgm}_i(X)} |I(\gamma)|. \tag{2.15}$$

For $i = 0$, this quantity admits a closed form in terms of the MST: each edge of $\mathrm{MST}(X)$ corresponds to a 0-dimensional persistence pair, with the edge weight equal to the persistence of the associated feature. Consequently:

$$E^0(X) \;=\; L\left(\mathrm{MST}(X)\right). \tag{2.16}$$

Persistent Homology Dimension

Moreover, Persistent Homology (PH) has been used to define a family of fractal dimensions [2], $\dim_{\mathrm{PH}}^i(M)$, for each homological degree $i \geq 0$. In particular, for $i = 0$ this coincides with the MST-based dimension, i.e., $\dim_{\mathrm{PH}}^0(M) = \dim_{\mathrm{MST}}(M)$. The PH dimension can be derived from entropy computations and has already been used in several dimension-estimation applications [17, 57, 181].

MST Optimization

Carriere et al. [28] and Leygonie, Oudot, and Tillmann [116] provided the necessary mathematical framework for optimizing persistence-based functionals, such as the *total persistence of the VR filtration in degree* 0, or equivalently $L\left(\mathrm{MST}(X)\right)$, and establishing their convergence guarantees. First, since $L\left(\mathrm{MST}(X)\right)$ is Lipschitz continuous [31], Rademacher theorem guarantees it is differentiable almost everywhere, with derivatives given by the following closed-form expression:

$$\forall x \in X, \quad \nabla_x L\left(\mathrm{MST}(X)\right) = \sum_{\substack{(x,z) \text{ edge} \\ \text{of } \mathrm{MST}(X)}} \nabla_x \left\|x - z\right\|_2 = \sum_{\substack{(x,z) \text{ edge} \\ \text{of } \mathrm{MST}(X)}} \left\|x - z\right\|_2^{-1} (x - z). \tag{2.17}$$

In particular, Equation (2.17) shows that the two endpoints of each MST edge exert a repulsive force on one another when maximizing $L\,(\mathrm{MST}(X))$. Furthermore, under standard assumptions on the learning rate, stochastic gradient descent is guaranteed to converge almost surely to critical points of the functional [28, 116].

2.4.3 MST Computation

Given a finite point set $X \subset \mathbb{R}^d$, several classic sequential procedures exist to compute $\mathrm{MST}(X)$, notably Kruskal's, Prim's, and Boruvka's algorithms, which all have at least quadratic running time in the size of X.

Kruskal's algorithm [108] proceeds by sorting all pairwise edges by weight and greedily adding each edge to the tree, provided it does not create a cycle. Cycle detection is maintained efficiently via a union–find data structure, yielding an overall complexity of $\mathcal{O}(n^2 \log n)$ for a complete graph on n points.

Furthermore, although Kruskal's main loop is sequential, preprocessing (computing the distance matrix and sorting its entries) dominates in practice and can be efficiently optimized (as in `torchph` [20] or ripser [13]).

Chapter 3

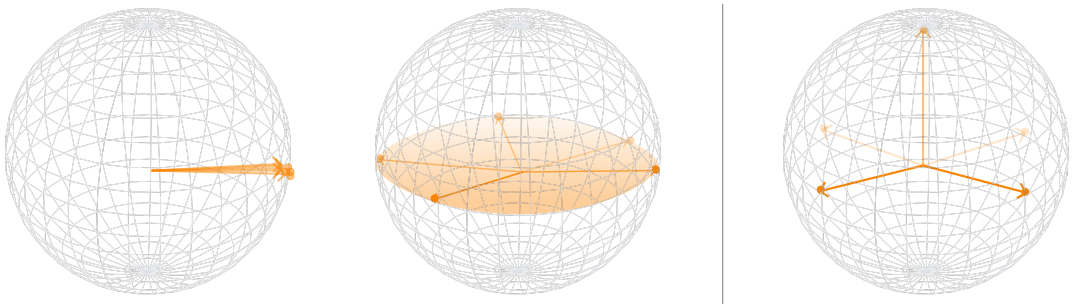
Minimum Spanning Tree Regularization for Self-Supervised Learning

As discussed in the previous section (Section 2.2), the main challenge of *Joint-Embedding Self-Supervised Learning* (JE-SSL) methods is to avoid the trivial solution known as *representation collapse*, whereby the network maps all inputs to a constant vector (Figure 3.1, *left*). In this chapter, we focus on its more insidious variant: *dimensional collapse*. We argue that a principled notion of *uniformity* offers a promising avenue to address it.

Section 3.1 motivates the problem, and Section 3.2 provides theoretical and empirical analyses demonstrating that T-REGS simultaneously mitigates dimensional collapse and promotes distributional uniformity on arbitrary compact Riemannian manifolds. Section 3.3 then validates the approach on classical SSL benchmarks, while Section 3.4 offers further experimental analyses.

This chapter is based on
T-REGS: Minimum Spanning Tree Regularization for Self-Supervised Learning
J. Mordacq, D. Loiseaux, V. Kalogeiton, S. Oudot
Spotlight, NeurIPS 2025

3.1 Introduction



× *Representational collapse*: Embeddings collapse to a single point.

× *Dimensional collapse*: Embeddings occupy a lower-dimensional subspace.

✓ *Uniformly distributed points*: Embeddings span the entire space.

Figure 3.1: **Illustration of collapse modes in JE-SSL versus uniformly sampled points.** *Representational collapse* (*left*) and *dimensional collapse* (*centre*) are the two main collapse modes in joint-embedding self-supervised learning: embeddings either degenerate to a single point or concentrate on a low-dimensional subspace, diminishing the ability to discriminate between inputs. An ideal regime (*right*) corresponds to embeddings that are *uniformly spread* across the representation space, preserving maximal information for downstream tasks.

Dimensional collapse [90, 96, 81, 117] (Figure 3.1, *center*) occurs when the embeddings span only a lower-dimensional subspace of the representation space, leading to high feature correlations and reduced representational diversity. Such a collapse can significantly impair the model's ability to capture the full complexity of the data, limiting performance on downstream tasks [68]. Another crucial aspect of representation quality is *uniformity*, which measures how evenly the embeddings are distributed across the representation space. It ensures that the learned representations preserve the maximum amount of information from the input data and avoid clustering in specific regions of the space (Figure 3.1, *right*). This property is fundamental because it helps maintain the discriminative power of the representations, thereby preventing dimensional collapse and allowing for better generalization to downstream tasks [63, 202, 161, 66].

While existing methods have made progress in mitigating dimensional collapse and enforcing uniformity, they present limitations: *contrastive methods* are sensitive to the number of negative samples [68, 12], and require large batch sizes, which can be computationally expensive. *Redundancy reduction* methods, which enforce the covariance matrix to be close to the identity matrix, only leverage the second moment of the data distribution and are blind to, e.g., concentration points of the density, which can prevent convergence to the uniform density (Figure 3.7); and

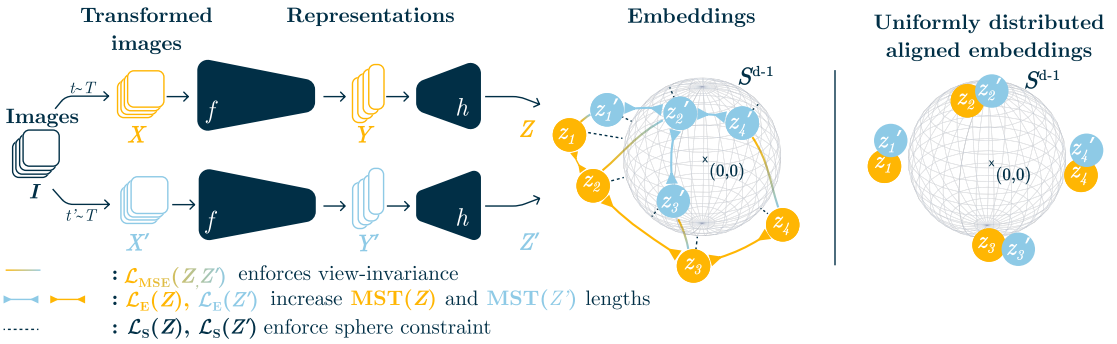


Figure 3.2: **Overview of T-REGS.** *(Left)* Two augmented views X, X' are encoded by f and projected by h into embeddings Z, Z'. Training jointly: *(i)* minimizes the Mean Squared Error, $\mathcal{L}_{\text{MSE}}(Z, Z')$, to enforce view invariance (or alternatively the objective function of a given SSL method, $\mathcal{L}_{\text{SSL}}(Z, Z')$, when used as an auxiliary term); *(ii)* maximizes the minimum-spanning-tree length on each branch, $\mathcal{L}_{\text{E}}(Z)$ and $\mathcal{L}_{\text{E}}(Z')$, repelling edge-connected points in MST(Z) and MST(Z'); and *(iii)* applies sphere constraints $\mathcal{L}_{\text{S}}(Z)$ and $\mathcal{L}_{\text{S}}(Z')$. *(Right)* As a result, T-REGS induces uniformly distributed embeddings without dimensional collapse.

asymmetric methods lack theoretical grounding to explain how the asymmetric network helps prevent collapse [206].

Given these limitations, Fang et al. [63] suggested rethinking the notion of good SSL regularization, and proposed an Optimal Transport-based metric that satisfies a set of four principled properties (*instance permutation*, *instance cloning*, *feature cloning*, and *feature baby constraints*) that prevent dimensional collapse and promote sample uniformity. Their approach has its drawbacks: the optimal transport distances are costly to compute in general, and the proposed closed formula for accelerating the computation holds only on the sphere and requires square roots over SVD computations, which may lead to numerical instabilities.

To address these limitations, we propose T-REG, a novel regularization approach that is conceptually simple, easy to implement, and computationally efficient. T-REG naturally satisfies the four principled properties of Fang et al. [63] (Section 3.2.2) and provably prevents dimensional collapse while promoting sample uniformity (Figure 3.3a). These properties make T-REG suitable for joint-embedding self-supervised learning, where it can be applied independently to each branch, yielding T-REGS (see Figure 3.2).

The central idea of T-REG is to maximize the length of the minimum spanning tree (MST) of the samples in the embedding space. It has strong theoretical connections to the line of work on statistical dimension estimation via entropy maxi-

mization [176] (Section 3.2.1). Since the length of MST(Z) scales under rescaling of Z, maximizing it alone leads the points to diverge (see Figure 3.3b). To prevent trivial scaling, T-REG constrains embeddings to a compact manifold, encouraging full use of the representation dimension and a uniform distribution.

The objective of spreading N points uniformly on a hypersphere is closely related to the *Tammes problem* [180]: given N points on the unit sphere $\mathbb{S}^{d-1}$, find the configuration that maximizes the minimum pairwise distance (or alternatively, the average pairwise distances). Intuitively, a well-regularized embedding space should resemble such a maximally spread configuration, ensuring that no two representations are unnecessarily close, providing uniformly sample points. However, naively maximizing the *minimum* pairwise distance yields very sparse gradients, as only the single closest pair of points contributes at each step. Conversely, maximizing the *average* pairwise distance distributes gradients across all pairs, treating already well-separated points equally to close ones. On the other hand, the MST length offers a middle ground between both minimum and average pairwise distances. In particular, Equation (2.17) shows that each pair of points forming an edge in the MST exerts a repulsive force on the other during optimization, achieving both gradient density and targeted repulsion.

The main contributions can be summarized as follows:

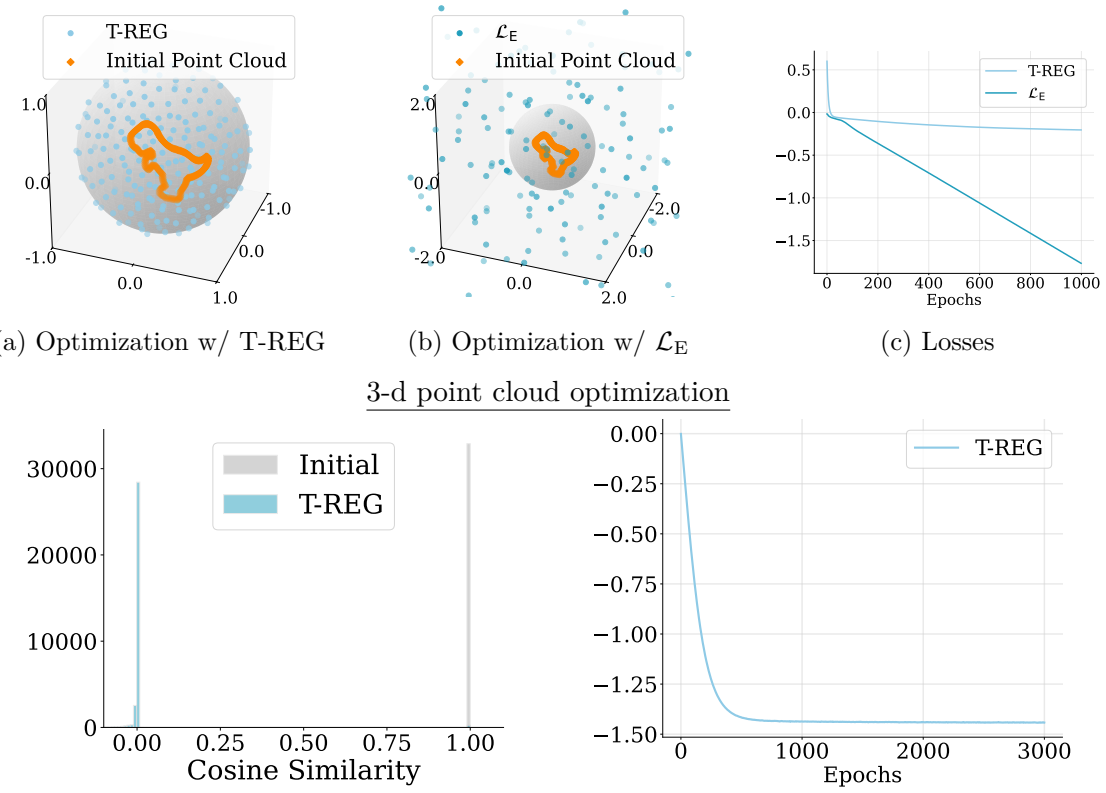
i) We introduce T-REG (Equation (3.3)), a regularization technique that maximizes the length of the minimum spanning tree (MST) while constraining embeddings to lie on a hypersphere.

ii) We show both theoretically and empirically that T-REG naturally prevents dimensional collapse while enforcing sample uniformity (Section 3.2).

iii) We apply T-REG to SSL either as standalone regularization, combining directly with view-invariance, or as an auxiliary loss to existing methods, yielding the T-REGS framework, whose effectiveness is evaluated through experiments on standard JE-SSL benchmarks (Section 3.3).

3.2 T-REG: Minimum Spanning Tree based Regularization

Our regularization T-REG has two terms: a length-maximization loss $\mathcal{L}_{\mathrm{E}}$ that decreases with the length of the minimum spanning tree, and a soft sphere-constraint $\mathcal{L}_{\mathrm{S}}$ that increases with the distance to a fixed sphere $\mathbb{S}$. These two terms combined force the embeddings to lie on $\mathbb{S}$ (or close to it), while spreading them out along $\mathbb{S}$.



(a) Optimization w/ T-REG (b) Optimization w/ $\mathcal{L}_E$ (c) Losses

(d) Optimization w/ T-REG (e) Loss

Figure 3.3: **Illustration of T-REG with synthetic data.** *(a-c) 3-d point cloud analysis:* (a) T-REG successfully spreads points uniformly on the sphere by combining MST length maximization and sphere constraint, (b) using only MST length maximization leads to excessive dilation, (c) stable convergence of T-REG whereas $\mathcal{L}_E$ alone fails to converge. *(d-e) Higher-dimensional analysis (256-d):* (d) T-REG enforces effective convergence to the 255-d regular simplex (Theorem 3.2.1), (e) stable optimization behavior of T-REG.

Formally, given $Z = \{z_1, ..., z_n\} \subseteq \mathbb{R}^d$, the MST length maximization loss is defined as:

$$\mathcal{L}_{\mathrm{E}}(Z) = -\frac{1}{n} L\left(\mathrm{MST}(Z)\right), \tag{3.1}$$

where $L\left(\mathrm{MST}(Z)\right)$ denotes the length of the MST of Z. The soft sphere-constraint is given by:

$$\mathcal{L}_{\mathrm{S}}(Z) = \frac{1}{n} \sum_i (\|z_i\|_2 - 1)^2. \tag{3.2}$$

It penalizes points that move away from the unit sphere. Maximizing the MST length alone would cause the points to diverge to infinity; the sphere constraint prevents this by keeping the embeddings within a fixed region around $\mathbb{S}$ (Figures 3.3b and 3.3c). The overall T-REG loss combines these two terms:

$$\mathcal{L}_{\text{T-REG}}(Z) = \gamma\, \mathcal{L}_{\mathrm{E}}(Z) + \lambda\, \mathcal{L}_{\mathrm{S}}(Z), \tag{3.3}$$

where γ and λ are hyperparameters controlling the trade-off between spreading out the embeddings and maintaining them on the sphere.

The remainder of this section is organized as follows. Section 3.2.1 provides a theoretical analysis of T-REG, Section 3.2.2 verifies the uniformity properties of Fang et al. [64], and Section 3.2.3 offers empirical evaluations.

3.2.1 Theoretical analysis

Behavior on small samples

We begin by considering the case where $n \leq d + 1$. It is particularly relevant since, in SSL, batch sizes are often smaller than or comparable to the ambient dimension. In order to account for the effect of the soft sphere constraint, we assume the points of X lie inside some fixed closed Euclidean d-ball B of radius r centered at the origin (see below for the explanation).

Theorem 3.2.1. *Under the above conditions, the maximum of $L\left(\mathrm{MST}(X)\right)$ over the point sets $X \subset B$ of fixed cardinality n is attained when the points of X lie on the sphere $S = \partial B$, at the vertices of a regular $(n-1)$-simplex that has S as its smallest circumscribing sphere.*

Recall that a *k-simplex* is the convex hull of a set of $k+1$ points that are affinely independent in $\mathbb{R}^d$—which is possible only for $k \leq d$. The simplex is *regular* if all its edges have the same length, i.e., all the pairwise distances between its vertices are equal. In such a case, we have the following relation between its edge length a and the radius r of its smallest circumscribing sphere:

$$a = r\sqrt{\frac{2(k+1)}{k}}. \tag{3.4}$$

Theorem 3.2.1 explains the behavior of T-REG as follows: first, minimizing the term $\mathcal{L}_{\mathrm{E}}$ in Equation (3.3) expands the point cloud until the sphere constraint term $\mathcal{L}_{\mathrm{S}}$ becomes the dominating term (which happens eventually since $\mathcal{L}_{\mathrm{S}}$ grows quadratically with the scaling factor, versus linearly for $\mathcal{L}_{\mathrm{E}}$); at that stage, the points stop expanding and start spreading themselves out uniformly along the sphere of directions. The amount of expansion before spreading is prescribed by the strength of the sphere constraint term versus the term in the loss, which is driven by the ratio between their respective mixing parameters λ and γ.

The proof of Theorem 3.2.1 relies on the following two ingredients: a standard result in convex geometry (Theorem 3.2.2), and a technical lemma relating the length of the MST to the sum of pairwise distances (Theorem 3.2.3).

Proposition 3.2.2 (Eq. (14.25) in Apostol and Mnatsakanian [6])**.** *Under the conditions of Theorem 3.2.1, and assuming $n = d+1$, the sum of pairwise distances $\sum_{1\leq i<j\leq n} \|z_i - z_j\|_2$ is maximal when the points of X lie on the bounding sphere S, at the vertices of a regular d-simplex.*

Lemma 3.2.3. *For any points $z_1, \ldots, z_n \in \mathbb{R}^d$:*

$$L\left(\mathrm{MST}\left(\{z_1, \ldots, z_n\}\right)\right) \leq \frac{2}{n} \sum_{1\leq i<j\leq n} \|z_i - z_j\|_2 .$$

We first prove the result in the case $n = d + 1$.

Proof of Theorem 3.2.1 ($n = d + 1$). Let $z_1^*, \ldots, z_n^* \in \mathbb{S}$ lie at the vertices of a regular d-simplex. Then, for any points $z_1, \ldots, z_n \in B$:

$$\begin{aligned}
L\left(\mathrm{MST}\left(\{z_1, \ldots, z_n\}\right)\right) &\overset{\text{Theorem 3.2.3}}{\leq} \tfrac{2}{n} \textstyle\sum_{1\leq i<j\leq n} \|z_i - z_j\|_2 \\
&\overset{\text{Theorem 3.2.2}}{\leq} \tfrac{2}{n} \textstyle\sum_{1\leq i<j\leq n} \left\|z_i^* - z_j^*\right\|_2 \\
&\overset{\text{Eq. (3.4)}}{=} \tfrac{2}{n} \tfrac{n(n-1)}{2}\, r\sqrt{\tfrac{2(d+1)}{d}} = (n-1)\, r\sqrt{\tfrac{2(d+1)}{d}} \\
&= L\left(\mathrm{MST}\left(\{z_1^*, \ldots, z_n^*\}\right)\right).
\end{aligned}$$

□

We now prove the result in the case $n < d + 1$ is the same modulo some extra technicalities.

Proof of Theorem 3.2.1 (case $n < d + 1$). Let $z_1, \ldots, z_n \in B$, and let $H \subset \mathbb{R}^d$ be an n-dimensional affine space containing $z_1, \ldots, z_n$. Let B_H be the $(n-1)$-dimensional

Euclidean ball $B \cap H$, and $S_H = S \cap H$ its bounding sphere. Finally, let $r_H \leq r$ be the radius of B_H (and of S_H). Inside H, the same calculation as in the case $n = d + 1$ shows that

$$L\left(\mathrm{MST}\left(\{z_1, \ldots, z_n\}\right)\right) \leq L\left(\mathrm{MST}\left(\{z_1^*, \ldots, z_n^*\}\right)\right) = (n-1)\, r_H \sqrt{\frac{2n}{n-1}}$$

for any points $z_1^*, \ldots, z_n^*$ lying at the vertices of a regular $(n-1)$-simplex inscribed in S_H. The quantity on the right-hand side of the equality is bounded above by $(n-1)r\sqrt{\frac{2n}{n-1}}$, which is attained when $r_H = r$, i.e., when the affine subspace H contains the origin, or equivalently, when S is the smallest circumscribing sphere of $z_1^*, \ldots, z_n^*$. □

Proof of Theorem 3.2.3. We fix d and proceed by induction on n.

For $n = 1$ we have:

$$L\left(\mathrm{MST}\left(\{z_1\}\right)\right) = 0 = \frac{2}{1} \sum_{1 \leq i < j \leq n} \|z_i - z_j\|_2 .$$

Assume now that the result holds for all n up to some $n_0 \geq 1$, and let us prove it for $n = n_0 + 1$. Let $e = (z_i, z_j)$ be an edge of $\mathrm{MST}\left(\{z_1, \ldots, z_n\}\right)$ of maximum length. Then, the graph $G = \mathrm{MST}\left(\{z_1, \ldots, z_n\}\right) \setminus \{e\}$ has two connected components C and D, each of which is a tree with less than n vertices. Up to a relabeling of the points of Z, we can assume without loss of generality that $j = i + 1$ and that the vertices of C are the points $z_1, \ldots, z_i$ while the vertices of D are the points $z_{i+1}, \ldots, z_n$. We can then make the following observations:

(i) $C = \mathrm{MST}\left(\{z_1, \ldots, z_i\}\right)$ and $D = \mathrm{MST}\left(\{z_{i+1}, \ldots, z_n\}\right)$. Indeed, otherwise, replacing C by $\mathrm{MST}\left(\{z_1, \ldots, z_i\}\right)$ and D by $\mathrm{MST}\left(\{z_{i+1}, \ldots, z_n\}\right)$, and connecting them with edge e, would yield a spanning tree of $\{z_1, \ldots, z_n\}$ of strictly smaller length than $G \cup \{e\}$, which would contradict the fact that $G \cup \{e\} = \mathrm{MST}\left(\{z_1, \ldots, z_n\}\right)$.

(ii) For all $k \leq i < l$, we have $\|z_k - z_l\|_2 \geq \|z_i - z_{i+1}\|_2$, for otherwise the graph $G \cup \{(z_k, z_l)\}$ would be a spanning tree of $\{z_1, \ldots, z_n\}$ of strictly smaller length than $G \cup \{e\}$, again contradicting the fact that $G \cup \{e\} = \mathrm{MST}\left(\{z_1, \ldots, z_n\}\right)$.

Then, *(i)* and the induction hypothesis imply:

$$\sum_{k<l\leq i} \|z_k - z_l\|_2 \geq \frac{i}{2} E(C), \tag{3.5}$$

$$\sum_{i<k<l} \|z_k - z_l\|_2 \geq \frac{n-i}{2} E(D). \tag{3.6}$$

Meanwhile, *(ii)* implies:

$$\sum_{k\leq i<l} \|z_k - z_l\|_2 \geq i(n-i)\ \|z_i - z_{i+1}\|_2 . \tag{3.7}$$

And since $e = (z_i, z_{i+1})$ is an edge of MST $(\{z_1, \ldots, z_n\})$ of maximum length, we have:

$$\|z_i - z_{i+1}\|_2 \geq \frac{1}{i-1} E(C), \tag{3.8}$$

$$\|z_i - z_{i+1}\|_2 \geq \frac{1}{n-i-1} E(D). \tag{3.9}$$

Hence:

$$\begin{aligned} \sum_{k\leq i<l} \|z_k - z_l\|_2 &\overset{\text{Eq. (3.7)}}{\geq} i(n-i)\ \|z_i - z_{i+1}\|_2 \\ &= \left(\tfrac{n}{2} + \tfrac{(n-i)(i-1)}{2} + \tfrac{i(n-i-1)}{2}\right) \|z_i - z_{i+1}\|_2 \\ &\overset{\text{Eqs. (3.8)-(3.9)}}{\geq} \tfrac{n}{2}\ \|z_i - z_{i+1}\|_2 + \tfrac{n-i}{2} E(C) + \tfrac{i}{2} E(D). \end{aligned} \tag{3.10}$$

It follows:

$$\begin{aligned} \sum_{1\leq k<l\leq n} \|z_k - z_l\|_2 &= \sum_{k\leq i<l} \|z_k - z_l\|_2 + \sum_{k<l\leq i} \|z_k - z_l\|_2 \\ &\quad + \textstyle\sum_{i<k<l} \|z_k - z_l\|_2 \\ &\overset{\text{Eqs. (3.10) and (3.5)-(3.6)}}{\geq} \tfrac{n}{2}\ \|z_i - z_{i+1}\|_2 + \tfrac{n-i}{2} E(C) + \tfrac{i}{2} E(D) \\ &\quad + \tfrac{i}{2} E(C) + \tfrac{n-i}{2} E(D) \\ &= \tfrac{n}{2}\ \|z_i - z_{i+1}\|_2 + \tfrac{n}{2} E(C) + \tfrac{n}{2} E(D) \\ &= \tfrac{n}{2} E\left(\text{MST}\left(\{z_1, \ldots, z_n\}\right)\right). \end{aligned}$$

□

Asymptotic behavior on large samples

We now consider the case where $n > d+1$, focusing specifically on the asymptotic behavior as $n \to \infty$. We analyze the constant C in Equation (2.13), which can be

made independent of the density of the sampling X. This, in particular, allows us to show that uniform and dimension-maximizing densities are asymptotically optimal for $L(\mathrm{MST}(\cdot))$. We fix a compact Riemannian d-manifold, $\mathcal{M}$, equipped with the d-dimensional Hausdorff measure μ.

We first recall Theorem 2.4.4 (Equation (2.14)) from Section 2.4.1 for the reader's convenience:

Theorem (Costa and Hero [40, Corollary 5])**.** *Let X_n be an iid n-sample of a probability measure on $\mathcal{M}$ with density f_X w.r.t. μ. Then, there exists a constant C' independent of f_X and of $\mathcal{M}$ such that:*

$$n^{-\frac{d-1}{d}} \cdot L\left(\mathrm{MST}(X_n)\right) \xrightarrow[n\to\infty]{} C' \int f_X^{\frac{d-1}{d}} \,\mathrm{d}\mu \quad \text{almost surely.}$$

As pointed out by Costa and Hero [40], the limit in Equation (2.14) is related to the *intrinsic Rényi $\frac{d-1}{d}$-entropy*:

$$\varphi_{\frac{d-1}{d}}(f) = \frac{1}{1-\frac{d-1}{d}} \log \int f^{\frac{d-1}{d}} \,\mathrm{d}\mu, \tag{3.11}$$

which is known to converge to the Shannon entropy as $\frac{d-1}{d} \to 1$ [19]. The Shannon entropy, in turn, achieves its maximum at the uniform distribution on compact sets [146]. This result can be shown by directly studying the map $\phi\colon f \mapsto \int f^p \,\mathrm{d}\mu$, which shows that an optimal density function f_X maximizes the dimensionality of the sampling X_n. Given a compact set $K \subseteq \mathcal{M}$, we consider the space $\mathcal{D}_K$ of positive, continuous probability densities f on K.

Proposition 3.2.4. *For any $0 < p < 1$ and any compact set $K \subseteq \mathcal{M}$, the map $\phi|_{\mathcal{D}_K}$ admits a unique maximum at the uniform distribution U_K on K. Furthermore, we have $\phi(U_A) < \phi(U_B)$ for all sets $A, B \subseteq M$ such that $\mu(A) < \mu(B)$.*

Proof. The map ϕ is strictly concave, as the composition of the strictly concave function $x \mapsto x^p$ with the linear map $x \mapsto \int x \,\mathrm{d}\mu$. Since density functions integrate to 1 over K, maximizing ϕ corresponds to an optimization problem under constraint, which can be solved using the Lagrangian:

$$\mathcal{L}_\lambda(f) := \int_K f^p \,\mathrm{d}\mu - \lambda\left(\int_K f \,\mathrm{d}\mu - 1\right),$$
$$\text{with differential } \mathrm{d}\left(\mathcal{L}_\lambda\right)_f(h) = \int_K \left(pf^{p-1} - \lambda\right) h \,\mathrm{d}\mu.$$

Then, $\mathrm{d}\left(\mathcal{L}_\lambda\right)_f = 0$ is equivalent to $pf^{p-1} - \lambda$ being 0 almost everywhere, i.e. f being equal to $\left(\frac{\lambda}{p}\right)^{\frac{1}{p-1}}$ (a constant determined by $\int f \,\mathrm{d}\mu = 1$) almost everywhere, i.e. f being the density of the uniform measure on K.

Now, recalling that $0 < p < 1$, we have:

$$\phi(U_K) = \int \left(\frac{\mathrm{d}U_K}{\mathrm{d}\mu}\right)^p \mathrm{d}\mu = \int_K \frac{1}{\mu(K)^p}\,\mathrm{d}\mu = \int \frac{\mu(K)}{(\mu(K))^p}\,\mathrm{d}U_K = \mu(K)^{1-p},$$

which proves the second part of the statement. □

A direct consequence of Theorem 3.2.4 and the fact that $\mathcal{D}_{\mathcal{M}}$ is dense in the set of probability densities of $L^p(\mathcal{M})$ is the following corollary.

Corollary 3.2.5. *Let $\mathcal{D}$ be the set of probability densities over $\mathcal{M}$, and $p \in (0,1)$. Then, the map $f \in \mathcal{D} \mapsto \int f^p\,\mathrm{d}\mu$ reaches its maximum at the density $f := \frac{\mathrm{d}U_{\mathcal{M}}}{\mathrm{d}\mu}$.*

3.2.2 Uniformity properties

Recall from Equation (3.1) that $\mathcal{L}_{\mathrm{E}}(Z) = -L\left(\mathrm{MST}(Z)\right)/|Z|$. In practice, for datasets of fixed size, minimizing $\mathcal{L}_{\mathrm{E}}$ is equivalent to minimizing $-L\left(\mathrm{MST}(Z)\right)$ itself. In this section, we show that, up to a renormalization, $-L\left(\mathrm{MST}(Z)\right)$ satisfies the four principled properties for uniformity metrics introduced in [63, Section 3.1].

A *uniformity metric* $\mathcal{U}$ is a scalar score function on n-samples in $\mathbb{R}^d$ that is *large* on uniform-like point clouds and *small* on degenerate or almost-degenerate point clouds. In the following, we fix an n-sample $Z = (z_1, \ldots, z_n)$ in $\mathbb{R}^d$.

We define our uniformity metric as $-L\left(\mathrm{MST}(\cdot)\right)$ normalized by the edge length of the regular d-simplex σ_d^0 with vertices on the unit sphere $\mathbb{S}^{d-1}$:

$$\mathcal{U}_{\text{T-REG}}(Z) := -\frac{L\left(\mathrm{MST}(Z)\right)}{\left(\frac{2(d+1)}{d}\right)^{1/2}}. \tag{3.12}$$

In practice, for a fixed ambient dimension d, minimizing $-L\left(\mathrm{MST}(Z)\right)$ or minimizing $\mathcal{U}_{\text{T-REG}}(Z)$ is equivalent. The motivation behind renormalization is to measure the length of the minimum spanning tree relative to a *reference* length on the unit sphere $\mathbb{S}^{d-1}$, to make it a dimensionless quantity.

We now show that our uniformity score $\mathcal{U}_{\text{T-REG}}$ satisfies the desired uniformity properties:

1. *Instance permutation constraint:* $\forall \pi \in \mathfrak{S}_n, \mathcal{U}\left((z_{\pi_1}, \ldots, z_{\pi_n})\right) = \mathcal{U}(Z)$.
   By construction, $\mathrm{MST}(Z)$ and its length are invariant under permutations of the points' indices, therefore $L\left(\mathrm{MST}(\cdot)\right)$ and $\mathcal{U}_{\text{T-REG}}$ are also permutation invariant.

2. *Instance cloning constraint:* if $Z' = Z$, then $\mathcal{U}((z_1, \ldots, z_n, z'_1, \ldots, z'_n)) = \mathcal{U}(Z)$.
   If $G = (Z, E)$ is an MST of Z, then

$$G' = ((z_1, \ldots, z_n, z'_1, \ldots, z'_n), E \cup \{(z_i, z'_i) : 1 \leq i \leq n\}) \tag{3.13}$$

   is a spanning tree of $(z_1, \ldots, z_n, z'_1, \ldots, z'_n)$, and it is minimal for E since G itself is minimal on Z and $E(G') = E(G) + n \times 0^1 = E(G)$. Therefore, $\mathcal{U}_{\text{T-REG}}((z_1, \ldots, z_n, z_1, \ldots, z_n)) = \mathcal{U}_{\text{T-REG}}(Z)$ since the ambient dimension is constant.

3. *Feature cloning constraint:* $\mathcal{U}(z_1 \oplus z_1, \ldots, z_n \oplus z_n) < \mathcal{U}(z)$.
   Feature cloning corresponds to pushing the points of Z to the diagonal in $\mathbb{R}^{2d}$, which impacts the pairwise distances by a uniform scaling by a factor of $\sqrt{2}$. In particular, the MST remains the same combinatorially and we have:

$$-L\left(\text{MST}(Z \oplus Z)\right) = -2^{1/2} L\left(\text{MST}(Z)\right) < -L\left(\text{MST}(Z)\right).$$

   Meanwhile, since $\varphi \colon d \mapsto \left(\frac{2(d+1)}{d}\right)^{-1/2}$ is an increasing function and $\mathcal{U}_{\text{T-REG}} \leq 0$, we have:

$$\mathcal{U}_{\text{T-REG}}(Z \oplus Z) = -L\left(\text{MST}(Z \oplus Z)\right) \varphi(2d) < -L\left(\text{MST}(Z)\right) \varphi(d) = \mathcal{U}_{\text{T-REG}}(Z).$$

4. *Feature baby constraint:* $\forall k \in \mathbb{N}_+, \mathcal{U}(Z \oplus \mathbf{0}^k) < \mathcal{U}(Z)$.
   Adding constant features does not impact the pairwise distances, hence does not impact the minimum spanning tree and its length We thus have $L\left(\text{MST}\left(Z \oplus \mathbf{0}^k\right)\right) = L\left(\text{MST}(Z)\right)$.

   As in the previous case, since $\varphi \colon d \mapsto \left(\frac{2(d+1)}{d}\right)^{-1/2}$ is an increasing function and $\mathcal{U}_{\text{T-REG}} \leq 0$, we have:

$$\mathcal{U}_{\text{T-REG}}(Z \oplus \mathbf{0}^k) = -L\left(\text{MST}\left(Z \oplus \mathbf{0}^k\right)\right) \varphi(d+k) < -L\left(\text{MST}(Z)\right) \varphi(d) = \mathcal{U}_{\text{T-REG}}(Z).$$

3.2.3 Empirical evaluation

We conduct empirical studies on synthetic data to validate T-REG's ability to prevent dimensional collapse and promote sample uniformity.

Preventing dimensional collapse

Following Fang et al. [63], we assess T-REG's effectiveness against dimensional collapse by measuring how sensitive its loss $\mathcal{L}_{\text{E}}$ is to simulated collapse. Specifically, we generate $10,000$ data points in dimension 1024 from an isotropic Gaussian

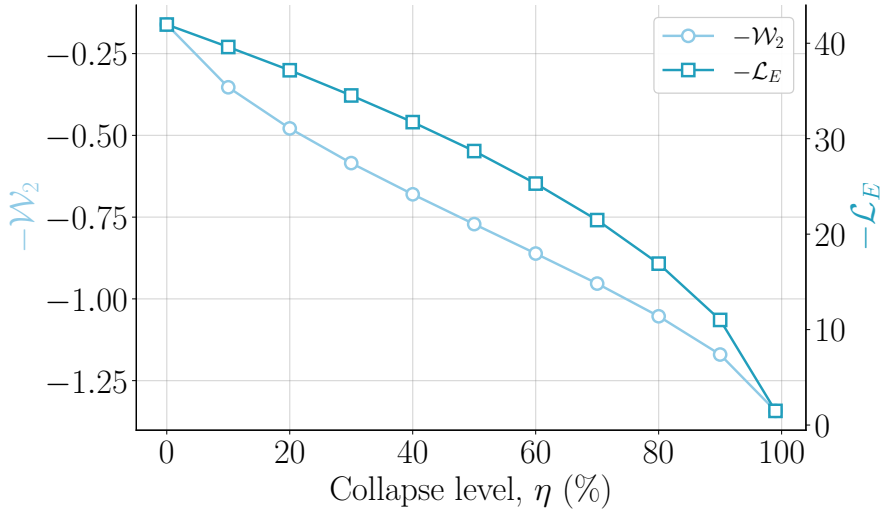


Figure 3.4: **Sensitivity to dimensional collapse.** Collapse is simulated by zeroing out a fraction η of the coordinates of 10,000 points drawn from an isotropic Gaussian in $\mathbb{R}^{1024}$. Both $-\mathcal{W}_2$ and $-\mathcal{L}_{\mathrm{E}}$ decrease as η increases, showing that $\mathcal{L}_{\mathrm{E}}$ is as sensitive to dimensional collapse as the $\mathcal{W}_2$ loss of Fang et al. [63].

distribution, then zero out a fraction η of their coordinates to control the collapse level.

As shown in Figure 3.4, the sensitivity of the T-REG loss to η is similar to that of the $\mathcal{W}_2$ loss from Fang et al. [63], indicating that T-REG effectively penalizes dimensional collapse.

Promoting sample uniformity

We apply T-REG alone to optimize a given point cloud and analyze its behavior in both low-dimensional and high-dimensional scenarios (Figure 3.3).

For this, we use two different input point clouds: *(i)* a degenerate set of 256 points on a 1-d curve (corresponding to the orange dots in Figures 3.3a and 3.3b); *(ii)* a set of 256 points in $\mathbb{R}^{256}$, initially concentrated around a specific point on the unit sphere, where each point x_i is sampled as $x_i = e_1 + \varepsilon_i$, with ε_i drawn uniformly from a ball of radius 0.001.

As illustrated in Figure 3.3a, optimization with T-REG successfully transforms the initial point cloud in a 3-d space into a uniformly distributed point cloud on the sphere, as per Corollary 3.2.5. This is achieved through the combination of MST length maximization and sphere constraint. The sphere constraint is crucial here: when optimizing only the MST length, $\mathcal{L}_{\mathrm{E}}$, (see Figure 3.3b), the optimization

fails to converge.

In high dimensions (Figure 3.3d), we analyze the distribution of cosine similarities between the embeddings. The initial distribution shows a sharp peak near 1, indicating highly correlated samples on the sphere. After optimization with T-REG, the distribution becomes almost a Dirac slightly below 0, indicating that the configuration of the points is close to that of the vertices of the regular simplex, as per Theorem 3.2.1.

3.3 T-REGS: T-REG for Self-supervised learning

T-REGS extends T-REG to Joint-Embedding Self-Supervised Learning. For an input image i, two transformations t, t' are sampled from a distribution $\mathcal{T}$ to produce two augmented views $x = t(i)$ and $x' = t'(i)$. These transformations are typically random crops and color distortions. We compute embeddings $z = h(f(x))$ and $z' = h(f(x'))$ using a backbone f and projector h. T-REGS acts as a regularization applied separately to the embedding batches $Z = [z_1, ..., z_n]$ and $Z' = [z'_1, ..., z'_n]$.

Specifically, embeddings from each view batch, Z and Z', are treated as points in a high-dimensional space, and Kruskal's algorithm [108] is used to construct two Minimum Spanning Tree (MST), one for each view batch. These MSTs yield two T-REG regularization terms, which are combined into the T-REGS objective as follows:

$$\mathcal{L}_{\text{T-REGS}}(Z, Z') = \underbrace{\gamma\,\mathcal{L}_{\text{E}}(Z) + \lambda\,\mathcal{L}_{\text{S}}(Z)}_{\mathcal{L}_{\text{T-REG}}(Z)} + \underbrace{\gamma\,\mathcal{L}_{\text{E}}(Z') + \lambda\,\mathcal{L}_{\text{S}}(Z')}_{\mathcal{L}_{\text{T-REG}}(Z')}. \tag{3.14}$$

where γ, λ control the contribution of each term.

In practice, T-REGS can be used as *(i)* a standalone regularization, combined directly with an invariance term such as the Mean Squared Error: $\mathcal{L}(Z, Z') = \beta\,\mathcal{L}_{\text{MSE}}(Z, Z') + \mathcal{L}_{\text{T-REGS}}(Z, Z')$, where $\mathcal{L}_{\text{MSE}}(Z, Z') = \frac{1}{n}\sum_i \|z_i - z'_i\|_2^2$ and β is a mixing parameter; or *(ii)* as an auxiliary loss to existing SSL methods: $\mathcal{L}(Z, Z') = \beta\,\mathcal{L}_{\text{SSL}}(Z, Z') + \mathcal{L}_{\text{T-REGS}}(Z, Z')$, where $\mathcal{L}_{\text{SSL}}(Z, Z')$ denotes the objective function of a given SSL method, and β is a mixing parameter. An overview of T-REGS is presented in Figure 3.2.

Algorithm 2 summarizes the full T-REGS training loop, combining the view-invariance loss with the MST-length and sphere-constraint regularizers.

Algorithm 2 T-REGS combined with view invariance using PyTorch pseudocode

```
# f:  encoder network, h:  projection network
# β, γ, λ:  coefficients of the invariance, MST length and sphere
constraint losses
for x in loader: do # load a batch with N samples
   # two randomly augmented versions of x
   x_1, x_2 = augment(x), augment(x)
   # compute the representations and the embeddings
   y_1, y_2 = f(x_1), f(x_2)
   z_1, z_2 = h(y_1), h(y_2)
   inv_loss = L_MSE(z_1,z_2) # invariance loss
   length_mst_loss = L_E(z_1) + L_E(z_2) # MST length
   sphere_loss = L_S(z_1) + L_S(z_2) # soft sphere constraint loss
   # total loss
   loss = β inv_loss + γ length_mst_loss + λ sphere_loss
   # optimization step
   loss.backward()
   optimizer.step()
end for
```

The remainder of the section provides evaluations of T-REGS on standard SSL benchmarks in Section 3.3.1. Further analyses and implementation details are provided in Sections 3.4 and 3.5.

3.3.1 Evaluation on standard SSL benchmark

We evaluate the representations obtained after training with T-REGS, either directly combined with view invariance or integrated with existing methods (i.e., BYOL, and Barlow Twins) on CIFAR-10/100 [106], ImageNet-100 [187], and ImageNet [47]. Our implementation is based on `solo-learn` [43], and we use `torchph` [20] for the MST computations. For T-REGS as a standalone regularizer, we use $\beta = 10$, $\gamma = 0.2$, $\lambda = 8e - 4$.

Evaluation on CIFAR-10/100

We first focus on comparisons with Fang et al. [63] ($\mathcal{W}_2$-regularized methods), following the same protocol on CIFAR-10/100 [106] with ResNet-18. As shown in Table 3.1, T-REGS demonstrates strong standalone performance, achieving results within 0.1% of the best $\mathcal{W}_2$-regularized approach on CIFAR-10. Additionally, using T-REGS as an auxiliary loss consistently improves performance over the respective baselines, and over variants that use $\mathcal{L}_u$ or $\mathcal{W}_2$ as additional regularization

Method		CIFAR-10	CIFAR-100
Zero-CL		91.3	**68.5**
	+ $\mathcal{L}_u$	91.3	68.4
	+ $\mathcal{W}_2$	91.4	**68.5**
MoCo v2		90.7	60.3
	+ $\mathcal{L}_u$	91.0	61.2
	+ $\mathcal{W}_2$	91.4	63.7
BYOL		89.5	63.7
	+ $\mathcal{L}_u$	90.1	62.7
	+ $\mathcal{W}_2$	90.1	65.2
	+ $\mathcal{L}_{\text{T-REGS}}$	90.4	65.7
Barlow Twins		91.2	68.2
	+ $\mathcal{L}_u$	91.4	68.4
	+ $\mathcal{W}_2$	91.4	**68.5**
	+ $\mathcal{L}_{\text{T-REGS}}$	**91.8**	**68.5**
$\mathcal{L}_{\text{MSE}}$	+ $\mathcal{L}_{\text{T-REGS}}$	91.3	67.4

Table 3.1: **Comparison with $\mathcal{W}_2$ regularization [63] on CIFAR-10/100.** The table is inherited from Fang et al. [63], and we follow the same protocol: ResNet-18 models are pre-trained for 500 epochs on CIFAR-10/100, with a batch size of 256, followed by linear probing. We report Top-1 accuracy (%). **Boldface** indicates best performance.

terms.

Evaluation on ImageNet-100/1k

To assess the scalability of T-REGS, we evaluate our model on ImageNet-100 and ImageNet-1k using ResNet-18 and ResNet-50, respectively, following the standard linear evaluation protocol on ImageNet and comparing with the state of the art. We report Top-1 accuracy. As shown in Table 3.2, T-REGS is competitive with methods that use the same number of views (e.g., INTL), and improves existing methods when used as an auxiliary loss.

3.4 Ablation Studies and Analyses of T-REGS

In this section, we conduct a comprehensive set of ablation experiments to assess the robustness and versatility of T-REGS. These experiments cover various aspects, including projector architecture, batch sizes, and seed variability and computational cost.

# views	Method		Imagenet-100 Top-1	ImageNet-1k Batch Size	ImageNet-1k Top-1
8	SwAV		74.3	4096	66.5
	FroSSL		79.8	-	-
	SSOLE		**82.5**	256	**73.9**
2	SimCLR		77.0	4096	66.5
	MoCo v2		79.3	256	67.4
	SimSiam		78.7	256	68.1
	W-MSE		69.1	512	65.1
	Zero-CL		79.3	1024	**68.9**
	VICReg		79.4	1024	68.3
	CW-RGP		77.0	512	67.1
	INTL		**81.7**	512	**69.5**
	BYOL		80.3	1024	66.5
		+ $\mathcal{L}_{\text{T-REGS}}$	**80.8**	1024	67.2
	Barlow Twins		80.2	2048	67.7
		+ $\mathcal{L}_{\text{T-REGS}}$	**80.9**	2048	67.8
	$\mathcal{L}_{\text{MSE}}$	+ $\mathcal{L}_{\text{T-REGS}}$	80.3	512	**68.8**

Table 3.2: **Linear Evaluation on ImageNet-100/1k.** We report Top-1 accuracy (%). The top-4 methods are **boldfaced**. For ImageNet-100, ResNet-18 are pre-trained for 400 epochs using a batch size of 256; for ImageNet-1k, ResNet-50 is pre-trained for 100 epochs. The table is mostly inherited from Weng et al. [206].

3.4.1 Loss coefficients

We determine the final coefficients for T-REGS as a standalone regularizer on ImageNet-1k as follows (the same approach was applied when combining T-REGS with existing methods). Initial experiments revealed that maintaining $\beta \geq \gamma \geq \lambda$ was essential to prevent representation collapse. To efficiently explore the parameter space while managing computational costs, we fixed β (the largest coefficient) and systematically varied the ratios $\frac{\beta}{\gamma}$ and $\frac{\gamma}{\lambda}$ using 50-epoch online probing [67]. As shown in Table 3.3, both $\mathcal{L}_{\text{E}}$ and $\mathcal{L}_{\text{S}}$ contribute to performance, with $\mathcal{L}_{\text{S}}$ requiring a smaller weight. This suggests that the actual radius of the sphere is not critical, thereby validating our choice of a soft sphere constraint instead of a hard one.

Coefficients			Scaling		
β	γ	λ	$\frac{\beta}{\gamma}$	$\frac{\gamma}{\lambda}$	**Top-1**
1	-	-	-	-	collapse
1	1	-	1	-	collapse
10	1	1	10	1	collapse
10	0.5	5e-2	20	10	25.7
10	0.2	2e-2	50	10	45.4
10	0.5	2.5e-3	20	200	65.0
10	0.2	1e-3	50	200	65.3
10	0.5	2e-3	20	250	64.9
10	0.2	8e-4	50	250	**66.1**
10	0.02	8e-5	100	300	63.3

Table 3.3: **Impact of coefficients.** $\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{T-REGS}}$ top-1 accuracy (%) on ImageNet-1k with online evaluation protocol over 50 epochs. **Boldface** indicates best performance.

Method	**Complexity**	B **range**	D **range**	**Wall-clock time**
SimCLR [34]	$\mathcal{O}(B^2 \cdot D)$	[2048-4096]	[256-1024]	0.22 ± 0.03
VICReg [12]	$\mathcal{O}(B \cdot D^2)$	[1024-4096]	[4096-8192]	0.23 ± 0.02
$\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{T-REGS}}$	$\mathcal{O}(B^2(D \cdot \log B))$	[512-1024]	[512-2048]	0.20 ± 0.001

Table 3.4: **Complexity and computational cost.** Comparison between different methods is performed, with training on ImageNet-1k distributed across 4 Tesla H100 GPUs. The wall-clock time (sec/step) is averaged over 500 steps. B, D ranges are reported from Bardes, Ponce, and Lecun [12] and Garrido et al. [67].

3.4.2 Computational overhead of T-REGS

We evaluate the computational cost of T-REGS. The MSTs are computed with Kruskal's algorithm [108], whose worst-case time is $\mathcal{O}(B^2(D \cdot \log B))$, with B the batch size and D the embedding dimension. Although Kruskal's main loop is sequential, preprocessing (computing the distance matrix and sorting its entries) dominates in practice and can be efficiently parallelized on GPUs (as in `torchph` [20], used in our implementation). Empirically, T-REGS matches the per-step wall-clock of VICReg and SimCLR (averaged over 500 steps during training on ImageNet-1k with $B = 512, D = 1024$; see Table 5.2).

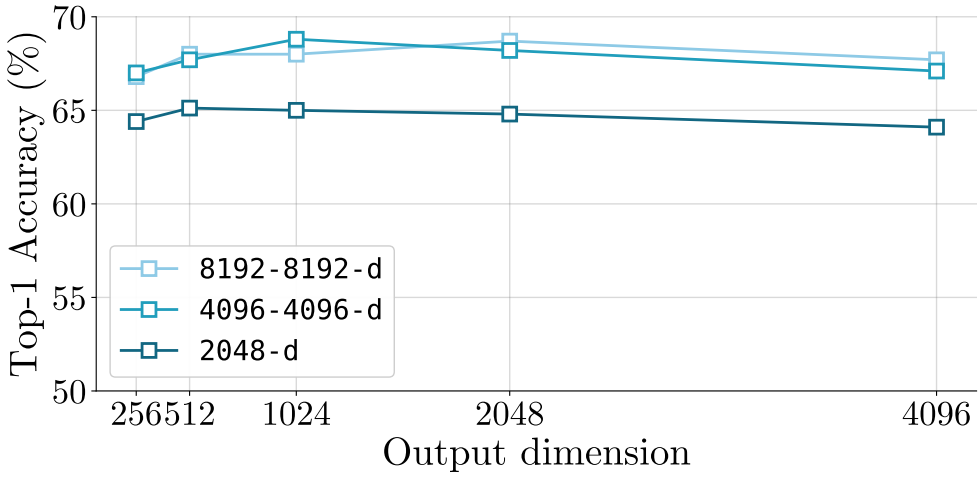


Figure 3.5: **Impact of the projector architecture.** $\mathcal{L}_{\text{MSE}}$ +$\mathcal{L}_{\text{T-REGS}}$ top-1 accuracy (%) on the linear evaluation protocol with 100 pretraining epochs.

3.4.3 Projector architecture

An essential difference between methods lies in how the projector (h_ϕ, in Figure 3.2) is designed [68]. To assess the impact of the projector architecture on T-REGS performance (when combined with view invariance), we train models for 100 epochs on ImageNet-1k with different projector architectures. We describe projector architectures using the notation `X-Y-Z`, where each number represents the dimension of a linear layer in sequence. Each layer (except the last) is followed by a `ReLU` activation and batch normalization. The final layer has no activation, batch normalization, or bias. We evaluate three projector architectures:

- `2048-d`: the projector used in SimCLR
- `8192-8192-d`: the projector used in VICReg
- `4096-4096-d`: a smaller variant of the VICReg projector

with `d` varying from 256 to 4096. The remaining hyperparameters are the same as reported in Section 3.5. Our results demonstrate that the choice of projector architecture significantly impacts the model's performance, with `8192-8192-d` and `4096-4096-d` consistently outperforming the `2048-d` variant.

We also observe that the embedding dimension `d` (i.e., the output dimension of the projector h_ϕ) has minimal impact on performance. This is particularly noteworthy as other methods, such as VICReg [12] and Barlow Twins [213], are known to be sensitive to embedding dimension, typically requiring dimensions larger than

2048 for optimal performance. Our results show that T-REGS maintains high accuracy even with small embedding dimensions, demonstrating its robustness to this hyperparameter.

3.4.4 Batch size

Batch Size	128	256	512	1024
Top-1	66.3	67.2	68.7	68.0

Table 3.5: **Impact of batch size.** $\mathcal{L}_{\text{MSE}}$ +$\mathcal{L}_{\text{T-REGS}}$ top-1 accuracy (%) using linear evaluation after 100 pre-training epochs.

Many SSL methods are known to be sensitive to batch sizes, for instance contrastive methods suffer from the need of many negative samples which can translate into the need of large batch sizes [34, 25]. In Table 3.5, we study how batch size affects T-REGS performance (when combined with view invaraince). We train models for 100 epochs on ImageNet-1k with batch sizes ranging from 128 to 1024, using the same hyperparameters as in Table 3.7. T-REGS maintains good performance even with small batch sizes (e.g., 128), demonstrating its robustness to different batch size configurations.

3.4.5 Normalization

	CIFAR-10	CIFAR-100
soft-constraint	91.2	66.8
hard-constraint	89.2	64.7

Table 3.6: **Impact of normalization.** $\mathcal{L}_{\text{MSE}}$ +$\mathcal{L}_{\text{T-REGS}}$ top-1 accuracy (%).

Some popular SSL frameworks, such as SimCLR and BYOL, employ explicit normalization of features to the unit sphere, enforcing a hard constraint. Others do not implement such constraints explicitly, and instead rely on soft mechanisms—such as VICReg, which incorporates variance and covariance regularization terms in its loss, implicitly constraining the distribution (and to some extent, the norms) of the embeddings. We chose a soft sphere constraint for the following reasons: *(i)* a hard normalization has been shown to ignore the importance of the embedding norm for gradient computation, whereas a soft constraint enables better embedding optimization [215, 216], *(ii)* during our initial experiment, we found that relaxing the sphere constraint from a hard one to a soft one provides more leeway for optimization and leads to improved results, as shown in Table 3.6.

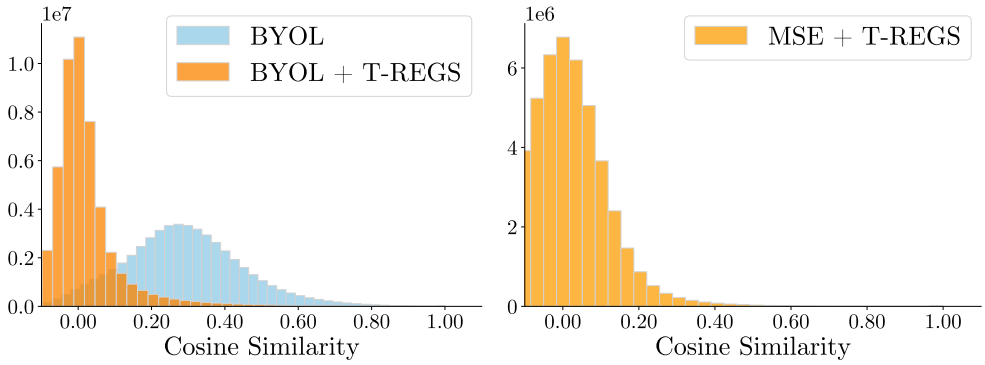


Figure 3.6: **Histograms of embeddings' cosine similarities on CIFAR-10.** With T-REGS as a standalone regularization (orange) or as an auxiliary loss (dark orange), the distribution of pairwise cosine similarities becomes concentrated around zero, indicating that the embeddings are highly decorrelated and approach a regular simplex configuration (Theorem 3.2.1).

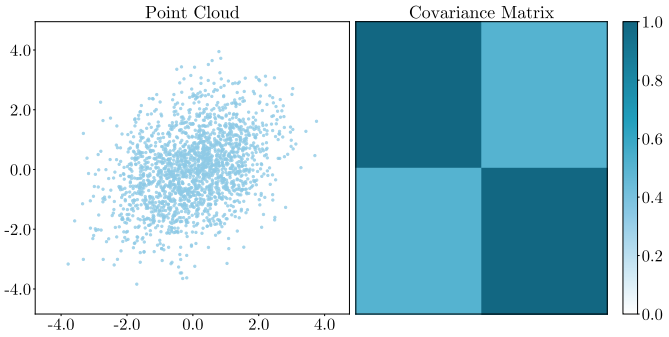
3.4.6 Study of the embeddings

We analyze the learned embeddings by computing pairwise cosine similarity between embeddings on CIFAR-10. As shown in Figure 3.6, BYOL yield embeddings with mean similarities significantly above zero (≈ 0.3). This indicates a concentration within a cone rather than uniformity on the hypersphere. In contrast, T-REGS yields mean similarities near zero, reflecting more uniformly distributed and decorrelated embeddings, with some values slightly negative– suggesting an arrangement close to a regular simplex, as per Theorem 3.2.1. Additionally, applying T-REGS as an auxiliary loss effectively shifts the mean cosine similarity towards zero, as illustrated in Figure 3.6, thus indicating its effectiveness.

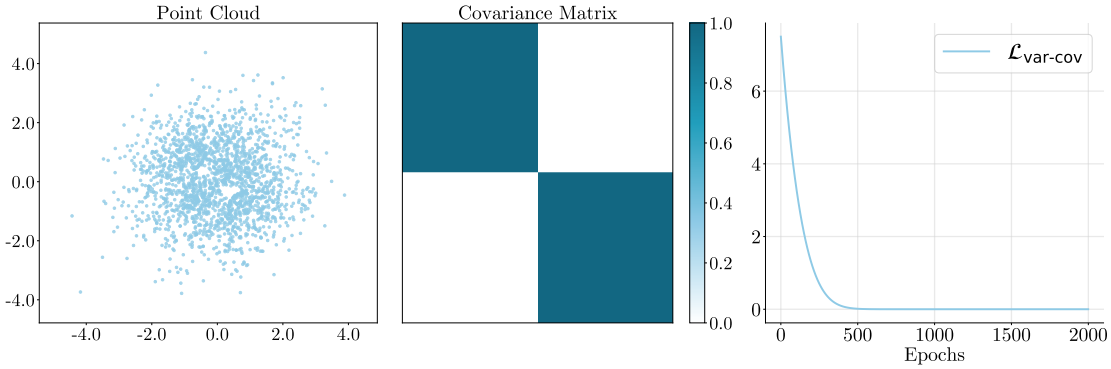
3.4.7 Study of redundancy-reduction methods

As discussed in Section 2.2.4, redundancy-reduction methods [12, 213, 206, 60] aim to produce embedding dimensions that are mutually decorrelated. They do so by maximizing the informational content of embeddings through regularization of their empirical covariance matrix.

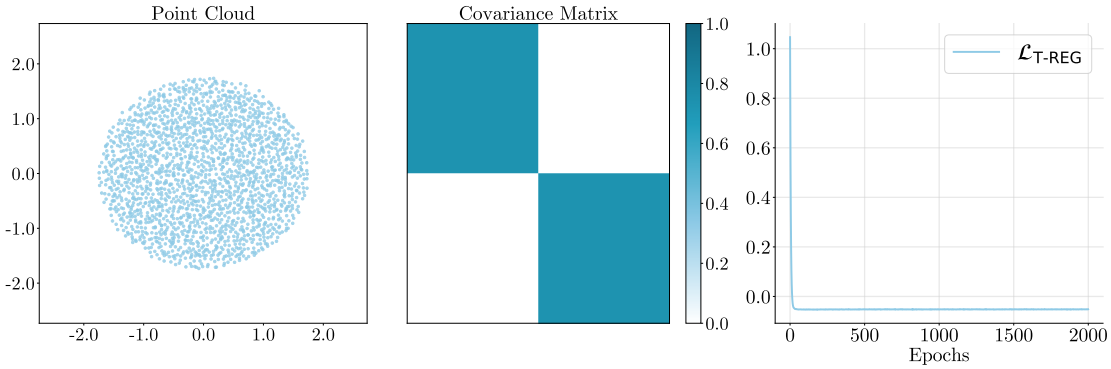
For instance, VICReg [12] combines *(i)* a variance term that encourages the standard deviation of each embedding dimension within a batch to remain above 1, preventing dimensional collapse; and *(ii)* a covariance regularization that penalizes the off-diagonal entries of the empirical covariance matrix, decorrelating the different embedding dimensions. Their combined regularization (see Section 2.2.4



(a) Initial point cloud: sampling of a non-isotropic Gaussian measure



(b) Optimization w/ $\mathcal{L}_{\text{var-cov}}$



(c) Optimization w/ $\mathcal{L}_{\text{T-REG}}$ on the ball

Figure 3.7: **Limitations of redundancy-based methods for non-isotropic Gaussian measure.** (a) Initial sampling from a non-isotropic Gaussian distribution. (b) After $\mathcal{L}_{\text{var-cov}}$ optimization: despite achieving a near-identity covariance matrix (center), the point cloud remains concentrated around its mean, with visible artifacts: holes (left). (c) $\mathcal{L}_{\text{T-REG}}$ optimization achieves both uniform distribution on the disk and near-identity correlation matrix.

and Eq. (2.8)) writes:

$$\mathcal{L}_{\text{var-cov}} = \nu\,\mathcal{L}_{\text{var}} + \tau\,\mathcal{L}_{\text{cov}} \tag{3.15}$$

where ν, τ are scalar hyperparameters.

Limitations of redundancy-reduction methods. In Figure 3.7 we study a limitation of redundancy-reduction methods. We sample 2000 points from a non-isotropic Gaussian distribution, and observe the resulting point cloud after optimization with $\mathcal{L}_{\text{var-cov}}$ (using $\nu = 25, \tau = 1$, as in Bardes, Ponce, and Lecun [12]). Since Gaussian distributions are fully characterized by their mean and covariance matrix, we expect that by optimizing the empirical covariance matrix, the initial point cloud will converge towards a sample of the standard Gaussian distribution, thus far from a uniform distribution.

Figure 3.7b shows the optimization of Figure 3.7a using the covariance-based approach from VICReg (Equation (3.15)). The optimized empirical distribution is, as expected, closer to the standard Gaussian distribution, but exhibits artifacts such as low-dimensional concentration points or even holes suggesting local instabilities. These artifacts persist across different parameter choices, though their specific manifestations may vary with the learning rate.

Figure 3.7c shows the optimization of Figure 3.7a using $\mathcal{L}_{\text{T-REG}}$. We leverage Theorem 3.2.5, whose guarantees are valid for any Riemannian manifold, and in particular the disk. This ensures that the limiting distribution for T-REG is the uniform distribution on the disk, which prevents dimensional collapse and naturally decorrelates the distribution while maintaining the uniform distribution.

3.5 Implementation Details

Our implementation is available at `t-regs`. It is based on `solo-learn` [43], which is released under the MIT License. To compute the length of the minimum spanning tree, we rely on `torchph` and `Gudhi` [150], both released under MIT Licenses.

We follow the guidance of Da Costa et al. [43] for selecting baseline hyperparameters and use the same seed: 5. Table 3.7 lists each dataset's architectural and training details.

3.6 Computational Cost

Experiments were conducted on NVIDIA H100 and V100 GPUs. Training a single T-REGS model requires:

	CIFAR-10	CIFAR-100	ImageNet-100	ImageNet-1k
Backbone				
Backbone	ResNet-18	ResNet-18	ResNet-18	ResNet-50
Projector				
Layers	3 layers with BN and ReLU			
Hidden dim.	2048	2048	4096	4096
Output dim.	1024			
Pre-training				
Batch size	256	256	256	512
Optimizer	LARS			
Learning rate	base_lr $\times$ batch size/256			
base_lr	0.4	0.4	0.3	1
LR warm-up	2 epochs	2 epochs	2 epochs	10 epochs
LR schedule	cosine decay			
Weight decay	10^{-4}	10^{-4}	10^{-4}	10^{-5}
Linear evaluation				
Batch size	256			
Optimizer	SGD			
base_lr	0.1			
LR schedule	cosine decay			

Table 3.7: **Training hyperparameters.**

- **ImageNet-1k:** 15 hours using 4 H100 GPUs (totaling 60 GPU-hours);
- **ImageNet-100:** 7 hours using 1 H100 GPU;
- **CIFAR-10/100:** 7 hours using 1 V100 GPU.

The total computational resources consumed for this project, including all baseline evaluations, hyperparameter tuning, and ablation studies, amounted to approximately 10,000 H100 and 5,000 V100 GPU-hours.

3.7 Discussion

While effective on the settings considered, T-REGS has several limitations.

The theoretical analysis characterizes maximizers of the MST length for a point

cloud on a compact set (Theorems 3.2.1 and 3.2.5). These results describe the geometric bias of T-REG; they do not account for its joint optimization with view-invariance through a deep encoder. In practice, preventing collapse still requires a careful balance of loss coefficients (Table 3.3).

Additionally, the empirical study is restricted to standard vision JE-SSL benchmarks with ResNet backbones and linear probing. Broader validation in the multimodal setting is developed in Chapter 4.

3.8 Conclusion

In this chapter, we introduced T-REG, a regularization approach that prevents dimensional collapse and promotes sample uniformity. Our method maximizes the length of the minimum spanning tree (MST), coupled with a sphere constraint.

Our analysis connects MST optimization to entropy maximization and uniformity on compact manifolds, providing theoretical guarantees corroborated by empirical results.

We extend T-REG to Self-supervised learning, yielding T-REGS. On CIFAR-10/100 and ImageNet-100/1k, T-REGS is competitive with $\mathcal{W}_2$-regularized and state-of-the-art methods, both as a standalone regularizer and as an auxiliary term, underscoring its effectiveness.

Chapter 4

Anchored Multimodal Representations for Physiological Change Detection under Missing Modalities

Building upon joint-embedding architectures, this chapter presents a concrete application of T-REGS (introduced in Chapter 3) to multimodal learning in the medical domain, where heterogeneous sources, e.g., imaging, physiological signals, and clinical records, are combined to build richer representations, demonstrating the broader relevance of T-REGS beyond standard vision benchmarks.

We begin by introducing the problem (Section 4.1) and providing background on handling missing modalities (Section 4.2), before discussing the use of T-REGS for multimodal data (Section 4.3) and presenting ADAPT (Section 4.4). We then report experimental results (Section 4.5).

This chapter is an updated version of
ADAPT: Multimodal Learning for Detecting Physiological Changes under Missing Modalities
J. Mordacq, L. Milecki, M. Vakalopoulou, S. Oudot, V. Kalogeiton
MIDL 2024 & CVPR-W 2024

It has been reworked to leverage T-REGS to close the modality gap during anchoring (Chapter 3).

4.1 Introduction

Monitoring physiological responses to external stimuli is crucial for assessing individual well-being in medical and safety-critical settings. Relevant examples include drowsiness during driving, a physiological state that contributes to road accidents and fatalities [177], and stress, a response to emotional, mental, and physical challenges [162] that can trigger or aggravate various pathological conditions [49]. Such responses manifest across multiple channels, they can be captured visually (videos), acoustically (audio), or through biomedical signals (e.g., electrocardiograms), naturally motivating a multimodal approach.

Yet this multimodal setting introduces three intertwined challenges: balancing the contributions of heterogeneous modalities under limited data, ensuring robustness when modalities are absent, and closing the *modality gap* [121], a form of collapse where modality embeddings occupy separate subspaces rather than spanning a shared representational space, which limits the quality and transferability of learned representations. Existing methods address these challenges only partially, suffering from *(i)* bias toward the most available modalities, which leads to sub-optimal performance [103], *(ii)* dependence on complete modalities during training [129, 32], *(iii)* limited generalizability to more than two modalities [128, 127], and *(iv)* reliance on a shared encoder for same-dimensional inputs, which complicates extension to heterogeneous modalities such as imaging and biomedical signals [103].

To address these issues, we introduce the Anchored multimodal Physiological Transformer (ADAPT), designed to operate robustly under missing modalities during both training and inference. ADAPT comprises two key components. First, we embed all modalities into a shared feature space by aligning each to a single frozen reference modality, the *anchor*, via a contrastive objective, achieving linear rather than quadratic scaling in the number of modalities. Yet contrastive alignment alone is prone to the *modality gap*, degrading cross-modal representations. To counteract this, we apply T-REGS (Chapter 3) independently to each modality's embeddings as an auxiliary loss, promoting uniformity across the full space and

preventing modality-specific dimensional collapse. We refer to this stage as *anchoring*. Second, ADAPT leverages a Masked Multimodal Transformer that exploits inter- and intra-modality correlations to produce a unified representation. Following Ma et al. [127] and Milecki et al. [133], we adopt masked self-attention [196]: when a modality is unavailable, its feature representation is masked, ensuring flexibility under arbitrary missing patterns.

ADAPT is evaluated on the detection of stress triggered by specific stimuli, using the multimodal StressID dataset [29], with missing modalities both during training and inference.

Our main contributions can be summarized as follows:

i) We introduce ADAPT, a modular anchoring-then-fusion framework that aligns all modalities to a common rich feature space using T-REGS to prevent the modality gap (Sections 4.3 and 4.4).

ii) We propose a masked-attention fusion strategy that handles missing modalities during both training and inference (Section 4.4).

iii) We demonstrate state-of-the-art results on StressID with extensive evaluations highlighting ADAPT's superiority (Section 4.5).

4.2 Background: Multimodal Learning under Missing Modalities

4.2.1 Handling missing modalities

Missing modalities pose a persistent challenge in multimodal learning, particularly in medical imaging, due to privacy concerns or impractical data acquisition [120, 10]. Various strategies have been explored to address this challenge.

Knowledge distillation [89, 129, 201] transfers representations from a teacher network trained on complete modality data to a student network that operates on incomplete inputs, guiding it to mimic full-modality embeddings at inference time. *Generative modeling* leverages generative adversarial networks or variational autoencoders to synthesize plausible substitutes for absent modalities, thereby imputing missing inputs before feature extraction [164]. Both approaches require complete modality observations at training time, which limits their robustness when full observations are unavailable during training itself.

A complementary line of work is *joint-embedding modelling*, which learns a shared

latent space from partially available modalities [128, 103, 200]. SMIL [128] perturbs the latent feature space to approximate the embedding of missing modalities, but is restricted to bimodal settings. ShaSpec [200] extends this to more than two modalities by learning shared and modality-specific features; however, its shared encoder complicates generalization to heterogeneous modalities of varying dimensionality. Moreover, shared latent space modeling may introduce biases toward the most available modalities [103]. Cross-modal contrastive learning has shown strong results [219, 134, 119] by aligning heterogeneous data in a joint embedding space. ImageBind [72] scales this idea to six modalities by anchoring all representations to image-paired data, demonstrating that exhaustive pairwise alignment is unnecessary to bind multiple modalities together.

Our proposed ADAPT builds on this anchoring paradigm by training unimodal encoders supervised by a single anchor modality and aligning them in a joint embedding space. It ensures that every modality contributes to the final representation, even when largely absent during training.

4.2.2 Multimodal transformers

Transformers [196] have become the de facto approach for multimodal tasks [153, 175]. They rely on the self-attention mechanism to model long-range dependencies and can accommodate incomplete inputs by masking absent modality tokens. Milecki et al. [133] and Ma et al. [127] demonstrate this by efficiently handling missing data in temporal sequences and bimodal settings via masked attention.

ADAPT generalizes this to more than two modalities by leveraging attention to fuse modalities and to capture their inter- and intra-modal correlations while masking missing ones. We further conduct a systematic study of missing modalities across training and test conditions, demonstrating the versatility and potential of ADAPT for real-life scenarios.

4.3 Motivation: T-REGS for Multimodal Alignment

T-REGS (Chapter 3) is a regularizer that aims to promote uniformity in the representation space by operating independently on each encoder branch, making it a natural fit for multimodal settings where modality-specific encoders can produce non-uniform or poorly calibrated embeddings that hinder cross-modal alignment.

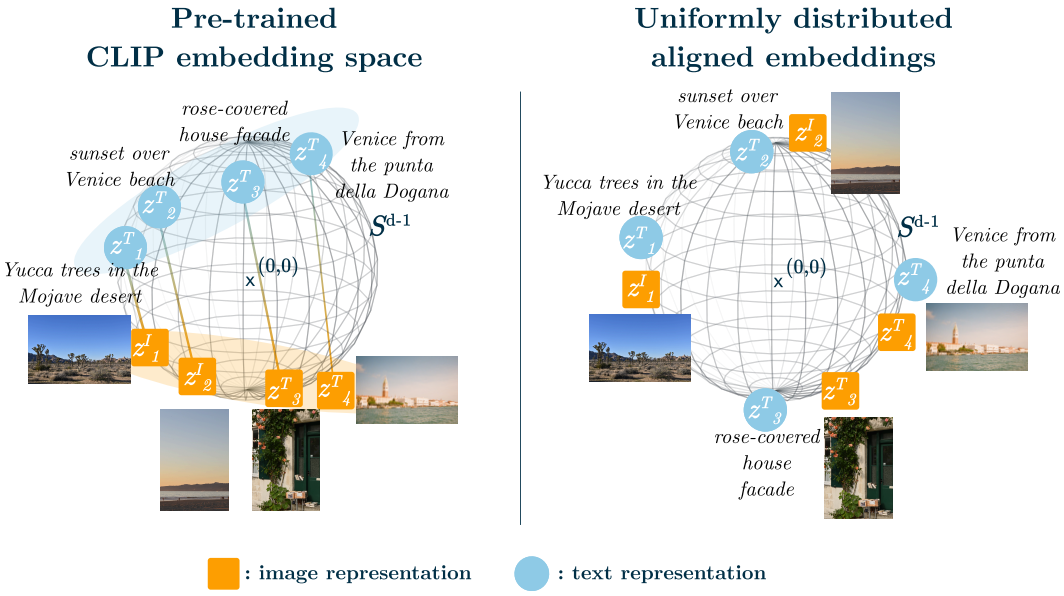


Figure 4.1: **illustration of the modality gap.** *(Left)* Pre-trained CLIP exhibits two separate clusters for images and text: *the modality gap*, a form of dimensional collapse where each modality's embeddings are confined to a distinct cone on the hypersphere, limiting cross-modal transferability. *(Right)* Enforcing uniformity spreads embeddings across the hypersphere, closing the modality gap. Figure adapted from Oh et al. [140].

4.3.1 The modality gap

Pre-trained constrastive multimodal models such as CLIP [151] provide broadly transferable representations across modalities. However, despite being trained to align paired image and text embeddings, such model do not map the two modalities to a shared region of the embedding space. Instead, image and text embeddings tend to occupy disjoint clusters on the unit hypersphere, separated by an approximately constant offset: the *modality gap* [121, 140, 101]. The *modality gap* can itself be understood as a form of dimensional collapse, where the effective dimensionality of the each modalities embedding is reduced by its confinement to its own cone. An illustration is presented in Figure 4.1.

Liang et al. [121] show that this gap arises from two compounding factors. First, the representations of a standard deep neural network are intrinsically confined to a narrow cone at initialization, so that in a two-encoder model the two modality cones are already spatially separated before any training takes place. Second, temperature-scaled contrastive optimization reinforces this separation: by pushing all unpaired embeddings apart, including those *within* each modality, it causes each modality's embeddings to collapse toward a distinct region of the hypersphere

rather than spreading uniformly across it, and the width of the resulting gap is directly modulated by the temperature parameter in the contrastive loss.

Prior analyses [140, 211] formalize this gap further: it is tightly coupled to low embedding uniformity, which limits cross-modal transferability.

4.3.2 T-REGS for CLIP fine-tuning

	Flickr30k [147]				**MS-COCO** [122]			
	i → t		t → i		t → i		t → i	
Method	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Zero-Shot	71.1	90.4	68.5	88.9	31.9	56.9	28.5	53.1
Finetune	81.2	95.5	80.7	95.8	36.7	63.6	36.9	63.9
ES [117]	71.8	90.0	68.5	88.9	31.9	56.9	28.7	53.0
i-Mix [113]	72.3	91.7	69.0	91.1	34.0	63.0	34.6	62.2
Un-Mix [165]	78.5	95.4	74.1	91.8	38.8	66.2	33.4	61.0
m^3-Mix [140]	82.3	95.9	**82.7**	96.0	41.0	68.3	39.9	67.9
$\mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{T-REGS}}$	**83.2**	**96.0**	80.8	**96.4**	**41.6**	**68.7**	**41.5**	**68.7**

Table 4.1: **Cross-Modal Retrieval after finetuning CLIP.** Image-to-text (i → t) and text-to-image (t → i) retrieval results (top 1/5 Recall: R@1, R@5). The table is mostly inherited from Oh et al. [140]. **Boldface** indicates the best performance.

Since the modality gap is rooted in low embedding uniformity, T-REGS is a natural candidate for mitigating it in the multimodal setting: operating independently on each encoder branch, it targets uniformity directly and applies to the bimodal image-text case without any architectural modification. We test this hypothesis on cross-modal retrieval, fine-tuning CLIP by combining $\mathcal{L}_{\text{T-REGS}}$, applied independently to the image and text branches, with the standard $\mathcal{L}_{\text{CLIP}}$ objective [151]:

$$\mathcal{L} = \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{T-REGS}}(Z^{\text{img}}) + \mathcal{L}_{\text{T-REGS}}(Z^{\text{txt}}), \tag{4.1}$$

where Z^{img} and Z^{txt} denote the image and text embeddings, respectively, and with the standard $\mathcal{L}_{\text{CLIP}}$ objective (Eq. (2.12)) [151]:.

Following the experimental protocol of m^3-Mix [140], we fine-tune CLIP ViT-B/32 with Adam for 9 epochs on the target dataset. We compare against baselines that each address the modality gap differently: zero-shot CLIP [151] as a reference, embedding shift [121], which corrects the gap via a post-hoc additive offset, *i*-Mix [113] and Un-Mix [165], which apply mixing strategies to improve cross-modal alignment, and m^3-Mix [140], which uses geodesic interpolation on the hypersphere to bridge the two modality clusters.

We report R@1 and R@5 for image-to-text (i→t) and text-to-image (t→i) retrieval on Flickr30k and MS-COCO in Table 4.1. $\mathcal{L}_{\text{CLIP}}$ + T-REGS achieves the best

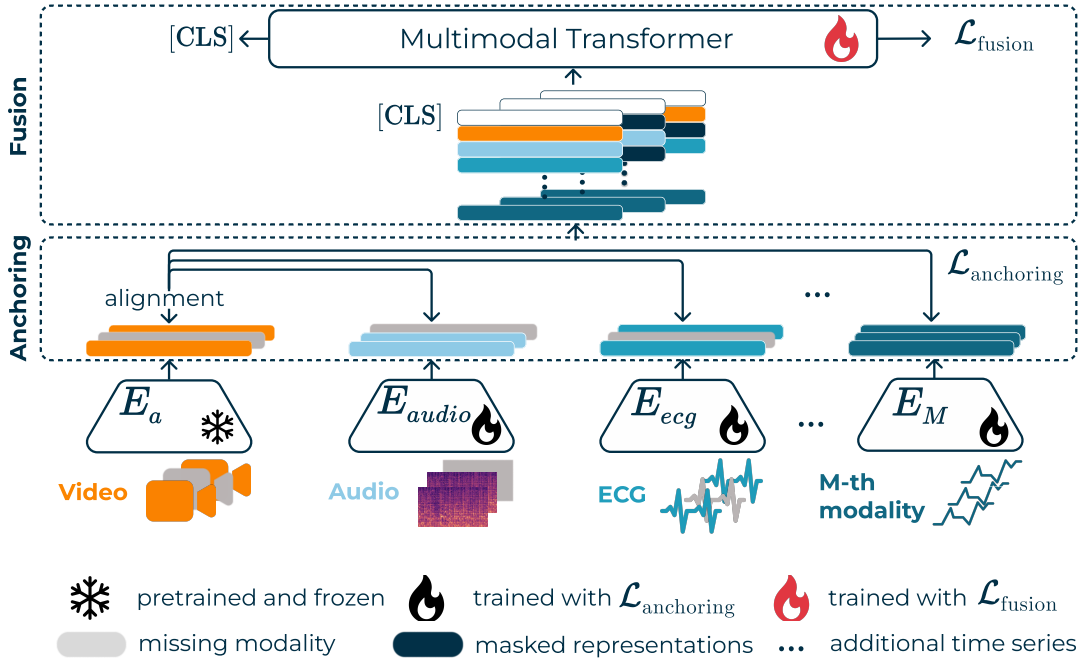


Figure 4.2: **Overview of ADAPT.** ADAPT takes up to M modalities, including video, audio, and biosignals, as input to produce a modality-agnostic representation for downstream tasks. It is trained in two steps. *(i) Anchoring.* We align the representations of all modalities via contrastive learning to the one of an *anchor* modality, i.e., the strongest and richest modality; here the video. *(ii) Fusion.* The encoders' features are concatenated and fed into the Masked Multimodal Transformer. When a modality is unavailable, the transformer masks its corresponding feature representations. The final representation (i.e., `[CLS]` token output) is used for downstream tasks.

performance on the majority of metrics.

These results confirm that T-REGS transfers naturally to the cross-modal setting: promoting embedding uniformity acts as a simple yet effective auxiliary objective that consistently improves retrieval performance, motivating its use beyond standard unimodal self-supervised benchmarks.

4.4 Method

ADAPT (Anchored multimodal Physiological Transformer) is a two-stage framework for detecting physiological changes from heterogeneous modalities, video, audio, and biomedical signals, that remains robust when modalities are missing at training or inference time. ADAPT first aligns all modality-specific representations to a shared space via anchoring (Section 4.4.1), then fuses them through a masked

multimodal transformer (Section 4.4.2). An overview is provided in Figure 4.2.

Notations

Let $\mathcal{D} = \{(x_m^i)_{m=1}^M, y^i\}_{i=1}^N$ denote our training dataset, with M modalities and N labeled observations and $x^i = (x_m^i)_{m=1}^M$ the i-th observation (i.e., a family of m modality values) with $y^i \in \mathcal{Y} = \{0, .., J\}$ its corresponding label (i.e., a physiological state). Given this input, we seek to train a neural network $\mathcal{F}$, that associates to any observation, with any missing modality, a target label $y \in \mathcal{Y}$.

4.4.1 Anchoring modality-specific encoders

We train modality-specific encoders with a contrastive objective that aligns each modality's representations to those of a designated *anchor* modality, embedding all modalities into a common feature space. The anchor is chosen as the modality with the richest unimodal signal; in this work, video, as it captures visually distinguishable physiological changes, however, any modality can be the anchor.

Modality-specific encoders

Each modality is encoded using a dedicated encoder. For *video*, we use the pre-trained Hiera [159] encoder. For *audio*, each sample is encoded into a mel-spectrogram (a 2D acoustic time-frequency representation of sound), fed to BYOL-A [138] to obtain a 1-d feature. *Biomedical signals* are processed using 1D CNNs [203, 92]. A modality-specific linear projection head maps each encoder's output to a fixed-size d-dimensional embedding $z_i^m \in \mathbb{R}^d$. ADAPT can be extended to other modalities by adding their respective encoders.

Anchoring

We consider a pair of modalities with aligned observations $(\mathcal{A}, \mathcal{M}_m)$, where $\mathcal{A}$ represents the anchor (video) and $\mathcal{M}_m$ another modality. The anchor video x_a^i and its corresponding observation x_m^i are encoded using $f_a^i{=}E_a(x_a^i)$ and $f_m^i{=}E_m(x_m^i)$, respectively, where E_a is a pre-trained and frozen video encoder and E_m a neural network. Projection heads map the embeddings to $z_a^i, z_m^i{\in}\mathbb{R}^d$. The loss is computed on z_a^i and z_m^i [72]:

$$\mathcal{L}_{\mathcal{A},\mathcal{M}_m} = -\sum_{i=1}^{B} \log \frac{\exp(\cos(z_a^i, z_m^i)/\tau)}{\sum_{k=1}^{B} \exp(\cos(z_a^i, z_m^k)/\tau)} \quad , \tag{4.2}$$

where $\tau \in \mathbb{R}^+$ is a temperature, $\cos(\cdot,\cdot)$ the cosine similarity, and B the batch size. In practice, we symmetrize by adding $\mathcal{L}_{\mathcal{M}_m,\mathcal{A}}$. The full anchoring objective

sums over all non-anchor modalities:

$$\mathcal{L}_{\text{anchoring}} = \sum_{\substack{m=1 \\ \mathcal{M}_m \neq \mathcal{A}}}^{M} (\mathcal{L}_{\mathcal{A},\mathcal{M}_m} + \mathcal{L}_{\mathcal{M}_m,\mathcal{A}}) \tag{4.3}$$

To alleviate the modality gap [121] (see Section 4.3), we augment $\mathcal{L}_{\text{anchoring}}$ with T-REGS applied independently to each modality's embedding batch:

$$\mathcal{L} = \mathcal{L}_{\text{anchoring}} + \sum_{m=1}^{M} \mathcal{L}_{\text{T-REGS}}(Z_m), \tag{4.4}$$

where Z_m denotes the embedding batch for modality m.

4.4.2 Masked multimodal transformer

To fuse the anchored modality representations into a single modality-agnostic embedding, we use a transformer [196] with N_L attention blocks.

For each sample, we stack the modality-specific representations, $z_m^i \in \mathbb{R}^d, \forall m \in [1, M]$, into a single matrix and prepend a special token `[CLS]`, yielding a matrix $Z \in \mathbb{R}^{(M+1)\times d}$. Similarly to Liu, Courant, and Kalogeiton [123] and Nagrani et al. [137], the query, key and value are derived from F via: $Q = W^Q Z$, $K = W^K Z$ and $V = W^V Z$ where $Q, K \in \mathbb{R}^{(M+1)\times d_k}$ and $V \in \mathbb{R}^{(M+1)\times d_v}$. Our modelization of inter-modal interactions differs from the usual cross-attention [33, 93], which asymmetrically combines two separate embedding sequences of the same dimension. Using stacked features Z allows generalization to any number of modalities, with linear scalability in the number of modalities instead of quadratic.

Handling missing modalities

Inspired by Milecki et al. [133] for missing follow-up patient examinations, we apply our strategy to deal with missing modalities to the scaled dot-product, core of each multi-head self-attention sub-layer. We consider one sub-layer with one head ($h = 1$) for simplicity. We use a masking binary matrix $\Omega \in \mathbb{R}^{(M+1)\times(M+1)}$ that specifies which modalities are missing: $\omega_{ij} = 1$ if i and j are available, else $\omega_{ij} = 0$. The output $O \in \mathbb{R}^{(M+1)\times d_v}$ of the attention mechanism is, for O_i each line of O:

$$O_i = \sum_j \omega_{ij} \frac{\exp(Q_i^T K_j / \sqrt{d_k})}{\sum_{\{j', z_{ij'}=1\}} \exp(Q_i^T K_{j'} / \sqrt{d_k})} V_j \quad . \tag{4.5}$$

When $h > 1$, queries, keys, and values are linearly projected h times with different, learned linear projections, concatenated, and once again projected after the scaled-dot product.

Modality dropout

We train the Masked Multimodal Transformer with a multi-view contrastive objective [34]. Drawing inspiration from Shi et al. [166], we mitigate the model's over-reliance on a single modality while enhancing its robustness in the absence of modalities through an augmentation technique called *modality dropout*. We leverage the masking scheme at the attention level to randomly mask input modalities. Given a batch $\mathcal{Z}$, we create two simultaneous view $\mathcal{Z}'$ and $\mathcal{Z}''$. For each observation within $\mathcal{Z}'$, we hide up to $M-1$ modalities following a uniform probability, M the number of modalities. Additionally, motivated by Han et al. [80], who recently showed the effect of additive noise on electrocardiograms, we add $\varepsilon \sim \mathcal{N}(0, \sigma)$ on the biomedical signals to each view independently. We chose σ based on the amplitude of the signal. We use the infoNCE loss to enforce the similarity between the two views [34] on the final representation output given by the `[CLS]` token. Since the two representations are already mapped to the same dimension, following Jing et al. [96], we directly optimize the representations to enforce scalability and mitigate dimension collapse: $\mathcal{L}_{\text{fusion}} = -\sum_{i=1}^{B} \log \frac{\exp(\cos(\text{CLS}^i, \text{CLS}^{i'})/\tau)}{\sum_{k=1}^{B} \exp(\cos(\text{CLS}^i, \text{CLS}^{k'})/\tau)}$.

4.5 Experiments

4.5.1 Dataset

`StressID` [29] is a dataset specifically designed for stress identification from unimodal and multimodal data. It contains videos of facial expressions, audio recordings (AUDIO), and physiological signals, namely electrocardiography (ECG), electrodermal activity (EDA), and respiration (RR), all acquired in a synchronized manner using a wearable device paired with an RGB camera and integrated microphone. The dataset includes recordings from 65 participants performing 11 tasks, encompassing emotional video clips, cognitive exercises, and public speaking scenarios, with self-reported ratings of perceived stress, relaxation, arousal, and valence.

`StressID` has 18% and 46% of missing video and audio recordings, respectively. We denote the entire training and testing sets as X_{train}, X_{test} (considering samples with and without missing modalities); and X^*_{train}, X^*_{test} for the train and test sets where all modalities are available.

4.5.2 Metrics

We use the balanced accuracy (ACC) and weighted F1 score (F1). We report metrics in the format `mean(std)` in %.

4.5.3 Comparison to the state of the art

We report the performance of ADAPT and competing methods in Table 4.2, evaluated in the presence of all modalities. We consider unimodal baselines for video, audio, and biomedical signals (rows 1–3) from Chaptoukaev et al. [29], and the following multimodal methods:

1. *Feature-level fusion* [29] (row 4): unimodal features are concatenated into a single high-dimensional vector and fed to an MLP, trained with cross-entropy loss.

2. *Decision-level fusion* [29] (row 5): independent SVMs are trained per modality and their predictions are combined via ensemble rules (sum, average, product, maximum); the best-performing rule is reported.

3. *ShaSpec+* [200] (row 6): ShaSpec learns shared and modality-specific features to maximize the use of available modalities. Since applying a shared encoder directly to heterogeneous raw inputs of varying dimensionality (e.g., 3D video, 1D biomedical signals) is nontrivial, we feed *encoded* inputs to the shared encoder instead, hence the "+" variant. For fair comparison, we use the same encoders as ADAPT: Hiera [159] for video, BYOL-A [138] for audio, and a 1D CNN [203] for biomedical signals.

ADAPT achieves the best results on X^*_{test}, outperforming all methods from Chaptoukaev et al. [29] by a notable margin, for instance, surpassing decision-level fusion by 6.3 points in ACC and 6.7 in F1. ADAPT also outperforms ShaSpec+, the strongest baseline, by 1.1 points in ACC and 3.0 in F1.

4.5.4 Ablations

Robustness to missing modalities during inference

Beyond training-time robustness, we evaluate ADAPT when entire modalities are withheld at inference (Table 4.3). Row 1 establishes the full-modality baseline on X_{test} (ACC = 74.6, F1 = 74.7); rows 2–4 successively remove video, audio, or both. Withholding video causes only a marginal drop ($|\Delta| = 1.3$ in ACC, 1.6 in F1), a direct consequence of ADAPT's anchoring strategy: by projecting all modalities into the video latent space during training, the remaining modalities

	ACC	F1
Unimodal		
Video	62(4)*	67(3)*
Biomedical signals	58(4)*	66(5)*
Audio	62(4)*	67(4)*
Multimodal		
Feature Fusion [29]	61(3)*	66(4)*
Decision-level fusion [29]	65(5) *	72(5)*
ShaSpec+ [200]	<u>70.2(3.7)</u>	<u>75.7(5.3)</u>
ADAPT	**71.3(4.2)**	**78.7(4.6)**

Table 4.2: **Comparison with state-of-the-art methods on StressID.** We report ACC and F1 (mean ± std) against unimodal and multimodal baselines from Chaptoukaev et al. [29] and an additional multimodal baseline designed for missing modalities: ShaSpec+ [200]. **Bold**: best; <u>underlined</u>: second best. *From Chaptoukaev et al. [29].

Video	Audio	ACC	F1
✓	✓	74.6(1.9)	74.7(2.0)
×	✓	73.3(1.6)	73.1(1.6)
✓	×	67.5(3.4)	65.5(4.8)
×	×	66.9(3.3)	65.2(3.8)

Table 4.3: **ADAPT under missing-modality conditions on the StressID test set.** ✓/× denote modality availability at test time. We report ACC and F1 (mean ± std).

acquire representations that closely mirror the video anchor and can therefore substitute for it at inference. This result also validates the choice of video as anchor: since video is the strongest unimodal modality (ACC = 66.8, F1 = 71.3 on X^*_{test}, Table 4.5), anchoring in a rich unimodal space maximizes the information available to the other modalities when that anchor is absent at test time. Removing audio, or both audio and video, leads to a more pronounced but still contained degradation. Even in this most challenging setting, ADAPT retains ACC above 66% and F1 above 65, whereas competing methods, having no mechanism to handle missing modalities, cannot be evaluated under these conditions at all.

Impact of T-REGS during anchoring

	X^*_{test}		X_{test}	
T-REGS	**ACC**	**F1**	**ACC**	**F1**
×	69.9(3.7)	77.0(4.3)	72.9(2.9)	72.9(3.1)
✓	**71.3(4.2)**	**78.7(4.6)**	**74.6(1.9)**	**74.7(2.0)**

Table 4.4: **Ablation of the T-REGS term in Eq. (4.4) on X^*_{test} and X_{test}.** We compare training with $\mathcal{L}_{\text{anchoring}}$ alone (×) versus the full objective that adds $\sum_m \mathcal{L}_{\text{T-REGS}}(Z_m)$ (✓). **Bold** indicates the best result per column.

We ablate the T-REGS term in Eq. (4.4) by comparing the anchoring stage trained with $\mathcal{L}_{\text{anchoring}}$ alone (Eq. (4.3)) against the full objective that augments it with $\sum_{m=1}^{M} \mathcal{L}_{\text{T-REGS}}(Z_m)$.

As shown in Table 4.4, the full objective yields consistent gains on both test sets. On X_{test}, it improves ACC and F1 from 72.9 to 74.6 and 74.7 (+1.7 ACC, +1.8 F1). On X^*_{test}, it brings ACC and F1 from 69.9 and 77.0 to 71.3 and 78.7 (+1.4 ACC, +1.7 F1).

These results confirm that T-REGS is beneficial beyond the unimodal setting of Chapter 3: enforcing per-modality representation uniformity during anchoring consistently improves performance, extending the motivation established in Section 4.3 to the multimodal domain-adaptation setting.

Unimodal performances

Unimodal performances are reported in Table 4.5. As expected, all modalities individually underperform the full multimodal pipeline, confirming that stress recognition benefits from complementary information across modalities. Among physiological signals, ECG achieves the strongest results, while EDA lags behind, consistent with its known sensitivity to noise and individual variability. Overall, Video is the best-performing single modality, outperforming all physiological

	X^*_{test}		X_{test}	
	ACC	**F1**	**ACC**	**F1**
EDA	51.1(6.4)	40.2(5.7)	54.8(1.9)	48.3(2.1)
RR	58.1(3.7)	67.0(4.8)	68.7(3.7)	68.6(3.9)
ECG	65.4(4.3)	71.6(3.0)	69.2(2.3)	69.5(2.3)
Audio	58.7(7.9)	67.8(6.4)	—	—
Video	66.8(5.4)	71.3(5.3)	—	—

Table 4.5: **Performance of unimodal encoders.** Each encoder is evaluated independently using the same backbone as in the full multimodal pipeline. Results are reported on X^*_{test} and X_{test} as `mean(std)` over ACC and F1. Audio and Video have no X_{test} results due to missing modalities.

signals on X^*_{test}. Audio and Video cannot be evaluated on X_{test} due to missing modalities in that split.

4.5.5 Implementation details

Our implementation is available at `adapt`. Table 4.6 lists the architectural details used for each dataset. The seed is fixed to 42. In our experiments, for both stages we use the AdamW [125] optimizer, with a weight decay of 0.05, a starting learning rate of $1e^{-4}$ following a cosine schedule, and preceded by a linear warm-up of 4 epochs on 1 NVIDIA Tesla H100.

4.6 Computational Cost

All training runs were performed on V100 or H100 GPUs, with a single ADAPT training taking approximately 15 minutes. The total computational cost of the project, including baseline training, experiments, and ablation studies, amounts to approximately 1,000 GPU-hours.

4.7 Discussion

While effective on the settings considered, ADAPT has several limitations.

The anchoring stage assumes that a privileged modality can be designated as the *anchor*: a frozen, pretrained encoder to which all other modalities are aligned. In this work, video plays this role because it is the strongest unimodal signal on StressID (Table 4.5). This choice, however, ties the quality of the joint space to the availability of a strong pretrained encoder for that modality, and to the

Component	Parameter	StressID
Unimodal Encoders	Audio	BYOL-A [138]
	Video	Hiera [159]
	Biomedical signals	FCN [92]
Anchoring	Anchor	Video
	d	64
	τ	Temperature schedule [109]
	γ	0.08
	Projection head	1 fully-connected layer
Multimodal Transformer	d_{model}	64
	n_{heads}	4
	N_L	2
	d_{ffn}	256
	d_v	64
	d_k	64
	Activation	GeLU
	σ	0.4
	Normalization	LayerNorm
Linear Classifier	Loss	Cross Entropy

Table 4.6: **Architectural details of ADAPT.**

assumption that the information relevant to the downstream task is well captured by it. Moreover, non-anchor modalities are aligned only to the anchor, not to one another: any information shared among weaker modalities but absent from the anchor is therefore learned only later, at the fusion stage. The contrastive objective itself also requires co-occurrence with the anchor, so samples where the anchor is missing cannot contribute to the corresponding pairing terms.

Additionally, the empirical study is restricted to stress detection on StressID, a cohort of 65 participants. Although ADAPT remains robust when entire modalities are withheld at inference (Table 4.3), this protocol is more structured than the random modality dropout used at training, and may not fully reflect the correlated missingness patterns of real deployments. Broader validation on other physiological tasks and cohorts would be needed to establish the generality of the anchoring-then-fusion methodology.

4.8 Conclusion

In this chapter, we proposed ADAPT, a modality-agnostic representation framework that operates effectively under missing modalities at both training and in-

ference time. ADAPT outperforms the state of the art on physiological change detection and remains robust under missing modalities. Furthermore, we evaluated the use of T-REGS as an auxiliary regularizer applied independently to each modality's embeddings during the anchoring stage. These results confirm that enforcing representation uniformity during anchoring consistently improves performance, validating the motivation established in Section 4.3.

Chapter 5

Assessing Self-Supervised Learning Representations via Intrinsic Dimension

Chapters 3 and 4 leveraged Minimum Spanning Trees (MST) to *shape* joint-embedding self-supervised learning (JE-SSL), enforcing uniformity and preventing dimensional collapse. This chapter asks a complementary question: *can the geometry of learned representations provide reliable, label-free insights into their quality for downstream tasks?*

We begin by motivating the approach (Section 5.1) and covering the necessary background (Section 5.2), before presenting the theoretical grounding (Section 5.3). Experimental results are reported in Section 5.4, and we conclude with a discussion of limitations in Section 5.7.

This chapter is based on

***IdEst**: Assessing Self-Supervised Learning Representations via Intrinsic Dimension*

J. Mordacq, V. Kalogeiton, S. Oudot

ICML 2026

5.1 Introduction

Regardless of the SSL method, evaluating learned representations typically relies on linear probing, a protocol that is computationally expensive, sensitive to hyperparameters, and limited in the insight it provides into the geometric structure of the representation space. Yet, since SSL representations inhabit high-dimensional spaces whose geometry is directly shaped by their training objectives, intrinsic properties of the learned manifold are natural label-free proxies for representation quality [5, 67].

Among them, intrinsic dimension (ID), originally introduced by Bennett [15], has emerged as a particularly informative quantity: it characterizes the effective number of degrees of freedom required to represent data in an embedding space. Recent studies have revealed a monotonic relationship between a model's generalization error and the intrinsic dimension of its representations [5, 104, 158], with lower intrinsic dimension often correlating with improved downstream accuracy. These findings, grounded in the *manifold hypothesis* [73], suggest that representation quality is governed not only by separability, but also by how efficiently information is compressed into a low-dimensional geometric structure. Yet, existing evidence is largely confined to supervised convolutional networks, leaving open how intrinsic dimension behaves across modern self-supervised methods and whether it indicates downstream performance in this setting.

In practice, estimating ID is far from trivial. Intrinsic dimension admits multiple mathematical formalizations, topological, fractal, and information-theoretic, each capturing different facets of data geometry. Moreover, since ID cannot be observed directly but must be estimated from finite samples, the choice of estimator matters. Most notably, nearest-neighbor-based methods such as TwoNN [61] and maximum likelihood estimators [115] have been widely adopted, but suffer from well-known limitations: they rely on strong locality and isotropy assumptions, are sensitive to noise and finite-sample effects, and become unstable in high-dimensional or highly structured representation spaces. These limitations are especially pronounced in SSL, which operates far from standard conditions. First, it operates in a non-asymptotic regime where $n \approx d$ rather than $n \to \infty$ with d fixed, where n is the number of samples and d is the ambient dimension. Second, SSL objectives introduce dependencies between data points, violating standard independence assumptions. For instance, as shown in Figure 5.2, TwoNN and MLE are sensitive to these inter-sample dependencies; they are unable to reliably detect the intrinsic dimension, with TwoNN even diverging.

To address these limitations, we consider a complementary ID estimator grounded

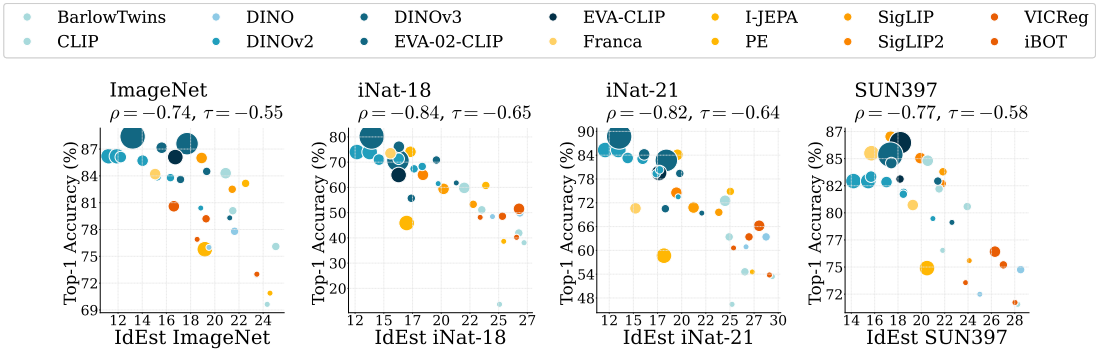


Figure 5.1: **Foundation models and IdEst**: Intra-dataset correlation. Linear probe accuracy of pretrained SSL models on ImageNet (left), iNat-18 (middle left), iNat-21 (middle right), SUN397 (right) versus IdEst on each respective dataset. Each point corresponds to a model checkpoint; point size reflects the number of parameters. We report Kendall's $\tau \in [-1, 1]$ and Spearman's $\rho \in [-1, 1]$. Correlations across all four benchmarks demonstrate IdEst's ability to provide insights into models' representation quality.

in the scaling behavior of minimum spanning trees [40]: $\dim_{\text{MST}}$. Asymptotically tied to the intrinsic Rényi entropy, this estimator balances local and global information, enjoying robustness to noise, sampling-density variations, and high ambient dimensionality. Furthermore, its reliability under sampling sparsity, a behavior consistent with known theoretical results on MST length in sparse regimes (e.g., Theorem 3.2.1 in Chapter 3), making it particularly well-suited to the $n \approx d$ regime inherent in vision SSL.

Building on this perspective, we propose IdEst (pronounced *id est*, Latin for 'that is'), standing for ID Estimation for SSL using minimum spanning Trees. IdEst is an unsupervised criterion for evaluating self-supervised representations based on intrinsic dimension estimation via minimum spanning trees. We show that IdEst strongly correlates with downstream performance across a wide range of self-supervised objectives, including joint-embedding, joint-predictive, and vision-language alignment (Figures 5.1, 5.3 and 5.4). Moreover, IdEst provides an efficient proxy for supervised linear probing, enabling practical hyperparameter selection without labels at a fraction of the computational cost of supervised probing.

Our main contributions can be summarized as follows:

i) We propose IdEst, an unsupervised criterion based on the intrinsic dimension of the representation (Section 5.3).

ii) We demonstrate that IdEst serves as an efficient proxy to assess the quality

of SSL representations (Section 5.4.2).

iii) We show that IDEST enables unsupervised hyperparameter selection across diverse SSL objectives and architectures (Sections 5.4.3 and 5.4.4).

5.2 Background: Label-Free Geometric Analysis of Representations

Our work builds on recent efforts to analyze geometric properties of deep neural networks (DNNs) in a label-free setting, and to understand their connection to generalization. Existing literature investigates these properties through two main lenses: spectral properties and intrinsic dimensionality.

5.2.1 Representation spectrum

α-ReQ [4] and RankMe [67] characterize the eigenspectrum of representations within SSL frameworks, by measuring the decay rate of empirical eigenvalues or calculating the effective rank of the representations, respectively. These studies showed that both metrics often strongly correlate with downstream performance and highlighted their utility for hyperparameter selection.

However, the power-law assumptions underlying α-ReQ are violated in the presence of representation collapse [81] (where networks output identical or non-informative vectors regardless of the input, resulting in rank-deficient representations), leading to unreliable performance estimates when the embedding space becomes rank-deficient [67]. Furthermore, RankMe is limited to the study of Joint-Embedding Architectures (JEAs) where two networks are trained to produce similar embeddings for different views of the same image. Notably, RankMe is tailored to identify a pivotal challenge of JEAs: representation collapse [95]. This narrow focus potentially limits the applicability of such metrics to other SSL paradigms less prone to this specific type of collapse. For instance, as shown in Table 5.1, RankMe is less effective on I-JEPA, a joint-predictive method.

More recently, Thilak et al. [184] proposed LiDAR, which quantifies the rank of the Linear Discriminant Analysis matrix associated with the surrogate SSL task, a measure that intuitively captures how much information the representation retains for solving that task. Crucially, LiDAR leverages SSL pretraining information (i.e., augmentations), whereas our work targets the harder setting, where only frozen representations are accessible, without any access to training data or pairing structure, aiming to gain a deeper understanding of SSL models beyond hyperparameter selection.

5.2.2 Intrinsic dimension

Intrinsic dimension (ID) has been linked to the generalization of DNNs through two primary lenses: *the optimization trajectory*, analyzing the sequence of model states and parameter updates during training [171, 17, 57, 181], and *the learned representations* [5, 104, 158]. In this work, we focus on the latter. Analyzing optimization trajectories is often computationally expensive, as it requires processing all network parameters (which can reach hundred of millions in current vision models, e.g., ViT-L [53]) and necessitates access to intermediate training checkpoints that are rarely available for large-scale, pre-trained models.

Regarding representational analysis, several studies have investigated the ID of Large Language Model (LLM) representations [3, 21, 191, 192, 199, 112, 158] demonstrating that ID provides critical insights into training dynamics and generalization. However, research on computer vision models has largely remained restricted to supervised convolutional neural networks (CNNs). For instance, Ansuini et al. [5] showed that the ID of supervised CNNs correlates with performance, while Konz and Mazurowski [104] proposed a generalization scaling law based on representational ID, though their validation was limited to supervised CNN architectures.

Despite these advances, the use of intrinsic dimension estimations to characterize representation quality remains largely unexplored in self-supervised learning.

5.3 Generalization and Dimension Estimation

This section first introduces the theoretical connection between intrinsic dimension and generalization (Section 5.3.1). Second, it presents standard intrinsic dimension estimators used in prior studies of deep neural network representations, along with their limitations, and motivates the use of an alternative: $\dim_{\text{MST}}$ (Section 5.3.2).

5.3.1 Theoretical connection to generalization

Consider a classification dataset $\mathcal{D}$, consisting of N_D points $x \in \mathbb{R}^n$ with target labels $y = \mathcal{F}(x)$ defined by an unknown function $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}^C$, where C is the number of classes. The dataset is split into a training set $\mathcal{D}_{\text{train}}$ and a test set $\mathcal{D}_{\text{test}}$. We analyze 'well-trained' models $f : \mathbb{R}^n \to \mathbb{R}^C$ that interpolate the training data, such that $f(x) = \mathcal{F}(x)$ for all $x \in \mathcal{D}_{\text{train}}$. Let $\mathcal{L}$ be a non-negative loss function (e.g., cross-entropy) satisfying $\mathcal{L}(f(x), \mathcal{F}(x)) = 0$ if $f(x) = \mathcal{F}(x)$. The generalization error is expressed as the expected loss over $\mathcal{D}_{\text{test}}$. The model can be decomposed as $f = h \circ g$, where g is an encoder (e.g., a pre-trained SSL backbone) that produces latent representations living on some d-dimensional manifold, and

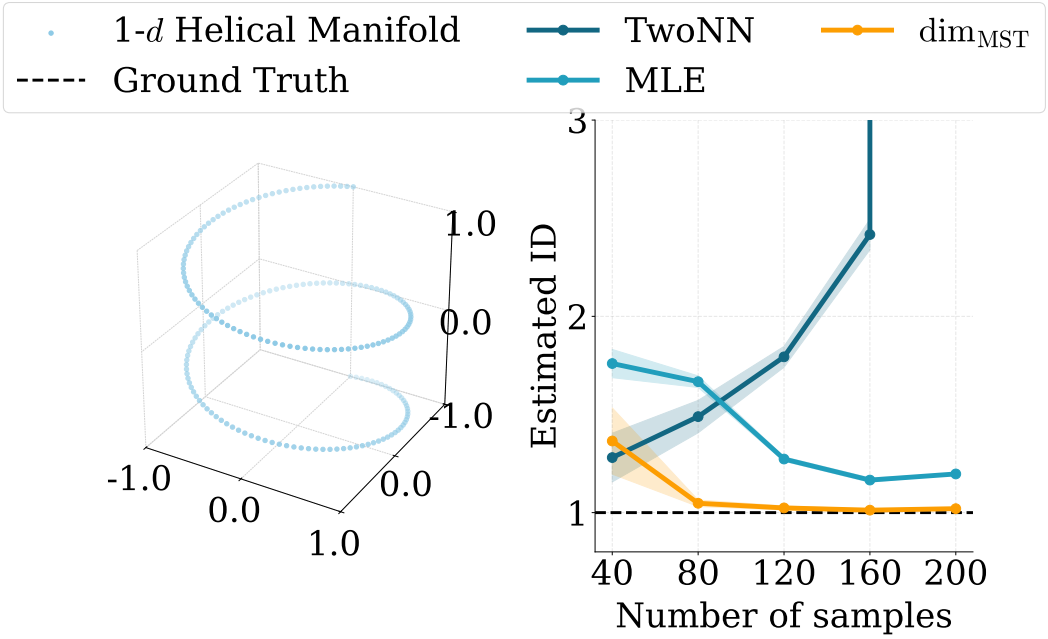


Figure 5.2: **Impact of sampling distribution on estimators.** *(Left)* 200 points sampled evenly along a 1-dimensional helix. *(Right)* Estimated ID as a function of sample size. While $\dim_{\text{MST}}$ converges accurately to the ground truth $d = 1$ as the sample size increases, MLE and TwoNN do not, with TwoNN even diverging to infinity.

where h is a classification head.

Theorem 5.3.1. *Konz and Mazurowski [104] Let K_L the Lipschitz constant of the loss function. Then:*

$$\mathcal{L} \simeq \mathcal{O}\left(K_L N_D^{-1/d}\right) \tag{5.1}$$

This relation suggests that, for a fixed dataset size N_D, the error is dominated by the $-\frac{1}{d}$ exponent. Consequently, a lower d implies a more efficient representation. We therefore use the estimated ID of the representations as an unsupervised criterion to assess the quality of the downstream task.

5.3.2 Intrinsic dimension estimators

The intrinsic dimension (ID) estimation problem assumes the data points are sampled on (or close to) some unknown d-submanifold of the ambient space. The goal is to estimate d from the data.

Parametric estimators

While many estimators exist [97, 183, 16], two main estimators have been adopted in prior studies of deep neural network representations [5, 148, 104, 158]: *Maximum Likelihood Estimation (MLE)* [115] and *TwoNN* [61]. Both are parametric estimators, founded on the assumption that the data points are sampled i.i.d. from a probability distribution supported on the submanifold, with locally constant density, and both treat the data points locally as a homogeneous Poisson process. Under these conditions, the number of points within a small ε-ball, and the ratio between the distances to the second and first nearest neighbors, follow specific parametric distributions whose parameters depend directly on d and can be inferred, either via maximum likelihood estimation for MLE or via regression for TwoNN.

These methods are inherently tied to specific data distributions, and they become unstable as the input data deviate from those, even in simple cases. Consider, for instance, the sample of Figure 5.2, composed of up to 200 points evenly spaced along a 1-dimensional helix embedded in $\mathbb{R}^3$. Despite the high regularity of both the submanifold and the sample, both methods fail to recover the correct intrinsic dimension, with TwoNN even diverging to infinity in the asymptotic regime. This behavior is consistent across standard choices of hyperparameters of the methods. In practice, this instability translates into a degradation of the performances in the presence of noise [191, 16], or on heavy-tailed distributions [17].

Metric invariants based estimators

The limitations of the parametric methods TwoNN and MLE motivate a shift toward estimators based on the theory of Euclidean functionals [212]. In particular, the asymptotic growth rate of the total length of a Minimum Spanning Tree (MST) is related to the Rényi entropy of the underlying distribution, yielding provably consistent dimension estimators under mild assumptions, i.e., compactness of the manifold and boundedness of the Lebesgue sampling density [40]. For instance, the *Minimum Spanning Tree dimension estimator*, or $\dim_{\mathrm{MST}}$ for short, successfully recovers the ground truth dimension $d = 1$ in the example of Figure 5.2.

We recall the $\dim_{\mathrm{MST}}$ estimator [40], previously introduced in Section 2.4.1 for the reader's convenience, covering its definition, practical estimation procedure, and theoretical stability guarantees.

Definition. Given a bounded metric space M, the MST dimension of M, denoted by $\dim_{\mathrm{MST}}(M)$, is the infimal exponent $d \in \mathbb{N}$ such that $L\left(\mathrm{MST}(X)\right)/|X|^{\frac{d-1}{d}}$ is

uniformly bounded for all finite subsets $X \subseteq M$:

$$\dim_{\mathrm{MST}}(M) := \inf\Big\{ d : \exists C \text{ s.t. } \frac{L\left(\mathrm{MST}(X)\right)}{|X|^{\frac{d-1}{d}}} \leq C \quad \text{for every finite subset } X \text{ of } M \Big\}.$$

In practice, d is estimated via log-log linear regression [17, 16]. Specifically, given subsamples X_{n_i} with increasing sizes n_i, we fit:

$$\log L\left(\mathrm{MST}(X_{n_i})\right) \approx \frac{d-1}{d} \log(n_i) + \log(C).$$

The resulting slope m yields the intrinsic dimension via the relation $d = 1/(1-m)$. The complete algorithm for computing $\dim_{\mathrm{MST}}$ is given in Algorithm 1 (Section 2.4.1).

Furthermore, the MST length, coincides with the total degree-0 persistence of the Rips filtration [144, 1], which is provably stable under perturbations of the underlying distribution [30]. Since $\dim_{\mathrm{MST}}$ is derived from the scaling behavior of MST length, it inherits this stability, providing additional theoretical grounding for its robustness.

5.4 IdEst: $\dim_{\mathrm{MST}}$ for Unsupervised Assessment of Self-Supervised methods

In this section, we apply $\dim_{\mathrm{MST}}$ to estimate the intrinsic dimension of self-supervised representations. This yields IDEST, which stands for Intrinsic Dimension Estimation for SSL using minimum spanning Trees. Our goal is to determine whether IDEST can yield unsupervised insights into downstream performances. Specifically, to empirically validate IDEST, we compare it against linear probing, the standard evaluation protocol for self-supervised learning (SSL).

Subsequently, we address the following research questions:

Q1. To what extent does IDEST reflect representation quality across pretrained SSL models?
Q2. Can IDEST yield insight along self-supervised pretraining?
Q3. Can IDEST be leveraged as a principled proxy for hyperparameter selection without requiring labeled data?

5.4.1 Implementation of IdEst

To satisfy the assumptions of Theorem 5.3.1, IDEST operates on the frozen representation passed to the classifier head, following each method's standard evaluation protocol (i.e., consistent with how linear probes are trained). For models without a class token (e.g., I-JEPA), we average-pool the patch tokens; for models with a [CLS] token (e.g., DINO, DINOv2), we use this token directly. We follow the implementation of Adams et al. [1] and Dupuis, Deligiannidis, and Simsekli [57] for $\dim_{\text{MST}}$. For each dataset and model, we compute the dimension estimator. Preprocessing (distance matrix computation and sorting) dominates in practice and is optimized via Ripser [188, 13], which implements an efficient sparse distance pipeline. To keep computation tractable, IDEST operates on a subsample of size $N \ll N_D$, where N_D denotes the full dataset size; we set $N = 50{,}000$ throughout.

5.4.2 IdEst and accuracy of foundation models

We begin by evaluating whether our proposed IDEST reflects representation quality across a diverse set of pretrained SSL models. Specifically, we compute IDEST on frozen representations and compare it against standard linear probing accuracy. We compute two rank correlation statistics: *Spearman's rank correlation coefficient* ρ [173] and *Kendall's rank correlation coefficient* τ [99]. Both capture monotonic relationships between rankings.

Set-up

We evaluate 14 SSL methods spanning four paradigms: pure joint-embedding (e.g., VICReg [12], DINO [26]), joint-predictive (e.g., I-JEPA [8]), combined objectives (e.g., iBOT [220], DINOv2 [143]), and vision-language alignment (e.g., CLIP [151], EVA-CLIP [178]).

For each method, we evaluate two main architectures: ResNet [84] and ViT [53], and we include various model scales (e.g., ViT-S, and ViT-G). This results in 33 different models. Table 5.3 (in supplementary) provides the complete list of models and architectures studied.

Results

We first examine Intra-Dataset Correlation in Figure 5.1, where IDEST is estimated on each dataset separately and compared against the corresponding linear probe accuracy. Figure 5.1 reports results on ImageNet [47] and three additional fine-grained datasets: iNat-18 [195], iNat-21 [194], and SUN397 [208]. Each point represents a pretrained SSL model, spanning diverse architectures, training objectives, and scales.

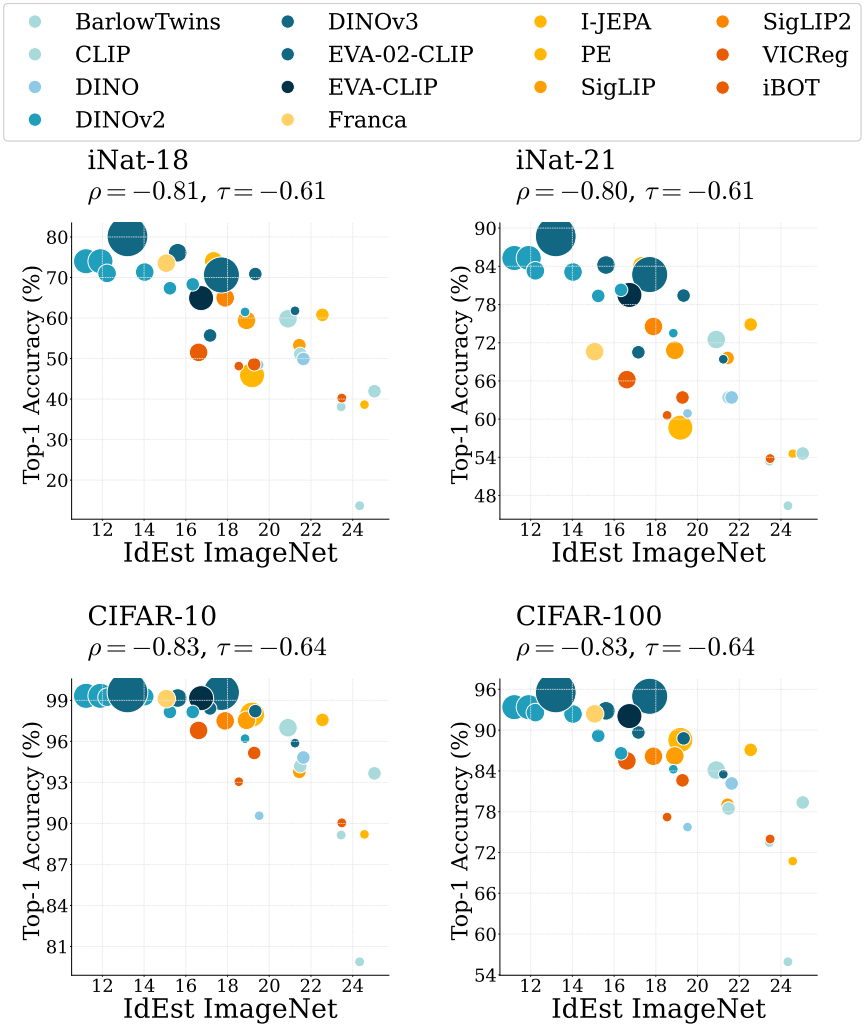


Figure 5.3: **Foundation models and `IdEst`:** Inter-dataset correlation. Linear probe accuracy of pretrained SSL models on iNat-18 (top right), iNat-21 (top left), CIFAR-10 (bottom right), CIFAR-100 (bottom left) versus `IdEst` computed on ImageNet. Strong correlations demonstrate that `IdEst` computed on a single reference dataset: ImageNet, is indicative of model quality across datasets.

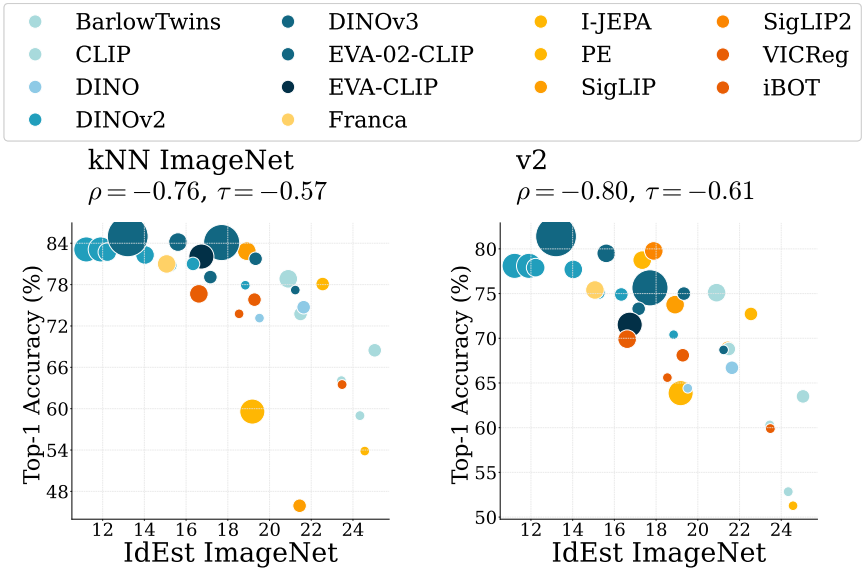


Figure 5.4: **Foundation models and IdEst:** Alternative evaluation protocol. Accuracy under alternative evaluation settings versus IdEst. Strong correlations demonstrate that IdEst computed on a single reference dataset: ImageNet, is indicative of model quality across evaluation protocols.

Across all datasets, Figure 5.1 reveals a consistent negative correlation between IdEst and downstream linear probing accuracy: models with lower intrinsic dimension tend to achieve higher performance. This trend holds on ImageNet (left) as well as on the fine-grained benchmarks iNat-18 (middle left), iNat-21 (middle right), and SUN397 (right). Notably, the relationship is consistent across a wide range of SSL paradigms, joint-embedding methods, joint-predictive methods, and vision-language pretraining, suggesting that intrinsic dimension acts as a unifying geometric descriptor of representation quality, independent of the training objective.

This is further supported by the correlation metrics reported atop each plot. Both Kendall's $\tau \approx -0.6$ and Spearman's $\rho \approx -0.8$ confirm that IdEst reliably preserves the relative ordering of models across all four datasets, in strong agreement with linear probing rankings.

While our analysis has first focused on Intra-Dataset Correlation, a natural question arises: do these findings persist when IdEst and accuracy are measured on *different* datasets, or under alternative evaluation protocols?

In Figure 5.3, we study Inter-Dataset Correlation, where IdEst is computed on

ImageNet and accuracy is evaluated on four target datasets: the large-scale fine-grained benchmarks iNat-18 and iNat-21, and the smaller-scale datasets CIFAR-10 and CIFAR-100. The correlation remains strong across four target datasets, suggesting that IDEST captures intrinsic properties of the learned representations rather than dataset-specific characteristics.

In Figure 5.4, we examine IDEST under Alternative Evaluation Protocols. IDEST remains indicative of performance on ImageNet under kNN evaluation, and on the complementary ImageNet-v2 validation set.

Overall, these results reveal that low intrinsic dimension is a consistent feature of well-performing representations, and suggest that IDEST serves as a simple, label-free proxy for downstream evaluation (Figures 5.1, 5.3 and 5.4), offering insight into the quality of self-supervised representations without annotated data.

Finding 1. IdEst negatively correlates with linear probing accuracy: low intrinsic dimension is a consistent geometric signature of strong representations across intra- and inter-dataset settings and alternative evaluation protocols.

5.4.3 Training dynamics: offline and online probing

Here, we study whether IDEST tracks representation quality during unsupervised training. To this end, we consider two complementary evaluation protocols: *(i) offline linear probing* and *(ii) online probing*. While both aim to assess downstream performance over the course of self-supervised pretraining, they differ in when and how the classifier is trained.

Offline linear probing

Offline linear probing is the standard evaluation protocol in self-supervised learning. The representation model is first trained without labels, after which a linear classifier is trained on frozen features. To analyze training dynamics, we extract multiple checkpoints during pretraining and perform linear probing independently for each checkpoint.

Figure 5.5 (top) reports results for VICReg [12] (5.5a), DINO [26] (5.5b) and I-JEPA [8] (5.5c), respectively on ImageNet. We show linear probing top-1 accuracy (y axis on the left, in orange) as a function of training epochs, together with IDEST (y axis on the right, in blue). Across both models, IDEST closely follows the evolution of downstream accuracy throughout training. As representations improve and linear probing accuracy increases, the intrinsic dimension consistently decreases.

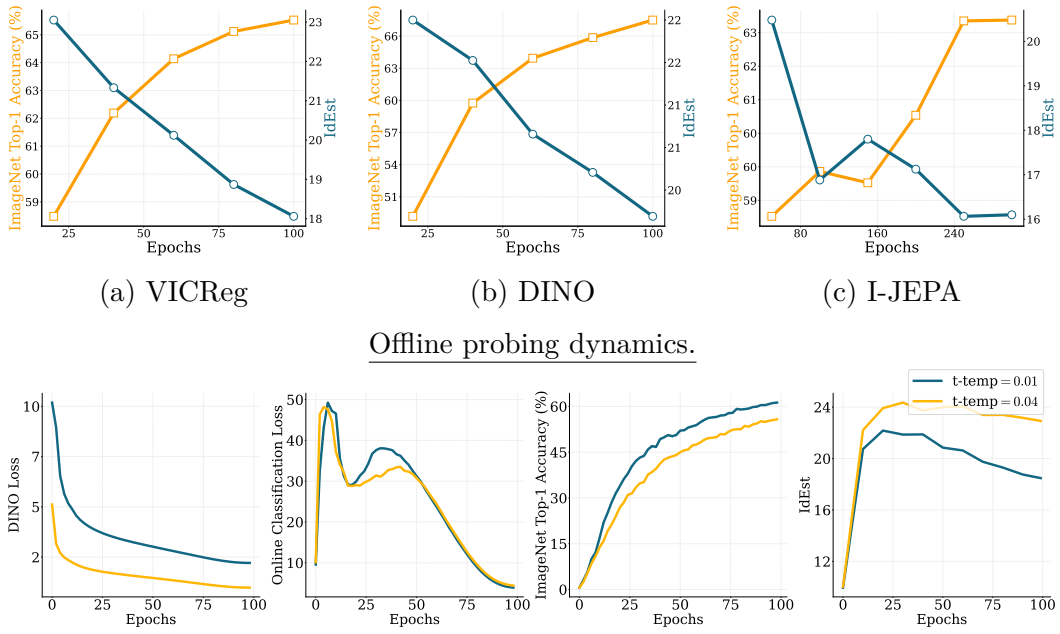


Figure 5.5: **Tracking training dynamics.** *(Top) Offline probing dynamics:* Evolution of ImageNet-1k linear probing top-1 accuracy (left y-axis, in orange) and IDEST (right y-axis, in blue) over self-supervised pretraining epochs for (a) VICReg (ResNet-50), (b) DINO (ViT-S), and (c) I-JEPA (ViT-B). As representations improve and linear probing accuracy increases, IDEST consistently decreases, demonstrating its ability to capture meaningful geometric changes during the evolution of the representation. *(Bottom) Online probing dynamics during self-supervised training:* Evolution of the self-supervised loss, classification loss, ImageNet-1k online classification top-1 accuracy, and IDEST over training epochs for DINO. While early-stage representations are highly constrained, IDEST progressively tracks improvements in downstream accuracy as training proceeds.

This strong temporal correlation indicates that IdEst captures meaningful geometric changes in the representation space as training progresses, without requiring labels.

Online linear probing

We consider online probing, where a linear classification head is attached to the representation and trained jointly during self-supervised pretraining. Importantly, gradients from the classifier do not backpropagate into the representation encoder, ensuring that the learned features remain purely self-supervised. This setting allows us to monitor downstream performance continuously during training.

Figure 5.5 (bottom) illustrates the training dynamics for DINO using a ViT-S, reporting the self-supervised loss (left), online classification loss and ImageNet top-1 accuracy (middle), and IdEst (right). We observe that during the early stages of training, particularly within the first 10 epochs, representations are highly constrained, rendering IdEst less informative. However, as training progresses, the self-supervised loss decreases and classification accuracy improves; concurrently, IdEst systematically drops, closely tracking the evolution of the other metrics. This demonstrates that IdEst faithfully reflects representation quality even in an online setting, capturing improvements in downstream performance as they emerge during the pretraining.

Overall, we observe that across both offline and online protocols, and across diverse SSL objectives, IdEst mirrors the evolution of downstream classification performance. These results further support its utility as a label-free indicator of representation quality, effective not only at convergence but throughout the pretraining process, once the initial training stages are surpassed. This leads us to our second finding:

Finding 2. IdEst can serve as a label-free proxy for SSL pretraining effectiveness: it decreases as downstream task accuracy improves over the course of unsupervised pre-training.

5.4.4 Label-free metric for model selection

Dataset	Method	Unsup.	VICReg (RN-50) lr	VICReg (RN-50) wd	VICReg (RN-50) var.	VICReg (RN-50) all	DINO (ViT-S) lr	DINO (ViT-S) s-temp.	DINO (ViT-S) t-temp.	DINO (ViT-S) all
ImageNet	ACC-1 Bounds		[62.2, 66.5]	[37.5, 69.1]	[65.5, 67.2]	[37.5, 69.1]	[63.6, 69.1]	[48.4,67.8]	[61.2 ,67.5]	[48.4, 69.1]
	α-ReQ	✓	65.0	53.5	**66.9**	53.5	63.6	58.6	61.2	58.6
	RankMe	✓	**66.5**	**69.1**	65.5	**69.1**	63.6	**67.8**	**67.5**	63.6
	LiDAR	×	65.0	**69.1**	66.8	65.0	**68.2**	65.6	65.0	**65.5**
	IDEST	✓	62.3	67.1	65.5	65.5	64.7	65.6	**67.5**	**65.5**
Fine-Grained	ImageNet Oracle	×	67.2	62.2	68.6	62.9	65.5	64.8	55.2	65.5
	α-ReQ	✓	66.7	47.8	**68.5**	66.5	63.8	59.4	60.6	59.4
	RankMe	✓	**67.2**	**62.2**	64.9	62.9	63.8	**64.8**	**64.5**	**63.8**
	LiDAR	×	66.7	**62.2**	66.6	**66.6**	68.3	62.4	64.3	62.4
	IDEST	✓	61.8	58.3	64.9	65.0	**63.9**	62.4	**64.5**	62.4

Dataset	Method	Unsup.	I-JEPA (ViT-B) lr	I-JEPA (ViT-B) target-size	I-JEPA (ViT-B) context-size	I-JEPA (ViT-B) all
ImageNet	ACC-1 Bounds		[61.9, 66.4]	[49.0, 66.4]	[61.3, 66.5]	[49.0,66.5]
	α-ReQ	✓	61.9	49.0	61.3	49.0
	RankMe	✓	61.9	55.9	61.3	61.9
	LiDAR	×	63.4	**66.4**	**66.0**	63.4
	IDEST	✓	**66.4**	**66.4**	**66.0**	**66.4**
Fine-Grained	ImageNet Oracle	×	60.0	60.7	58.3	60.0
	α-ReQ	✓	59.5	41.4	57.9	41.4
	RankMe	✓	59.5	56.9	57.9	59.5
	LiDAR	×	58.3	60.7	58.2	58.3
	IDEST	✓	**60.0**	**60.0**	**58.3**	**60.0**

Table 5.1: **Unsupervised model selection with IdEst.** We evaluate IDEST for hyperparameter selection against a supervised linear probe on ImageNet-1k, two unsupervised baselines: α-ReQ [4] and RankMe [67], and a weakly-supervised one: LiDAR [184]. For each SSL objective (and architecture), hyperparameters are jointly selected across according to Equation (5.2). 'Fine-grained' denotes the average performance across all furhter datasets (i.e. iNat-21,SUN, Aircraft, CUB, CIFAR-10) excluding ImageNet-1k. **Bold** values indicate the top-performing model selected by the criteria. IDEST is competitive with unsupervised and weakly-supervised baselines across SSL methods, including VICReg, despite RankMe directly mirroring its objective.

As stated in Theorem 5.3.1 and highlighted by previous experiments, performance on downstream tasks is bounded by the intrinsic dimension of the representations. Consequently, we evaluate IDEST as an hyperparameter selection criterion that may bypass the need for linear probing.

Selecting hyperparameters with `IdEst`

Given a set of candidate models $\mathcal{F} = \{f_1, \ldots, f_n\}$, each trained with varying hyperparameters, and their corresponding dimension estimates $\Delta = \{d_1, \ldots, d_n\}$, the selected model, f^*, is defined as:

$$f^* = f_{\arg\min_i d_i} \tag{5.2}$$

Set-up

We apply Equation (5.2) to identify the optimal model for a given hyperparameter configuration across several SSL frameworks including Joint-Embedding methods (VICReg [12], DINO [26]) and the Joint-Predictive architecture (I-JEPA [8]). Training details for each model are provided in Appendix 5.5.3.

We focus on various hyperparameters:

1. Optimization: learning rate (`lr`) and weight decay (`wd`);
2. Loss-specific coefficients: variance coefficients in VICReg (`var.`); the teacher and student temperatures in DINO (`t-temp.`, `s-temp.`);
3. Size of masking for I-JEPA: target block size (`target-size`) and context block size (`context-size`).

We evaluate performance on ImageNet and the average accuracy across several fine-grained classification datasets, comparing models selected by `IdEst` against those selected by supervised validation accuracy, and include the **`all`** setting, where selection is performed over the full hyperparameter pool, the more challenging configuration.

Results

As shown in Table 5.1, `IdEst` can recover most ImageNet oracle performance, achieving results near the upper bound of validation accuracy across various architectures (ResNets and ViTs) and pre-training objectives. Notably, `IdEst` outperforms α-ReQ in most settings without suffering from significant performance drops in the worst-case scenarios. For instance, α-ReQ selects the lower-bound accuracy for I-JEPA.

Additionally, while RankMe was originally designed for JE-SSL methods (e.g., VICReg, DINO), and in fact mirrors the VICReg objective by directly rewarding high effective rank, `IdEst` maintains competitive performance on these models while also generalizing to settings where RankMe struggles, such as I-JEPA. `IdEst` demonstrates versatility to other SSL paradigms. As I-JEPA is not a standard Joint-Embedding Method (making dimensional collapse less of a primary concern),

Model	Architecture	D	Param (M)	Linear Probe (min)	IdEst (min)
VICReg	ResNet-50	2048	24	64.5	2.1
DINOv2	ViT-S	384	22	65.8	1.8
	ViT-B	768	86	64.5	2.1
	ViT-L	1024	303	113.3	4.9
	ViT-G	1408	1000	322.7	12.8

Table 5.2: **Computational cost.** Wall-clock time (min) on ImageNet-1k to evaluate frozen representations using either linear probing (10 epochs, batch size 1024) or IDEST (single feature-extraction pass followed by $\dim_{\text{MST}}$ computation). Rows vary the backbone architecture, primarily changing the output dimension D and the number of parameters. Results are averages over 3 runs.

other metrics struggle, whereas IDEST successfully generalizes to this and other SSL paradigms.

Furthermore, LiDAR [184] is a strong baseline when pretraining augmentations are accessible. Nonetheless, across all four models in the `all` setting, IDEST achieves comparable model selection than LiDAR in a strictly unsupervised setting.

Finding 3. IdEst can serve as an efficient label-free criterion for hyperparameter selection: it performs comparably to supervised baselines across diverse architectures and SSL objectives.

5.4.5 Computational complexity and scalability

We evaluate the computational cost of IDEST and compare it to linear probing. Compared to training a linear probe, which requires multiple epochs over the training dataset features to obtain reasonable estimates of downstream accuracy, computing IDEST only requires feature extraction and a $\dim_{\text{MST}}$ computation. In Table 5.2, we compare the wall-clock compute time for IDEST and standard linear probe training on ImageNet (10 epochs across 2 GPUs), averaged over 3 runs. Computations were performed on H100 GPUs and an Intel Sapphire Rapids 8468 CPU. Across all architectures, IDEST is substantially faster than a 10-epoch linear probe ($B = 1024$, on 2 GPUs). Notably, for larger models, the cost is dominated by feature extraction, while the overhead of computing $\dim_{\text{MST}}$ remains minimal across various output dimensions.

5.4.6 Situating `IdEst` among collapse-tailored JE-SSL metrics

Prior chapters have shown that the dominant failure mode of *joint-embedding self-supervised learning* (JE-SSL) is *dimensional collapse* (see Figure 3.1): embeddings either degenerate to a single point or concentrate on a low-dimensional subspace, undermining the discriminative power of the representations. To diagnose and prevent this phenomenon, dedicated metrics have been proposed, centred on two properties: *uniformity* [202, 63] and *spectral diversity* [67]. These metrics have driven progress in JE-SSL methods [12] and directly motivated our proposed T-REGS (Chapter 3). Here, we ask whether they retain their diagnostic value beyond the JE-SSL setting, specifically, whether they can serve as reliable proxies for representation quality across a diverse set of foundation models.

Uniformity and spectral metrics

The intuition behind uniformity-metrics is that vectors should be roughly uniformly distributed on the unit hypersphere S^{m-1}, preserving as much information of the data as possible. Two main metrics have been proposed:

- $\mathcal{L}_u$ Wang and Isola [202], based on the gaussian pairwise kernel:

$$\mathcal{L}_u = \log \mathop{\mathbb{E}}_{\substack{\text{i.i.d.} \\ x,y \sim p_{\text{data}}}} \exp^{-t||f(x)-f(y)||_2^2},\ t > 0. \tag{5.3}$$

- $\mathcal{W}_2$ Fang et al. [64], the quadractic Wasserstein distance between the distribution of the learned representation and $\mathcal{N}(0, I_M/m)$:

$$\mathcal{W}_2 := \sqrt{||\hat{\mu}||_2^2 + 1 + \text{tr}(\hat{\Sigma}) - \frac{2}{m}\text{tr}(\hat{\Sigma}^{\frac{1}{2}})} \tag{5.4}$$

  where $\hat{\mu}, \hat{\Sigma}$ are the sample mean and covariance mean.

Neither metric was originally intended for cross-model comparison or hyperparameter selection; we nonetheless evaluate them on pretrained foundation models to test whether JE-SSL collapse diagnostics transfer as label-free proxies of representation quality.

RankMe [67] is formally the smooth rank measure, originally introduced by Roy

and Vetterli [156]:

$$\text{RankMe}(Z) = \exp\left(-\sum_{k=1}^{\min(N,K)} p_k \log p_k\right),$$
$$p_k = \frac{\sigma_k(Z)}{\|\sigma(Z)\|_1} + \varepsilon \tag{5.5}$$

where Z is the representations obtained.

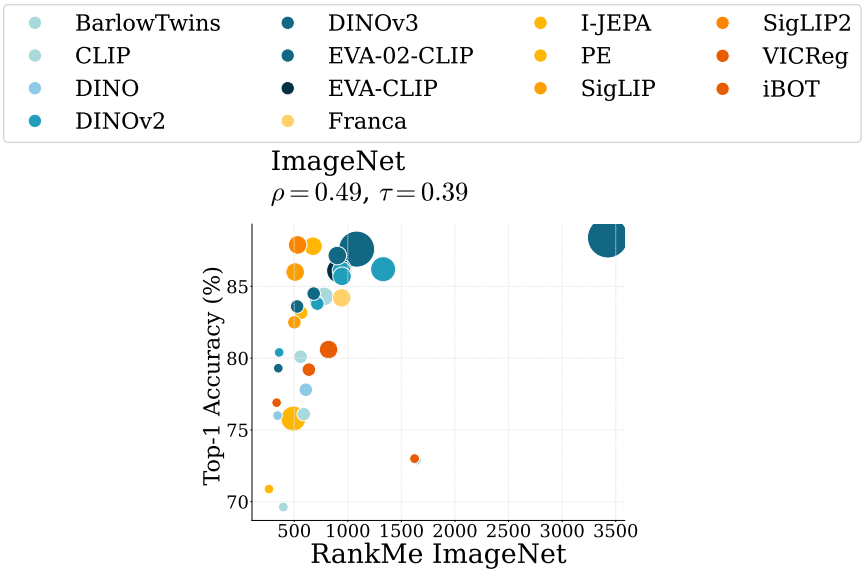
Uniformity and RankMe to study foundation models

In Figure 5.6 we study whether metrics initially proposed for JE-SSL can reflects representation quality across a diverse set of pretrained SSL models. Specifically, we compute RankMe, $\mathcal{L}_u$, $\mathcal{W}_2$ on frozen representations and compare it against standard linear probing accuracy on ImageNet.

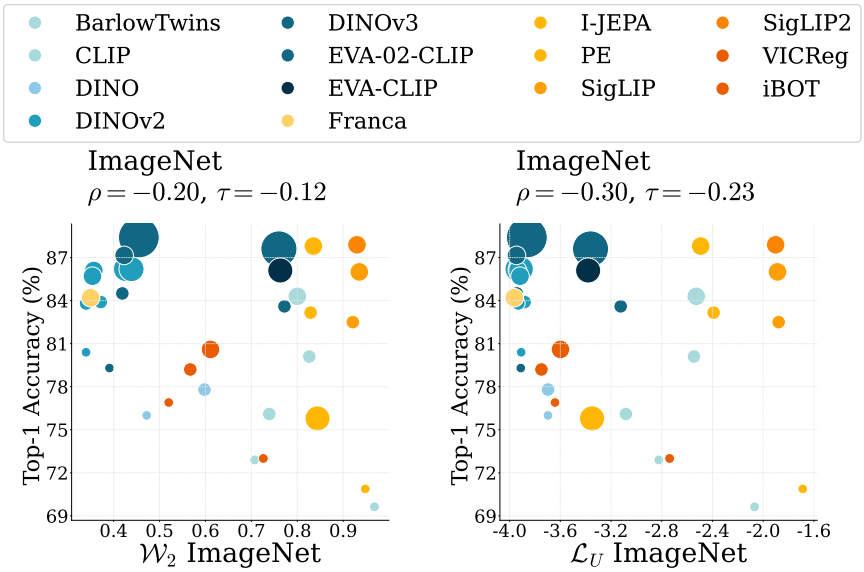
For RankMe [67], the original study noted that it is not suited for comparisons across architectures. A key reason is that RankMe is bounded by the output dimension of the model: two models of different architectures that both span their full representation space will yield different values, making cross-architecture comparisons unreliable. Figure 5.6a illustrates this limitation: DINOv3 ViT-7B is a clear outlier relative to DINOv3 ViT-L despite similar ImageNet accuracies, and models with comparable RankMe values (≈ 500) span linear probing accuracies from 70% to 87%.

Concerning uniformity metrics (see Figure 5.6b), neither $\mathcal{L}_u$ nor $\mathcal{W}_2$ correlates consistently with linear probing accuracy across models. Both remain only weakly aligned with accuracy across architectures, scales, and SSL paradigms. This is perhaps unsurprising: $\mathcal{L}_u$ and $\mathcal{W}_2$ were introduced as batch-level training diagnostics to prevent collapse in JE-SSL, not as cross-model selectors. On pretrained Foundation Models, collapse is largely already mitigated, so variation in uniformity mostly reflects objective-specific inductive biases rather than downstream usefulness. Multimodal joint-embedding models (e.g., PE-Core, SigLIP, SigLIP2) illustrate this point: they consistently rank among the least uniform models in Figure 5.6b, despite competitive linear probing performance. Prior work have related such non-uniformity to the *modality gap*, the separation of image and text subspaces in shared embedding spaces [121, 140, 211], a phenomenon we analyze further in Section 4.3, where low uniformity is shown to hinder cross-modal transferability. Uniformity metrics may thus conflate geometry induced by the training objective with representation quality, limiting their reliability as label-free selectors across SSL paradigms.

Overall, the three metrics retain some predictive signal, as reflected by non-trivial



(a) RankMe



(b) $\mathcal{L}_u$ and $\mathcal{W}_2$

Figure 5.6: **Alternative metrics and foundation models.** Linear probing accuracy on ImageNet versus *(Top)* RankMe and *(Bottom)* uniformity metrics ($\mathcal{L}_u$ and $\mathcal{W}_2$) for a diverse set of pretrained SSL models.

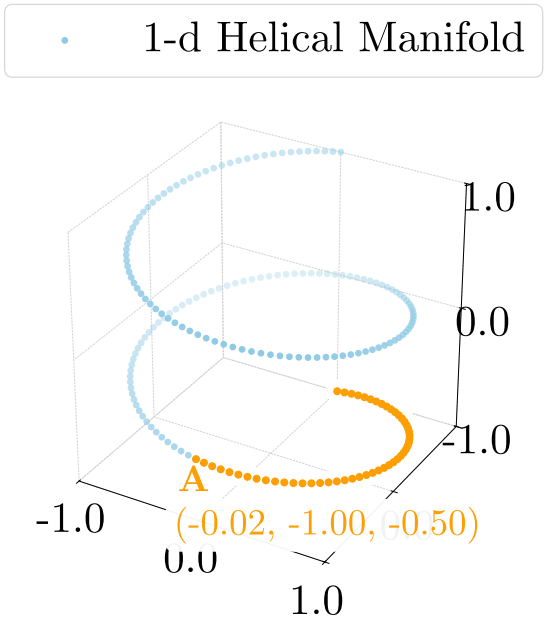


Figure 5.7: **Effective vs. intrinsic dimensionality.** A helix is embedded in a 3-dimensional ambient space (high effective dimension), yet is intrinsically a 1-dimensional manifold: point A requires three coordinates to specify its position in the ambient space, but a single arc-length coordinate suffices to locate it along the curve. The two notions thus capture complementary aspects of geometry, how spread out a representation is versus how many degrees of freedom it truly contains.

Kendall's τ and Spearman's ρ values, yet their limited correlation with accuracy motivates the search of alternative geometric proxies.

Effective and intrinsic dimensions

RankMe [67] and IDEST capture complementary notions of dimensionality. The former measures *effective rank*, the entropy of the singular-value distribution of the feature matrix, reflecting how uniformly the representation spreads across linear dimensions [156]. The latter estimates *intrinsic dimension*: the minimum number of degrees of freedom required to describe the data on its underlying manifold. Figure 5.7 illustrates this distinction with a helix, whose points require three ambient coordinates yet lie on a curve parameterized by a single value.

RankMe correlates less strongly with downstream accuracy across foundation models (Figure 5.6; $\rho = 0.49$, $\tau = 0.39$) than IDEST (Figure 5.1, $\rho = -0.74$, $\tau = -0.55$), confirming that intrinsic dimension captures information about representation quality beyond what linear spread alone can reveal. Yet RankMe's substantial

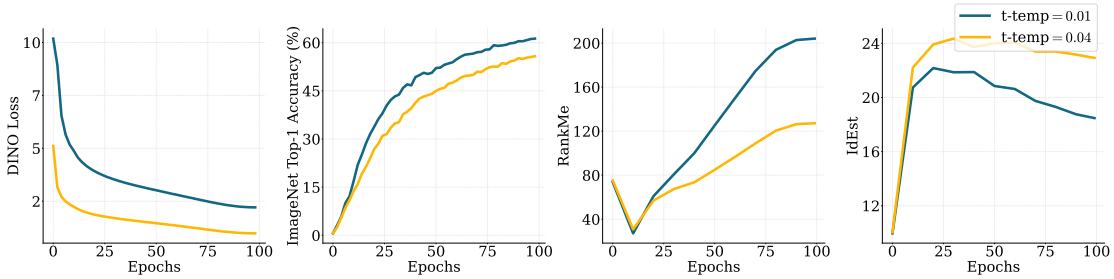


Figure 5.8: **Tracking training dynamics: RankMe and `IdEst`.** Evolution of the self-supervised loss (center), ImageNet-1k online classification top-1 accuracy (middle left), RankMe (middle right) and `IdEst` (right). `IdEst` decreases while RankMe increases throughout training.

correlation suggests that the effective rank of representations remains a useful proxy for downstream performance.

To further probe this distinction, we track the training dynamics of DINO (ViT-S) in Figure 5.8: `IdEst` decreases while RankMe increases throughout training. This inverse relationship admits a natural information-bottleneck interpretation: high-quality representations *compress* inputs onto a compact, low-dimensional manifold (low intrinsic dimension), while *spreading* embeddings uniformly across ambient dimensions (high effective rank) to prevent collapse and preserve discriminative information.

As hypothesized in [5], the gap between effective and intrinsic dimension relates to the *curvature* of the representation manifold: a flat manifold embedded in $\mathbb{R}^d$ has matching intrinsic and effective dimensions, whereas a highly curved one can occupy many ambient dimensions while remaining intrinsically low-dimensional, as illustrated by the helix in Figure 5.7.

5.5 Implementation Details

5.5.1 Image classification training details

Evaluation protocol

For each baseline, we follow the protocol of Siméoni et al. [168] and train a linear layer on the final frozen representation. To obtain the frozen representation, we follow each method's standard evaluation protocol, e.g., the `CLS` token after the layer norm, or `avgpool` if there is no `CLS` token. Specifically, we use SGD with a momentum of 0.9, and train for 10 epochs with a batch size of 1024, using random-resized-crop data. We perform the following grid search:

- **Learning Rate** in

  $\{0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, \quad 0.02, 0.05, 0.1, 0.2, 0.3, 0.5, 1\}$

For the fine-grained dataset (i.e., Aircraft), following [143], we use a lighter weight evaluation using scikit-learn's LogisticRegression implementation with the L-BFGS solver.

5.5.2 Models studied in Section 5.4.2

The checkpoints used were either downloaded from the original gihub or timm. The complete list of pretraiend models is reported in Table 5.3.

5.5.3 Models trained in Section 5.4.4

The complete list of models pretrained can be found below.

VICReg [12]. VICReg maximizes the informational content of embeddings by regularizing their empirical covariance matrix (see Section 2.2.4 for further details).

We pre-trained ResNet-50 for 100 epochs using LARS, the projector used is an MLP with intermediate dimensions (8192, 8192, 2048), with a batch size of 2048, following the protocol of Garrido et al. [67]

1. `lr`: wd $= 1e{-}6$, lr $\in \{0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5\}$, inv : 25, cov : 5, var : 25
2. `wd`: lr $= 0.3$, wd $\in \{1e{-}7,\ 1e{-}6,\ 1e{-}5,\ 1e{-}4,\ 1e{-}3\}$, inv : 25, cov : 5, var : 25
3. `cov`: lr $= 0.3$, wd $= 1e{-}6$, inv : 25, cov $\in \{0.4,\ 0.6,\ 0.8,\ 1,\ 4,\ 16\}$, var : 25

DINO [26]. DINO uses a student-teacher framework. Two versions of the same network (the student and the teacher) are fed different views of the same image. The student is trained to match the teacher's output probability distribution. To prevent the model from collapsing (i.e., giving the same output for every image), DINO uses a unique centering and sharpening operation on the teacher's outputs. The teacher's weights are updated as an exponential moving average of the student's weights (see Section 2.2.3 for further details).

We pre-trained ViT-S for 100 epochs using Adam-W. The projector used is an MLP with intermediate dimensions (8192, 8192, 32768), with a batch size of 2048:

1. `lr`: lr $\in \{1.25e-4, 2.25e-4, 0.0025, 0.002, 0.0075\}$, t-temp. $= 0.04$, s-temp $= 0.07$

Method	Architecture	Repo	Top-1 Accuracy (%) ImageNet	v2	SUN397	iNat-18	iNat-21
BarlowTwins [213]	ResNet-50	github	72.9	60.3	71.6	38.1	53.4
CLIP [151]	ResNet-50	timm	69.6	52.8	76.5	13.7	46.4
	ViT-B/16	timm	80.1	68.8	82.2	51.1	63.4
	ViT-B/32	timm	76.1	63.5	80.6	41.9	54.6
	ViT-L/14	timm	84.3	75.1	84.8	59.8	72.5
DINO [26]	ViT-S/16	github	76.0	64.4	72.5	48.4	60.9
	ViT-B/16	github	77.8	66.7	74.8	49.9	63.4
DINOv2 [143]	ViT-S/14	github	80.4	70.4	79.5	61.5	73.5
	ViT-B/14	github	83.8	74.9	81.7	68.3	80.3
	ViT-B/14-reg	github	83.9	75.1	81.9	67.4	79.4
	ViT-L/14	github	85.7	77.7	82.8	71.3	83.1
	ViT-L/14-reg	github	86.1	77.9	83.3	71.0	83.3
	ViT-G/14	github	86.2	78.1	82.9	74.0	85.3
DINOv3 [168]	ViT-S/16	github	79.3	68.7	79.1	61.8	69.4
	ViT-B/14	github	84.5	75.0	82.9	70.6	79.4
	ViT-L/16	github	87.2	79.5	84.6	76.1	84.2
	ViT-7B	github	88.4	81.4	85.4	80.2	88.7
EVA [178]	EVA01-g-14	timm	86.1	71.5	85.5	64.9	79.5
	EVA02-b-16	timm	83.6	73.3	83.1	55.7	70.5
	EVA02-E-14-plus	timm	87.6	75.6	86.5	70.6	82.7
Franca [197]	ViT-L/14 (In-21k)	github	84.2	75.4	80.7	73.5	70.6
iBoT [220]	ViT-S/16	github	76.9	65.6	73.6	48.1	60.6
	ViT-B/16	github	79.2	68.1	75.2	48.6	63.4
	ViT-L/16	github	80.6	69.9	76.4	51.5	66.2
I-JEPA [8]	ViT-G/16	github	75.8	63.8	74.9	45.9	58.7
PE-Core [18]	S16-336	timm	70.9	51.2	75.6	38.6	54.6
	B14-224	timm	83.2	72.7	83.8	60.8	74.9
	L14-336	timm	87.8	78.7	87.0	74.1	84.1
SigLIP [214]	ViT-B-16-SigLIP	timm	82.5	68.9	82.7	53.3	69.6
	ViT-L-16-SigLIP-256	timm	86.0	73.8	85.1	59.4	70.8
VICReg [12]	ResNet50	github	73.0	59.9	71.7	40.2	53.8

Table 5.3: **Foundation models studied.** Overview of the pretrained models evaluated in this work, spanning diverse SSL and vision-language objectives and architectures (ResNet, ViT). Top-1 accuracy is reported on ImageNet, ImageNet-v2, and additional fine-grained datasets SUN397, iNat-18, iNat-21. Model weights are loaded from the official repositories or `timm`.

2. `s-temp`: lr $= 0.002$, t-temp. $= 0.04$, s-temp $= \{0.07, 0.1, 0.2, 0.3, 0.4\}$,
3. `t-temp`: lr $= 0.002$, t-temp. $= \{0.01, 0.02, 0.03, 0.04, 0.05\}$, s-temp $= 0.07$,

I-JEPA [8] I-JEPA is a Joint-Predictive Architecture. It uses a **context block** to predict the representations of several target blocks from the same image. The context encoder is a Vision Transformer (ViT), which only processes the visible context patches. The predictor is a smaller ViT which takes the context encoder output and, conditioned on positional tokens, predicts the representations of a **target block** at a specific location. The weights of the target encoder are updated at each iteration via an exponential moving average of the context encoder weights (see Section 2.3.1 for further details).

We pre-trained ViT-B for 300 epochs with the same protocol as in the original papers with a batch size of 4096:

1. `lr`: wd sch $=$ [0.04–0.4], lr $\in \{4e{-}5,\ 8e{-}5,\ 9e{-}5,\ 1e{-}4,\ 1.25e{-}4,\ 2e{-}4,\ 3e{-}4\}$, target block size $= \{0.15, 0.2\}$, context block size $= \{0.85, 1.0\}$
2. `Target Size Block` (the size of the target block): wd sch $=$ [0.04–0.4], lr $= 1.25e{-}5$, target block size $\in \{\{0.075, 0.2\},\ \{0.1, 0.2\},\ \{0.125, 0.2\},\ \{0.2, 0.25\}\}$, context block size $= \{0.85, 1.0\}$
3. `Context Size` (the size of the context block): wd sch $=$ [0.04–0.4], lr $= 1.25e{-}5$, target block size $= \{0.15, 0.2\}$, context block size $\in \{\{0.4, 1.0\},\ \{0.5, 1.0\},\ \{0.65, 1.0\},\ \{0.75, 1.0\},\ \{0.90, 1.0\}\}$

5.6 Computational Cost

All trainings were performed on H100 GPUs. The total computational cost of the project, including training baselines, experiments, and ablation studies, amounts to approximately 12,000 GPU-hours.

5.7 Discussion

While versatile, IDEST has several limitations.

The formula in Theorem 5.3.1, provides only a bound on the convergence rate. Therefore, even if two models have similar intrinsic dimensions, their actual convergence rates may potentially differ in practice. Thus, IDEST should be viewed more as an indicator of accuracy than as a perfect predictor. This is indeed reflected in Figure 5.1 for instance, where $\rho \approx 0.8, \tau \approx 0.6$ confirms strong global

ranking ability, while acknowledging that the intrinsic dimension might not explain all the variance. In particular, ranking is inherently harder when model accuracies are tightly clustered, as in ImageNet or SUN397, than when they are more spread out, as in iNaturalist. Furthermore, one family deviates most visibly: vision-language models (e.g., CLIP, EVA-CLIP), where the cone effect introduces geometric misalignment between encoders that the contrastive objective preserves rather than resolves, constraining the representations [121].

Additionally, a more subtle limitation stems from the early-training regime of ViTs: IDEST is less informative during the first 10 training epochs, before representations develop stable geometric structure (Section 5.4.3).

5.8 Conclusion

In this chapter, we introduced IDEST, an unsupervised criterion for evaluating self-supervised representations based on the estimation of the intrinsic dimension via minimum spanning trees ($\dim_{\text{MST}}$). Building on the theoretical connection between intrinsic dimension and generalization, we demonstrated that IDEST serves as a robust and an efficient geometric proxy for downstream performance in SSL.

Our empirical evaluation across diverse datasets, architectures, and SSL objectives shows that IDEST correlates with supervised linear probing accuracy. Furthermore, IDEST provides a principled label-free metric for hyperparameter selection, performing on par with supervised oracles and generalizing across heterogeneous SSL paradigms.

By offering a unified framework for assessing SSL representations without requiring annotated data, our work highlights the potential of intrinsic geometric descriptors to complement standard evaluation protocols.

Chapter 6

Conclusion

6.1 Summary

This thesis aimed at capturing meaningful geometric descriptors suited to self-supervised learning along two critical axes: *shaping* representations against dimensional collapse, and *evaluating* representations beyond downstream performances. The work presented here demonstrates that Minimum Spanning Trees (MSTs) provide a powerful and versatile geometric tool for both objectives.

Our first line of contributions focused on the training side. Chapter 3 introduced T-REGS, which recasts the uniformity objective in joint-embedding SSL as the maximization of the MST length under a sphere constraint. Our analysis connects the optimization of MST length to entropy maximization and uniformity on compact manifolds, providing theoretical guarantees corroborated by empirical results. T-REGS thus enforces a geometrically grounded notion of uniformity that goes beyond existing formulations, consistently improving representation quality in JE-SSL. Second, Chapter 4 showed that these principles can be extended to the joint-embedding multimodal setting where a specific form of collapse arises: the modality gap. As a concrete biomedical application, we introduced ADAPT, which detects physiological changes from wearable sensors under missing modalities by anchoring representations to the strongest modality and closing the modality gap via T-REGS.

Finally, Chapter 5 introduced `IdEst`, completing this picture by turning the same geometric lens from a training tool into an evaluation criterion. A natural question *does a well-shaped representation space signal downstream quality?* finds a

concrete answer: IdEst establishes that the intrinsic dimension of SSL representations, estimated via $\dim_{\text{MST}}$, is a reliable, paradigm-agnostic proxy for downstream performance, replacing costly linear probing with a single unsupervised metric.

Together, these contributions reveal a common thread: geometric structure in the representation space can be harnessed both to guide training, and to assess representation, without relying on labels. Our work naturally gives rise to further discussion and avenues for future research.

6.2 Discussion

6.2.1 On the role of minimum spanning trees

```python
from scipy.spatial.distance import pdist, squareform
from scipy.sparse.csgraph import minimum_spanning_tree

dists = squareform(pdist(z))  # pairwise distances
mst = minimum_spanning_tree(dists)
length =  mst.sum()
```

MST length via `scipy`

```python
from ripser import ripser

pd = ripser(z, maxdim=0)["dgms"][0]
# remove the infinite bar (last connected component)
finite = pd[pd[:, 1] != np.inf]
length = (finite[:, 1] - finite[:, 0]).sum()
```

MST length via `ripser`

Figure 6.1: Two equivalent implementations of the MST length computation used throughout this thesis. *Top*: graph-based construction via (`scipy`). *Bottom*: topological construction via 0-dimensional persistent homology (`ripser`); the MST length equals the sum of finite bar lifetimes in the dimension-0 persistence diagram.

Throughout this thesis, the MST has served as a unifying computational tool across two distinct roles: *shaping* representations during training (Chapters 3 and 4) and *evaluating* them post hoc (Chapter 5). This dual role reflects genuine structural advantages of the MST. First, the length of the MST is differentiable and fast to compute, making it practical both as a training objective and as an evaluation tool. Moreover, several implementations exist making it straightforward to integrate into any existing pipeline (Figure 6.1). Second, and more fundamentally, it encodes intrinsic geometric properties of the underlying data distribution: asymptotically,

the MST length scales with the Rényi entropy of that distribution, providing a principled, sample-based way to characterize a measure.

At the same time, the MST is fundamentally a *zero-dimensional* topological summary: it captures spanning geometry and connectivity, but is insensitive to higher-order structure such as loops, voids, or the curvature of the representation manifold. Richer tools, for instance persistent homology in degrees ≥ 1 or Riemannian geometry approaches, could complement the MST by reflecting further geometric characteristics.

However, an important practical consideration is that moving beyond the MST typically comes at a steep computational cost. Computing higher-dimensional persistent homology or curvature estimates at the scale of modern SSL datasets remains non-trivial. The role of the MST in this thesis can therefore be understood as a principled and tractable first step towards a more complete geometric characterization of self-supervised representations.

6.2.2 On the interplay between identifying and enforcing geometric properties

The works of Wang and Isola [202] and Fang et al. [64] were pivotal in establishing uniformity as a key element of JE-SSL effectiveness, opening the door to a line of subsequent work aimed at explicitly enforcing such uniformity, most notably yielding our proposed T-REGS. This trajectory raises a broader question: when should an identified geometric property also become an explicit training objective?

We believe that explicit regularization is most justified when it addresses a known *limitation* of the learning objective. Uniformity, for instance, fits this criterion precisely: in joint-embedding architectures, it helps prevent dimensional and representational collapse, the central challenge whereby representations tend to respectively span a lower-dimensional subspace, or collapse to a single point. T-REGS is therefore not designed to enforce a geometric property for its own sake, but to provide a targeted regularization against a specific degenerate regime.

The findings of `IDEST`, by contrast, fall into a different category. Intrinsic dimension is not indicative of a failure mode per se, but rather a *quality indicator*: it reflects how well a given objective has succeeded at organizing representations, rather than signalling a degeneracy to be corrected. Explicitly optimizing for lower intrinsic dimension would therefore not remedy a failure mode of an underlying method, but risk conflating cause and effect: a lower intrinsic dimension is a consequence of good representations, not a target whose direct enforcement guarantees them. This distinction, between properties that *correct* and properties that *char-*

acterize, may offer a more principled basis for deciding when geometric objectives are warranted, and when they are better left as diagnostic tools.

6.3 Perspectives

The research presented in this thesis opens several promising avenues for future investigation.

Lead 1: Towards tightening numerical bounds for the Tammes problem

The objective of T-REGS, presented in Chapter 3, is to uniformly distribute points on the sphere. As discussed in Section 3.1, this objective connects naturally to the *Tammes problem* in combinatorial geometry: given $n \geq d+1$, find the greatest minimum pairwise distance achievable when placing n points on the sphere S^{d-1} [180].

This problem is notoriously difficult: in dimension 3, the exact optimum is known only for $n \leq 14$ or $n = 24$ points [39], and the dual sphere-packing problem (how many non-overlapping caps of radius r fit on the sphere) is similarly unsolved except in a handful of dimensions [38]. Any point configuration on the sphere immediately yields a lower bound on the Tammes solution via its minimum pairwise distance, so improving bounds amounts to finding near-optimal configurations via optimization.

A key challenge is the choice of objective: minimizing the minimum pairwise distance directly leads to very sparse gradients and slow convergence, whereas the MST length (alternatively, the total persistence) offers a natural compromise with gradients whose support scales linearly in the number of points. A future research direction could therefore be to first optimize functions whose gradients have large supports (e.g., sum of pairwise distances), then, using warm starts, to optimize other functions with more restricted supports (e.g., the MST length).

Lead 2: Asymptotic regime of the Tammes problem

Another regime in which the Tammes problem may become more tractable is the asymptotic regime $n \to \infty$. For the MST length, equivalently, total persistence, the work of Costa and Hero [40] establishes a connection, at the limit, with a geodesic version of the Rényi entropy, for which the uniform distribution on the sphere is the maximizer [136]. In other words, maximizing the MST length asymptotically drives point clouds towards uniform coverage of the sphere, the theoretical grounding underpinning T-REGS.

The analogous question for *minimal* persistence, the length of the shortest edge in the MST, remains open. A promising direction would be to characterize its asymptotic behaviour and to identify the optimal spreading distribution for this objective. Comparing that distribution to the uniform measure would quantify the loss incurred by substituting the MST for the exact Tammes objective, thereby providing a rigorous theoretical grounding for the relaxation underlying T-REGS. A further avenue would be to study the connections between minimal persistence and other classical functionals used to numerically bound the Tammes problem, such as Coulomb-type or Riesz potentials [114, 105], whose asymptotic behaviour is comparatively better understood and could offer additional analytical leverage.

Lead 3: Differential geometry to study SSL representations

In Chapter 5, we showed that intrinsic dimension provides a meaningful lens for assessing the quality of SSL representations on downstream tasks. As discussed in Section 5.4.6, the interplay between intrinsic and effective dimension, as captured by metrics such as RankMe [67], merits closer investigation. A pronounced gap between the two may indicate curvature in the representation manifold, suggesting that the learned geometry departs from a flat Euclidean structure. Differential geometry offers a framework to formalize this notion; incorporating curvature-aware metrics is therefore a natural direction for future work aimed at better characterizing the structure of SSL representations.

Lead 4: Beyond SSL, studying generative representations

While IDEST (Chapter 5) provides insights into the quality of SSL representations for classification tasks, extending this analysis to dense tasks, such as segmentation or generation, is a natural next step, as the underlying MST estimator is task-agnostic.

For generative modelling in particular, several representation spaces are natural candidates for study. In diffusion and flow matching models, the intermediate activations of the denoising or velocity-field network $v_\theta(x_t, t)$ vary continuously along the generative trajectory, raising the question of how intrinsic dimension evolves as $t \to 0$. In latent generative models (e.g., latent diffusion, FLUX), the VAE encoder's latent representation is itself a structured representation amenable to MST-based analysis. That said, the theoretical framework underlying IDEST is tailored to global vector representations extracted from frozen encoders, as typically used for linear classification, rather than dense ones. Generative models, however, produce representations of the latter kind, spatially arranged, sequence-structured, and high-dimensional in a token sense, which fall outside the theoretical framing as currently stated and would require a dedicated treatment.

Bibliography

[1] Henry Adams, Manuchehr Aminian, Elin Farnell, et al. "A Fractal Dimension for Measures via Persistent Homology". In: *Topological Data Analysis*. 2020.

[2] Henry Adams, Manuchehr Aminian, Elin Farnell, et al. "A fractal dimension for measures via persistent homology". In: *Topological Data Analysis: The Abel Symposium 2018*. 2020.

[3] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. "Intrinsic dimensionality explains the effectiveness of language model fine-tuning". In: *ACL-IJCNLP*. 2021.

[4] Kumar K Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. "α-ReQ: Assessing Representation Quality in Self-Supervised Learning by measuring eigenspectrum decay". In: *NeurIPS* (2022).

[5] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. "Intrinsic dimension of data representations in deep neural networks". In: *NeurIPS* (2019).

[6] Tom M Apostol and Mamikon A Mnatsakanian. *New horizons in geometry*. American Mathematical Soc., 2017.

[7] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. "Self-labelling via simultaneous clustering and representation learning". In: *ICLR* (2020).

[8] Mahmoud Assran, Quentin Duval, Ishan Misra, et al. "Self-supervised learning from images with a joint-embedding predictive architecture". In: *CVPR*. 2023.

[9] Guillaume Astruc, Nicolas Dufour, Ioannis Siglidis, et al. "OpenStreetView-5M: The Many Roads to Global Visual Geolocation". In: *CVPR* (2024).

[10] Reza Azad, Nika Khosravi, Mohammad Dehghanmanshadi, Julien Cohen-Adad, and Dorit Merhof. "Medical image segmentation on MRI images with missing modalities: A review". In: *arXiv* (2022).

[11] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. "BEiT: BERT Pre-Training of Image Transformers". In: *ICLR*. 2022.

[12] Adrien Bardes, Jean Ponce, and Yann Lecun. "VICReg: Variance-Invariance-Covariance Regularization For Self-Supervised Learning". In: *ICLR*. 2022.

[13] Ulrich Bauer. “Ripser: efficient computation of Vietoris–Rips persistence barcodes”. In: *Journal of Applied and Computational Topology* (2021).

[14] Suzanna Becker and Geoffrey E Hinton. “Self-organizing neural network that discovers surfaces in random-dot stereograms”. In: *Nature* (1992).

[15] Robert Bennett. “The intrinsic dimensionality of signal collections”. In: *IEEE Transactions on Information Theory* (1969).

[16] James AD Binnie, Paweł Dłotko, John Harvey, Jakub Malinowski, and Ka Man Yim. “A survey of dimension estimation methods”. In: *arXiv* (2025).

[17] Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. “Intrinsic dimension, persistent homology and generalization in neural networks”. In: *NeurIPS* (2021).

[18] Daniel Bolya, Po-Yao Huang, Peize Sun, et al. “Perception encoder: The best visual embeddings are not at the output of the network”. In: *arXiv* (2025).

[19] P. A. Bromiley, N. A. Thacker, and E. Bouhova-Thacker. “Shannon Entropy, Renyi Entropy, and Information”. In: *Statistics and Inf. Series (2004-004)* (2004).

[20] M. Dixit C. Hofer R. Kwitt and M. Niethammer. “Connectivity-Optimized Representation Learning via Persistent Homology”. In: *Proc. ICML*. 2019.

[21] Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. “Isotropy in the contextual embedding space: Clusters and manifolds”. In: *ICLR*. 2021.

[22] John Canny. “A computational approach to edge detection”. In: *IEEE PAMI* (1986).

[23] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. “Deep clustering for unsupervised learning of visual features”. In: *ECCV*. 2018.

[24] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. “Unsupervised pre-training of image features on non-curated data”. In: *ICCV*. 2019.

[25] Mathilde Caron, Ishan Misra, Julien Mairal, et al. “Unsupervised learning of visual features by contrasting cluster assignments”. In: *NeurIPS* (2020).

[26] Mathilde Caron, Hugo Touvron, Ishan Misra, et al. “Emerging properties in self-supervised vision transformers”. In: *CVPR*. 2021.

[27] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. “Human pose estimation with iterative error feedback”. In: *CVPR*. 2016.

[28] Mathieu Carriere, Frédéric Chazal, Marc Glisse, et al. “Optimizing persistent homology-based functions”. In: *Proc. ICML*. 2021.

[29] Hava Chaptoukaev, Valeriya Strizhkova, Michele Panariello, et al. “StressID: a Multimodal Dataset for Stress Identification”. In: *NeurIPS*. 2023.

[30] Frédéric Chazal, Vin De Silva, and Steve Oudot. “Persistence stability for geometric complexes”. In: *Geometriae Dedicata* (2014).

[31] Jeff Cheeger. “Differentiability of Lipschitz functions on metric measure spaces”. In: *Geometric & Functional Analysis GAFA* (1999).

[32] Cheng Chen, Qi Dou, Yueming Jin, Quande Liu, and Pheng Ann Heng. “Learning with privileged multimodal knowledge for unimodal segmentation”. In: *IEEE TMI* (2021).

[33] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. “Crossvit: Cross-attention multi-scale vision transformer for image classification”. In: *CVPR*. 2021.

[34] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. “A simple framework for contrastive learning of visual representations”. In: *Proc. ICML*. 2020.

[35] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. “Improved baselines with momentum contrastive learning”. In: *arXiv* (2020).

[36] Xinlei Chen and Kaiming He. “Exploring simple siamese representation learning”. In: *CVPR*. 2021.

[37] Xinlei Chen, Saining Xie, and Kaiming He. “An empirical study of training self-supervised vision transformers”. In: *ICCV*. 2021.

[38] Henry Cohn. *Kissing numbers*. https://cohn.mit.edu/kissing-numbers/. Accessed Feb. 16, 2026. 2026.

[39] Henry Cohn. *Spherical codes*. https://cohn.mit.edu/spherical-codes/. Accessed Feb. 16, 2026. 2026.

[40] Jose A. Costa and Alfred O. Hero. “Determining Intrinsic Dimension and Entropy of High-Dimensional Shape Spaces”. In: *Statistics and Analysis of Shapes*. 2006.

[41] Paul Couairon, Loick Chambon, Louis Serrano, et al. “JAFAR: Jack up any feature at any resolution”. In: *NeurIPS* (2025).

[42] George Cybenko. “Approximation by superpositions of a sigmoidal function”. In: *Mathematics of control, signals and systems* (1989).

[43] 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”. In: *J. Machine Learning Research* (2022).

[44] Navneet Dalal and Bill Triggs. “Histograms of oriented gradients for human detection”. In: *CVPR*. 2005.

[45] Lucas Degeorge, Arijit Ghosh, Nicolas Dufour, David Picard, and Vicky Kalogeiton. "How far can we go with ImageNet for Text-to-Image generation?" In: *arXiv* (2025).

[46] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, et al. "Scaling vision transformers to 22 billion parameters". In: *Proc. ICML*. 2023.

[47] Jia Deng, Wei Dong, Richard Socher, et al. "Imagenet: A large-scale hierarchical image database". In: *CVPR*. 2009.

[48] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *NAACL*. 2019.

[49] Joel E Dimsdale. "Psychological stress and cardiovascular disease". In: *Journal of the American College of Cardiology* (2008).

[50] Carl Doersch, Abhinav Gupta, and Alexei A Efros. "Unsupervised visual representation learning by context prediction". In: *CVPR*. 2015.

[51] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. "Adversarial feature learning". In: *arXiv* (2016).

[52] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv* (2020).

[53] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *ICLR*. 2021.

[54] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. "Discriminative unsupervised feature learning with convolutional neural networks". In: *NeurIPS* (2014).

[55] Nicolas Dufour, David Picard, Vicky Kalogeiton, and Loic Landrieu. "Around the World in 80 Timesteps: A Generative Approach to Global Visual Geolocation". In: *CVPR* (2025).

[56] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, et al. "Adversarially learned inference". In: *arXiv* (2016).

[57] Benjamin Dupuis, George Deligiannidis, and Umut Simsekli. "Generalization bounds using data-dependent fractal dimensions". In: *Proc. ICML*. 2023.

[58] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

[59] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. "Clap learning audio concepts from natural language supervision". In: *Proc. ICASSP*. 2023.

[60] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. "Whitening for self-supervised representation learning". In: *Proc. ICML*. 2021.

[61] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. "Estimating the intrinsic dimension of datasets by a minimal neighborhood information". In: *Scientific reports* (2017).

[62] Kenneth Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. 2013.

[63] Xianghong Fang, Jian Li, Qiang Sun, and Benyou Wang. "Rethinking the Uniformity Metric in Self-Supervised Learning". In: *ICLR*. 2024.

[64] Xianghong Fang, Jian Li, Qiang Sun, and Benyou Wang. "Rethinking the uniformity metric in self-supervised learning". In: *ICLR*. 2024.

[65] Kunihiko Fukushima. "Visual Feature Extraction by a Multilayered Network of Analog Threshold Elements". In: *Transactions on Systems Science and Cybernetics* (1969).

[66] Tianyu Gao, Xingcheng Yao, and Danqi Chen. "SimCSE: Simple Contrastive Learning of Sentence Embeddings". In: *EMNLP*. 2021.

[67] Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. "Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank". In: *Proc. ICML*. 2023.

[68] Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann Lecun. "On the duality between contrastive and non-contrastive self-supervised learning". In: *ICLR*. 2023.

[69] Chongjian GE, Jiangliu Wang, Zhan Tong, et al. "Soft Neighbors are Positive Supporters in Contrastive Visual Representation Learning". In: *ICLR*. 2023.

[70] Ridouane Ghermi, Xi Wang, Vicky Kalogeiton, and Ivan Laptev. "Long Story Short: Story-level Video Understanding from 20K Short Films". In: *arXiv* (2024).

[71] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. "Unsupervised representation learning by predicting image rotations". In: *arXiv* (2018).

[72] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, et al. "Imagebind: One embedding space to bind them all". In: *CVPR*. 2023.

[73] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press Cambridge, 2016.

[74] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. "Generative adversarial nets". In: *NeurIPS* (2014).

[75] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, et al. “Self-supervised pretraining of visual features in the wild”. In: *arXiv* (2021).

[76] Priya Goyal, Quentin Duval, Isaac Seessel, et al. “Vision models are more robust and fair when pretrained on uncurated images without supervision”. In: *arXiv* (2022).

[77] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. “Scaling and benchmarking self-supervised visual representation learning”. In: *ICCV*. 2019.

[78] Jean-Bastien Grill, Florian Strub, Florent Altché, et al. “Bootstrap your own latent-a new approach to self-supervised learning”. In: *NeurIPS* (2020).

[79] Michael Gutmann and Aapo Hyvärinen. “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models”. In: *AISTATS*. 2010.

[80] Xintian Han, Yuxuan Hu, Luca Foschini, et al. “Deep learning models for electrocardiograms are susceptible to adversarial attack”. In: *Nature medicine* (2020).

[81] Bobby He and Mete Ozay. “Exploring the gap between collapsed & whitened features in self-supervised learning”. In: *Proc. ICML*. 2022.

[82] Kaiming He, Xinlei Chen, Saining Xie, et al. “Masked autoencoders are scalable vision learners”. In: *CVPR*. 2022.

[83] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. “Momentum contrast for unsupervised visual representation learning”. In: *CVPR*. 2020.

[84] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”. In: *CVPR*. 2016.

[85] Geoffrey E Hinton and Ruslan R Salakhutdinov. “Reducing the dimensionality of data with neural networks”. In: *science* (2006).

[86] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, et al. “Learning deep representations by mutual information estimation and maximization”. In: *ICLR*. 2018.

[87] Toshiteru Homma, Les E Atlas, and Robert J Marks. “An artificial neural network for spatio-temporal bipolar patterns: Application to phoneme classification”. In: *NeurIPS*. 1987.

[88] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. “Multilayer feedforward networks are universal approximators”. In: *Neural networks* (1989).

[89] Minhao Hu, Matthis Maillard, Ya Zhang, et al. “Knowledge distillation from multi-modal to mono-modal segmentation networks”. In: *MICCAI*. 2020.

[90] Tianyu Hua, Wenxiao Wang, Zihui Xue, et al. “On feature decorrelation in self-supervised learning”. In: *CVPR*. 2021.

[91] Lun Huang, Qiang Qiu, and Guillermo Sapiro. “SSOLE: Rethinking Orthogonal Low-rank Embedding for Self-Supervised Learning”. In: *ICLR*. 2025.

[92] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. “Deep learning for time series classification: a review”. In: *Data mining and knowledge discovery* (2019).

[93] Andrew Jaegle, Felix Gimeno, Andy Brock, et al. “Perceiver: General perception with iterative attention”. In: *Proc. ICML*. 2021.

[94] C. Jia, Y. Yang, Y. Xia, et al. “Scaling up visual and vision-language representation learning with noisy text supervision”. In: *Proc. ICML*. 2021.

[95] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. “Understanding dimensional collapse in contrastive self-supervised learning”. In: *arXiv* (2021).

[96] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. “Understanding Dimensional Collapse in Contrastive Self-supervised Learning”. In: *ICLR*. 2022.

[97] Kerstin Johnsson, Charlotte Soneson, and Magnus Fontes. “Low bias local intrinsic dimension estimation from expected simplex skewness”. In: *IEEE PAMI* (2014).

[98] Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. “Learning visual features from large weakly supervised data”. In: *ECCV*. 2016.

[99] Maurice G Kendall. “A new measure of rank correlation”. In: *Biometrika* (1938).

[100] Agnan Kessy, Alex Lewin, and Korbinian Strimmer. “Optimal whitening and decorrelation”. In: *The American Statistician* (2018).

[101] Jeonghyeon Kim and Sangheum Hwang. “Enhanced OoD Detection through Cross-Modal Alignment of Multi-Modal Representations”. In: *CVPR*. 2025.

[102] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, et al. “Big transfer (bit): General visual representation learning”. In: *ECCV*. 2020.

[103] Aishik Konwer, Xiaoling Hu, Joseph Bae, et al. “Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation”. In: *CVPR*. 2023.

[104] Nicholas Konz and Maciej A Mazurowski. “The Effect of Intrinsic Dataset Properties on Generalization: Unraveling Learning Differences Between Natural and Medical Images”. In: *ICLR*. 2024.

[105] David A. Kottwitz. "The densest packing of equal circles on a sphere". In: *Acta Crystallographica Section A* (1991).

[106] Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of features from tiny images". In: *ICLR* (2009).

[107] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *NeurIPS* (2012).

[108] Joseph B Kruskal. "On the shortest spanning subtree of a graph and the traveling salesman problem". In: *Proc of the American Mathematical Society* (1956).

[109] Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. "Temperature Schedules for self-supervised contrastive methods on long-tail data". In: *ICLR*. 2023.

[110] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. "Learning representations for automatic colorization". In: *ECCV*. 2016.

[111] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. "Object recognition with gradient-based learning". In: *Shape, contour and grouping in computer vision*. 1999.

[112] Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. "Shared global and local geometry of language model embeddings". In: *arXiv* (2025).

[113] Kibok Lee, Yian Zhu, Kihyuk Sohn, et al. "*i*-Mix: A Domain-Agnostic Strategy for Contrastive Representation Learning". In: *ICLR*. 2021.

[114] John Leech. "Equilibrium of sets of particles on a sphere". In: *The Mathematical Gazette* (1957).

[115] Elizaveta Levina and Peter Bickel. "Maximum likelihood estimation of intrinsic dimension". In: *NeurIPS* (2004).

[116] Jacob Leygonie, Steve Oudot, and Ulrike Tillmann. "A framework for differential calculus on persistence barcodes". In: *Foundations of Computational Mathematics* (2022).

[117] Alexander C Li, Alexei A Efros, and Deepak Pathak. "Understanding collapse in non-contrastive siamese representation learning". In: *ECCV*. 2022.

[118] Fei-Fei Li. "Crowdsourcing, benchmarking & other cool things". In: *CMU VASC Semin* (2010).

[119] Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. "Frozen Language Model Helps ECG Zero-Shot Learning". In: *MIDL*. 2023.

[120] Paul Pu Liang, Yiwei Lyu, Xiang Fan, et al. "Multibench: Multiscale benchmarks for multimodal representation learning". In: *arXiv* (2021).

[121] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning". In: *NeurIPS*. 2022.

[122] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. "Microsoft coco: Common objects in context". In: *ECCV*. 2014.

[123] Zhisong Liu, Robin Courant, and Vicky Kalogeiton. "FunnyNet: Audio-visual Learning of Funny Moments in Videos". In: *Proc. Asian Conf. on Computer Vision*. 2022.

[124] David Loiseaux. "Multiparameter topological persistence for machine learning". PhD thesis. Université Côte d'Azur, 2024.

[125] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: *ICLR*. 2019.

[126] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *IJCV* (2004).

[127] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. "Are multimodal transformers robust to missing modality?" In: *CVPR*. 2022.

[128] Mengmeng Ma, Jian Ren, Long Zhao, et al. "Smil: Multimodal learning with severely missing modality". In: *AAAI*. 2021.

[129] Mayur Mallya and Ghassan Hamarneh. "Deep multimodal guidance for medical image classification". In: *MICCAI*. Springer. 2022.

[130] Benoit B Mandelbrot and Michael Aizenman. *Fractals: form, chance, and dimension*. 1979.

[131] Nissim Maruani, Peiying Zhang, Siddhartha Chaudhuri, et al. "Illustrator's Depth: Monocular Layer Index Prediction for Image Decomposition". In: *CVPR* (2026).

[132] Warren S McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* (1943).

[133] Leo Milecki, Vicky Kalogeiton, Sylvain Bodard, et al. "Contrastive Masked Transformers for Forecasting Renal Transplant Function". In: *MICCAI*. 2022.

[134] Leo Milecki, Vicky Kalogeiton, Sylvain Bodard, et al. "MEDIMP: 3D Medical Images with clinical Prompts from limited tabular data for renal transplantation". In: *MIDL*. 2023.

[135] Ishan Misra and Laurens van der Maaten. "Self-supervised learning of pretext-invariant representations". In: *CVPR*. 2020.

[136] Julie Mordacq, David Loiseaux, Vicky Kalogeiton, and Steve Oudot. "T-REGS: Minimum Spanning Tree Regularization for Self-Supervised Learning". In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2026. URL: https://openreview.net/forum?id=jvObbvshjE.

[137] Arsha Nagrani, Shan Yang, Anurag Arnab, et al. "Attention bottlenecks for multimodal fusion". In: *NeurIPS* (2021).

[138] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. "Byol for audio: Self-supervised learning for general-purpose audio representation". In: *IJCNN*. 2021.

[139] Mehdi Noroozi and Paolo Favaro. "Unsupervised learning of visual representations by solving jigsaw puzzles". In: *ECCV*. 2016.

[140] Changdae Oh, Junhyuk So, Hoyoon Byun, et al. "Geodesic multi-modal mixup for robust fine-tuning". In: *NeurIPS* (2023).

[141] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv* (2018).

[142] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. "Learning and transferring mid-level image representations using convolutional neural networks". In: *CVPR*. 2014.

[143] Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. "Dinov2: Learning robust visual features without supervision". In: *TMLR* (2023).

[144] Steve Y Oudot. *Persistence theory: from quiver representations to data analysis*. American Mathematical Soc., 2015.

[145] Steve Y Oudot. *Persistence theory: from quiver representations to data analysis*. American Mathematical Soc., 2017.

[146] Sung Y. Park and Anil K. Bera. "Maximum Entropy Autoregressive Conditional Heteroskedasticity Model". In: *Journal of Econometrics* (2009).

[147] Bryan A Plummer, Liwei Wang, Chris M Cervantes, et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models". In: *ICCV*. 2015.

[148] Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. "The Intrinsic Dimension of Images and Its Impact on Learning". In: *ICLR*. 2021.

[149] Robert Clay Prim. "Shortest connection networks and some generalizations". In: *The Bell System Technical Journal* (1957).

[150] The GUDHI Project. *GUDHI User and Reference Manual*. 3.11.0. GUDHI Editorial Board, 2025. URL: https://gudhi.inria.fr/doc/3.11.0/.

[151] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. “Learning transferable visual models from natural language supervision”. In: *Proc. ICML*. 2021.

[152] Marc'Aurelio Ranzato, Fu Jie Huang, Y-Lan Boureau, and Yann LeCun. “Unsupervised learning of invariant feature hierarchies with applications to object recognition”. In: *CVPR*. 2007.

[153] Adrià Recasens, Jason Lin, Joāo Carreira, et al. “Zorro: the masked multimodal transformer”. In: *arXiv* (2023).

[154] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *NeurIPS* (2015).

[155] Frank Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* (1958).

[156] Olivier Roy and Martin Vetterli. “The effective rank: A measure of effective dimensionality”. In: *EUSIPCO*. 2007.

[157] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *Nature* (1986).

[158] Benjamin Matthias Ruppik, Julius von Rohrscheidt, Carel van Niekerk, et al. “Less is More: Local Intrinsic Dimensions of Contextual Language Models”. In: *NeurIPS* (2025).

[159] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, et al. “Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles”. In: *Proc. ICML*. 2023.

[160] Emanuele Sansone, Tim Lebailly, and Tinne Tuytelaars. “Failure-Proof Non-Contrastive Self-Supervised Learning”. In: (2024).

[161] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. “A theoretical analysis of contrastive unsupervised representation learning”. In: *Proc. ICML*. 2019.

[162] Neil Schneiderman, Gail Ironson, and Scott D Siegel. “Stress and health: psychological, behavioral, and biological determinants”. In: *Annual Review of Clinical Psychology* (2005).

[163] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. “CNN features off-the-shelf: an astounding baseline for recognition”. In: *CVPR-W*. 2014.

[164] Anmol Sharma and Ghassan Hamarneh. “Missing MRI pulse sequence synthesis using multi-modal generative adversarial network”. In: *IEEE TMI* (2019).

[165] Zhiqiang Shen, Zechun Liu, Zhuang Liu, et al. “Un-mix: Rethinking image mixtures for unsupervised visual representation learning”. In: *AAAI*. 2022.

[166] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction". In: *ICLR*. 2022.

[167] Patrice Simard, Bernard Victorri, Yann LeCun, and John Denker. "Tangent prop-a formalism for specifying selected invariances in an adaptive network". In: *NeurIPS* (1991).

[168] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, et al. "Dinov3". In: *arXiv* (2025).

[169] Walter Simoncini, Spyros Gidaris, Andrei Bursuc, and Yuki M Asano. "No train, all gain: Self-supervised gradients improve deep frozen representations". In: *NeurIPS* (2024).

[170] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv* (2014).

[171] Umut Simsekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. "Hausdorff dimension, heavy tails, and generalization in neural networks". In: *NeurIPS* (2020).

[172] Oscar Skean, Aayush Dhakal, Nathan Jacobs, and Luis Gonzalo Sanchez Giraldo. "FroSSL: Frobenius Norm Minimization for Efficient Multiview Self-supervised Learning". In: *ECCV*. 2024.

[173] Charles Spearman. "The proof and measurement of association between two things." In: *The American Journal of Psychology* (1961).

[174] Jost Tobias Springenberg. "Unsupervised and semi-supervised learning with categorical generative adversarial networks". In: *arXiv* (2015).

[175] Siddharth Srivastava and Gaurav Sharma. "Omnivec: Learning robust representations with cross modal sharing". In: *CVPR*. 2024.

[176] J Michael Steele. "Growth rates of Euclidean minimal spanning trees with power weighted edges". In: *The Annals of Probability* 16.4 (1988), pp. 1767–1787.

[177] Timothy Stewart. *Overview of Motor Vehicle Traffic Crashes in 2021*. Tech. rep. National Highway Traffic Safety Administration, 2023.

[178] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. "Eva-clip: Improved training techniques for clip at scale". In: *arXiv* (2023).

[179] Christian Szegedy, Wei Liu, Yangqing Jia, et al. "Going deeper with convolutions". In: *CVPR*. 2015.

[180] Pieter Merkus Lambertus Tammes. "On the origin of number and arrangement of the places of exit on the surface of pollen-grains". In: *Recueil des travaux botaniques néerlandais* (1930).

[181] Charlie B Tan, Inés García-Redondo, Qiquan Wang, Michael M Bronstein, and Anthea Monod. "On the limitations of fractal dimension as a measure of generalization". In: *NeurIPS* (2024).

[182] Mingxing Tan and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *Proc. ICML*. 2019.

[183] Piotr Tempczyk, Rafał Michaluk, Lukasz Garncarek, et al. "Lidl: Local intrinsic dimension estimation using approximate likelihood". In: *Proc. ICML*. 2022.

[184] Vimal Thilak, Chen Huang, Omid Saremi, et al. "LiDAR: Sensing Linear Probing Performance in Joint Embedding SSL Architectures". In: *ICLR*. 2024.

[185] Bart Thomee, David A Shamma, Gerald Friedland, et al. "Yfcc100m: The new data in multimedia research". In: *Communications of the ACM* (2016).

[186] Yonglong Tian, Olivier J Henaff, and Aäron Van den Oord. "Divide and contrast: Self-supervised learning from uncurated data". In: *CVPR*. 2021.

[187] Yonglong Tian, Dilip Krishnan, and Phillip Isola. "Contrastive multiview coding". In: *ECCV*. 2020.

[188] Christopher Tralie, Nathaniel Saul, and Rann Bar-On. "Ripser. py: A lean persistent homology library for python". In: *JOSS* (2018).

[189] Michael Tschannen, Alexey Gritsenko, Xiao Wang, et al. "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features". In: *arXiv* (2025).

[190] Eleftherios Tsonis, Xi Wang, and Vicky Kalogeiton. "What Images Cannot Say: Language-Guided Olfactory Representation Learning". In: *ECCV*. 2026.

[191] Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, et al. "Intrinsic dimension estimation for robust detection of ai-generated texts". In: *NeurIPS* (2023).

[192] Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, et al. "The geometry of hidden representations of large transformer models". In: *NeurIPS* (2023).

[193] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *J. Machine Learning Research* (2008).

[194] Grant Van Horn, Elijah Cole, Sara Beery, et al. "Benchmarking representation learning for natural world image collections". In: *CVPR*. 2021.

[195] Grant Van Horn, Oisin Mac Aodha, Yang Song, et al. "The inaturalist species classification and detection dataset". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[196] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. “Attention is all you need”. In: *NeurIPS* (2017).

[197] Shashanka Venkataramanan, Valentinos Pariza, Mohammadreza Salehi, et al. “Franca: Nested matryoshka clustering for scalable visual representation learning”. In: *CVPR* (2026).

[198] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. “Extracting and composing robust features with denoising autoencoders”. In: *Proc. ICML*. 2008.

[199] Karthik Viswanathan, Yuri Gardinazzi, Giada Panerai, Alberto Cazzaniga, and Matteo Biagetti. “The geometry of tokens in internal representations of large language models”. In: *arXiv* (2025).

[200] Hu Wang, Yuanhong Chen, Congbo Ma, et al. “Multi-Modal Learning With Missing Modality via Shared-Specific Feature Modelling”. In: *CVPR*. 2023.

[201] Hu Wang, Congbo Ma, Jianpeng Zhang, et al. “Learnable Cross-modal Knowledge Distillation for Multi-modal Learning with Missing Modality”. In: *MICCAI*. 2023.

[202] Tongzhou Wang and Phillip Isola. “Understanding contrastive representation learning through alignment and uniformity on the hypersphere”. In: *Proc. ICML*. 2020.

[203] Yihe Wang, Yu Han, Haishuai Wang, and Xiang Zhang. “Contrast Everything: A Hierarchical Contrastive Framework for Medical Time-Series”. In: *NeurIPS*. 2023.

[204] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. “DeepFlow: Large displacement optical flow with deep matching”. In: *CVPR*. 2013.

[205] Xi Weng, Lei Huang, Lei Zhao, et al. “An investigation into whitening loss for self-supervised learning”. In: *NeurIPS* (2022).

[206] Xi Weng, Yunhao Ni, Tengwei Song, et al. “Modulate Your Spectrum in Self-Supervised Learning”. In: *ICLR*. 2024.

[207] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. “Unsupervised feature learning via non-parametric instance discrimination”. In: *CVPR*. 2018.

[208] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. “Sun database: Large-scale scene recognition from abbey to zoo”. In: *CVPR*. 2010.

[209] Junyuan Xie, Ross Girshick, and Ali Farhadi. “Unsupervised deep embedding for clustering analysis”. In: *Proc. ICML*. 2016.

[210] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated residual transformations for deep neural networks". In: *CVPR*. 2017.

[211] Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. "Post-pre-training for modality alignment in vision-language foundation models". In: *CVPR*. 2025.

[212] Joseph E Yukich. *Probability theory of classical Euclidean optimization problems*. Springer, 2006.

[213] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. "Barlow twins: Self-supervised learning via redundancy reduction". In: *Proc. ICML*. 2021.

[214] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. "Sigmoid loss for language image pre-training". In: *ICCV*. 2023.

[215] Dingyi Zhang, Yingming Li, and Zhongfei Zhang. "Deep metric learning with spherical embedding". In: *NeurIPS* (2020).

[216] Junming Zhang, Haomeng Zhang, Ram Vasudevan, and Matthew Johnson-Roberson. "Hyperspherical embedding for point cloud completion". In: *CVPR*. 2023.

[217] Richard Zhang, Phillip Isola, and Alexei A Efros. "Colorful image colorization". In: *ECCV*. 2016.

[218] Shaofeng Zhang, Feng Zhu, Junchi Yan, Rui Zhao, and Xiaokang Yang. "Zero-cl: Instance and feature decorrelation for negative-free symmetric contrastive learning". In: *ICLR*. 2021.

[219] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. "Contrastive learning of medical visual representations from paired images and text". In: *Machine Learning for Healthcare Conference*. 2022.

[220] Jinghao Zhou, Chen Wei, Huiyu Wang, et al. "iBOT: Image BERT Pre-Training with Online Tokenizer". In: *ICLR* (2022).

[221] Jinghao Zhou, Chen Wei, Huiyu Wang, et al. "Image BERT Pre-training with Online Tokenizer". In: *ICLR*. 2022.

[222] Bin Zhu, Bin Lin, Munan Ning, et al. "LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment". In: *ICLR*. 2024.

List of Figures

1.1 **Evolution of Visual Representation Learning.** Three paradigms for extracting representations from images: *(a)* handcrafted descriptors encode local gradient statistics around salient keypoints as fixed-length vectors (e.g., SIFT [126]); with *Deep Learning*, representations are learned end-to-end via backpropagation, where a head is trained on top of the encoder: *(b) deep supervised representations*: the head predicts the class label; *(c) self-supervised representations*: a pretext task provides a pseudo-label, enabling training without manual annotations. In both cases the head is discarded and the encoder and its representations are retained. 3

1.2 **Evolution of linear probing accuracy on ImageNet over time.** We compare fully Supervised Learning, Self-Supervised Learning (SSL) and Multimodal SSL. Although SSL emerged later, it improved rapidly and now approaches the saturated top-1 accuracy regime typical of recent Supervised Learning and Multimodal SSL architectures. Figure adapted from Siméoni et al. [168]. 5

1.3 **SSL models trained with different objectives produce geometrically distinct representation spaces.** CIFAR-10 representations from *(a)* CLIP ViT-B and *(b)* DINOv3 ViT-B, projected via t-SNE (left) and PCA (right). The two models structure the same frozen representations very differently: DINOv3 yields tight, well-separated clusters, while CLIP produces a more diffuse, overlapping arrangement. 7

2.1 **Self-supervised pretext tasks: Learning from self-generated signals by predicting withheld or distorted input properties.** Surrogate tasks are constructed by extracting supervisory signals directly from the unlabeled data itself, such as: *(a)* predicting the position of a target patch relative to a reference patch, *(b)* predicting pixel colors from a grayscale image and *(c)* rotating an image and predicting the corresponding angle. 13

2.2 **Semantic invariance under data augmentations.** Various random transforms applied to the original image *(left)* create distorted views that preserve the underlying semantic content [54]. 14

2.3 **Overview of joint-embedding architectures.** Two separate data augmentation operators are sampled from the same distribution ($t, t' \sim \mathcal{T}$) and applied to each image in a batch I to obtain two views, X and X'. An encoder network f and a projection head g are trained to maximize agreement between the resulting embeddings. After training, the encoder f and its representations Y are retained for downstream tasks. 16

2.4 **Overview of Mmultimodal joint-embedding self-supervised learning.** Each modality (e.g., image, text) is processed by a dedicated encoder and projection head. An alignment objective (typically contrastive) pulls together embeddings of paired samples across modalities while pushing apart unpaired ones, yielding a shared latent space in which semantically corresponding inputs from different modalities are co-located. 26

2.5 **Overview of a minimum spanning tree (**MST**).** *(Left)* a point cloud $X \subset \mathbb{R}^2$. *(Right)* The MST(X) connects all points without forming any cycle, while minimizing the total edge length. *Grey edges* indicate pairwise connections not retained in the MST: replacing a MST by one of them would either create a cycle or increase the total length. 29

2.6 **Vietoris-Rips filtration on a point cloud (degree** 0**).** At $t = 0$ every point forms its own connected component, so all 0-dimensional classes are *born* simultaneously at 0. As the filtration parameter t increases, points within distance t are connected by an edge, progressively merging connected components (e.g. at t_1). Each merger is the *death* of a 0-dimensional homology class, recorded as a point $(0, t_{\text{death}})$ in the persistence diagram *(right)*. 32

3.1 **Illustration of collapse modes in JE-SSL versus uniformly sampled points.** *Representational collapse* (*left*) and *dimensional collapse* (*centre*) are the two main collapse modes in joint-embedding self-supervised learning: embeddings either degenerate to a single point or concentrate on a low-dimensional subspace, diminishing the ability to discriminate between inputs. An ideal regime (*right*) corresponds to embeddings that are *uniformly spread* across the representation space, preserving maximal information for downstream tasks. 36

3.2 **Overview of T-REGS.** *(Left)* Two augmented views X, X' are encoded by f and projected by h into embeddings Z, Z'. Training jointly: *(i)* minimizes the Mean Squared Error, $\mathcal{L}_{\text{MSE}}(Z, Z')$, to enforce view invariance (or alternatively the objective function of a given SSL method, $\mathcal{L}_{\text{SSL}}(Z, Z')$, when used as an auxiliary term); *(ii)* maximizes the minimum-spanning-tree length on each branch, $\mathcal{L}_{\text{E}}(Z)$ and $\mathcal{L}_{\text{E}}(Z')$, repelling edge-connected points in MST(Z) and MST(Z'); and *(iii)* applies sphere constraints $\mathcal{L}_{\text{S}}(Z)$ and $\mathcal{L}_{\text{S}}(Z')$. *(Right)* As a result, T-REGS induces uniformly distributed embeddings without dimensional collapse. 37

3.3 **Illustration of T-REG with synthetic data.** *(a-c) 3-d point cloud analysis:* (a) T-REG successfully spreads points uniformly on the sphere by combining MST length maximization and sphere constraint, (b) using only MST length maximization leads to excessive dilation, (c) stable convergence of T-REG whereas $\mathcal{L}_{\text{E}}$ alone fails to converge. *(d-e) Higher-dimensional analysis (256-d):* (d) T-REG enforces effective convergence to the 255-d regular simplex (Theorem 3.2.1), (e) stable optimization behavior of T-REG. . . . 39

3.4 **Sensitivity to dimensional collapse.** Collapse is simulated by zeroing out a fraction η of the coordinates of 10,000 points drawn from an isotropic Gaussian in $\mathbb{R}^{1024}$. Both $-\mathcal{W}_2$ and $-\mathcal{L}_{\text{E}}$ decrease as η increases, showing that $\mathcal{L}_{\text{E}}$ is as sensitive to dimensional collapse as the $\mathcal{W}_2$ loss of Fang et al. [63]. 47

3.5 **Impact of the projector architecture.** $\mathcal{L}_{\text{MSE}}$ +$\mathcal{L}_{\text{T-REGS}}$ top-1 accuracy (%) on the linear evaluation protocol with 100 pretraining epochs. 53

3.6 **Histograms of embeddings' cosine similarities on CIFAR-10.** With T-REGS as a standalone regularization (orange) or as an auxiliary loss (dark orange), the distribution of pairwise cosine similarities becomes concentrated around zero, indicating that the embeddings are highly decorrelated and approach a regular simplex configuration (Theorem 3.2.1). 55

3.7 **Limitations of redundancy-based methods for non-isotropic Gaussian measure.** (a) Initial sampling from a non-isotropic Gaussian distribution. (b) After $\mathcal{L}_{\text{var-cov}}$ optimization: despite achieving a near-identity covariance matrix (center), the point cloud remains concentrated around its mean, with visible artifacts: holes (left). (c) $\mathcal{L}_{\text{T-REG}}$ optimization achieves both uniform distribution on the disk and near-identity correlation matrix. 56

4.1 **illustration of the modality gap.** *(Left)* Pre-trained CLIP exhibits two separate clusters for images and text: *the modality gap*, a form of dimensional collapse where each modality's embeddings are confined to a distinct cone on the hypersphere, limiting cross-modal transferability. *(Right)* Enforcing uniformity spreads embeddings across the hypersphere, closing the modality gap. Figure adapted from Oh et al. [140]. 64

4.2 **Overview of ADAPT.** ADAPT takes up to M modalities, including video, audio, and biosignals, as input to produce a modality-agnostic representation for downstream tasks. It is trained in two steps. *(i) Anchoring.* We align the representations of all modalities via contrastive learning to the one of an *anchor* modality, i.e., the strongest and richest modality; here the video. *(ii) Fusion.* The encoders' features are concatenated and fed into the Masked Multimodal Transformer. When a modality is unavailable, the transformer masks its corresponding feature representations. The final representation (i.e., `[CLS]` token output) is used for downstream tasks. 66

5.1 **Foundation models and `IdEst`**: Intra-dataset correlation. Linear probe accuracy of pretrained SSL models on ImageNet (left), iNat-18 (middle left), iNat-21 (middle right), SUN397 (right) versus `IdEst` on each respective dataset. Each point corresponds to a model checkpoint; point size reflects the number of parameters. We report Kendall's $\tau \in [-1, 1]$ and Spearman's $\rho \in [-1, 1]$. Correlations across all four benchmarks demonstrate `IdEst`'s ability to provide insights into models' representation quality. 78

5.2 **Impact of sampling distribution on estimators.** *(Left)* 200 points sampled evenly along a 1-dimensional helix. *(Right)* Estimated ID as a function of sample size. While $\dim_{\text{MST}}$ converges accurately to the ground truth $d = 1$ as the sample size increases, MLE and TwoNN do not, with TwoNN even diverging to infinity. . 81

5.3 **Foundation models and `IdEst`:** Inter-dataset correlation. Linear probe accuracy of pretrained SSL models on iNat-18 (top right), iNat-21 (top left), CIFAR-10 (bottom right), CIFAR-100 (bottom left) versus `IdEst` computed on ImageNet. Strong correlations demonstrate that `IdEst` computed on a single reference dataset: ImageNet, is indicative of model quality across datasets. 85

5.4 **Foundation models and IdEst:** Alternative evaluation protocol. Accuracy under alternative evaluation settings versus IdEst. Strong correlations demonstrate that IdEst computed on a single reference dataset: ImageNet, is indicative of model quality across evaluation protocols. 86
5.5 **Tracking training dynamics.** *(Top) Offline probing dynamics:* Evolution of ImageNet-1k linear probing top-1 accuracy (left y-axis, in orange) and IdEst (right y-axis, in blue) over self-supervised pretraining epochs for (a) VICReg (ResNet-50), (b) DINO (ViT-S), and (c) I-JEPA (ViT-B). As representations improve and linear probing accuracy increases, IdEst consistently decreases, demonstrating its ability to capture meaningful geometric changes during the evolution of the representation. *(Bottom) Online probing dynamics during self-supervised training:* Evolution of the self-supervised loss, classification loss, ImageNet-1k online classification top-1 accuracy, and IdEst over training epochs for DINO. While early-stage representations are highly constrained, IdEst progressively tracks improvements in downstream accuracy as training proceeds. 88
5.6 **Alternative metrics and foundation models.** Linear probing accuracy on ImageNet versus *(Top)* RankMe and *(Bottom)* uniformity metrics ($\mathcal{L}_u$ and $\mathcal{W}_2$) for a diverse set of pretrained SSL models. 95
5.7 **Effective vs. intrinsic dimensionality.** A helix is embedded in a 3-dimensional ambient space (high effective dimension), yet is intrinsically a 1-dimensional manifold: point A requires three coordinates to specify its position in the ambient space, but a single arc-length coordinate suffices to locate it along the curve. The two notions thus capture complementary aspects of geometry, how spread out a representation is versus how many degrees of freedom it truly contains. 96
5.8 **Tracking training dynamics: RankMe and IdEst.** Evolution of the self-supervised loss (center), ImageNet-1k online classification top-1 accuracy (middle left), RankMe (middle right) and IdEst (right). IdEst decreases while RankMe increases throughout training. 97

6.1 Two equivalent implementations of the MST length computation used throughout this thesis. *Top*: graph-based construction via (`scipy`). *Bottom*: topological construction via 0-dimensional persistent homology (`ripser`); the MST length equals the sum of finite bar lifetimes in the dimension-0 persistence diagram. 103

List of Tables

2.1 **Overview of self-supervised learning methods.** Joint-Embedding SSL (JE-SSL) methods are first grouped by the mechanism that prevents collapse: *contrastive* methods explicitly treat embeddings of different images as negative pairs; *clustering-based* methods balance assignments across prototypes; *self-distillation* methods rely on an asymmetric student–teacher architecture; and *redundancy-reduction* methods enforce decorrelation across embedding dimensions through soft or hard whitening. Finally, approaches *beyond JE-SSL* cover masked-image modeling, multimodal joint-embedding, and large-scale pretraining, which build on or complement these principles. 17

3.1 **Comparison with $\mathcal{W}_2$ regularization [63] on CIFAR-10/100.** The table is inherited from Fang et al. [63], and we follow the same protocol: ResNet-18 models are pre-trained for 500 epochs on CIFAR-10/100, with a batch size of 256, followed by linear probing. We report Top-1 accuracy (%). **Boldface** indicates best performance. 50

3.2 **Linear Evaluation on ImageNet-100/1k.** We report Top-1 accuracy (%). The top-4 methods are **boldfaced**. For ImageNet-100, ResNet-18 are pre-trained for 400 epochs using a batch size of 256; for ImageNet-1k, ResNet-50 is pre-trained for 100 epochs. The table is mostly inherited from Weng et al. [206]. 51

3.3 **Impact of coefficients.** $\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{T-REGS}}$ top-1 accuracy (%) on ImageNet-1k with online evaluation protocol over 50 epochs. **Boldface** indicates best performance. 52

3.4 **Complexity and computational cost.** Comparison between different methods is performed, with training on ImageNet-1k distributed across 4 Tesla H100 GPUs. The wall-clock time (sec/step) is averaged over 500 steps. B, D ranges are reported from Bardes, Ponce, and Lecun [12] and Garrido et al. [67]. 52

3.5 **Impact of batch size.** $\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{T-REGS}}$ top-1 accuracy (%) using linear evaluation after 100 pre-training epochs. 54

3.6 **Impact of normalization.** $\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{T-REGS}}$ top-1 accuracy (%). . 54

3.7 **Training hyperparameters.** 58

4.1 **Cross-Modal Retrieval after finetuning CLIP.** Image-to-text (i → t) and text-to-image (t → i) retrieval results (top 1/5 Recall: R@1, R@5). The table is mostly inherited from Oh et al. [140]. **Boldface** indicates the best performance. 65

4.2 **Comparison with state-of-the-art methods on StressID.** We report ACC and F1 (mean ± std) against unimodal and multimodal baselines from Chaptoukaev et al. [29] and an additional multimodal baseline designed for missing modalities: ShaSpec+ [200]. **Bold**: best; underlined: second best. *From Chaptoukaev et al. [29]. . . . 71

4.3 **ADAPT under missing-modality conditions on the StressID test set.** ✓/× denote modality availability at test time. We report ACC and F1 (mean ± std). 71

4.4 **Ablation of the T-REGS term in Eq. (4.4) on X^*_{test} and X_{test}.** We compare training with $\mathcal{L}_{\text{anchoring}}$ alone (×) versus the full objective that adds $\sum_m \mathcal{L}_{\text{T-REGS}}(Z_m)$ (✓). **Bold** indicates the best result per column. 72

4.5 **Performance of unimodal encoders.** Each encoder is evaluated independently using the same backbone as in the full multimodal pipeline. Results are reported on X^*_{test} and X_{test} as `mean(std)` over ACC and F1. Audio and Video have no X_{test} results due to missing modalities. 73

4.6 **Architectural details of ADAPT.** 74

5.1 **Unsupervised model selection with `IdEst`.** We evaluate `IdEst` for hyperparameter selection against a supervised linear probe on ImageNet-1k, two unsupervised baselines: α-ReQ [4] and RankMe [67], and a weakly-supervised one: LiDAR [184]. For each SSL objective (and architecture), hyperparameters are jointly selected across according to Equation (5.2). 'Fine-grained' denotes the average performance across all furhter datasets (i.e. iNat-21,SUN, Aircraft, CUB, CIFAR-10) excluding ImageNet-1k. **Bold** values indicate the top-performing model selected by the criteria. `IdEst` is competitive with unsupervised and weakly-supervised baselines across SSL methods, including VICReg, despite RankMe directly mirroring its objective. 90

5.2 **Computational cost.** Wall-clock time (min) on ImageNet-1k to evaluate frozen representations using either linear probing (10 epochs, batch size 1024) or IDEST (single feature-extraction pass followed by $\dim_{\mathrm{MST}}$ computation). Rows vary the backbone architecture, primarily changing the output dimension D and the number of parameters. Results are averages over 3 runs. 92
5.3 **Foundation models studied.** Overview of the pretrained models evaluated in this work, spanning diverse SSL and vision-language objectives and architectures (ResNet, ViT). Top-1 accuracy is reported on ImageNet, ImageNet-v2, and additional fine-grained datasets SUN397, iNat-18, iNat-21. Model weights are loaded from the official repositories or `timm`. 99

Appendix A

Résumé étendu

L'apprentissage auto-supervisé (SSL) s'est imposé comme un paradigme incontournable pour apprendre des représentations significatives sans annotations. En vision par ordinateur, ces représentations distillent l'espace des pixels, trop vaste pour les tâches en aval, en descripteurs compacts et transférables. Le SSL contourne le coût et l'ambiguïté des labels en dérivant le signal de supervision des données elles-mêmes, via des tâches prétextes : reconstruction de régions masquées, ou invariance à des transformations dans les architectures à plongement conjoint (JE-SSL), où deux réseaux rapprochent les plongements de vues distinctes d'une même image. Si le succès empirique de ces méthodes est indéniable, comprendre *pourquoi* ces représentations se transfèrent si bien, au-delà des résultats empiriques, reste une question ouverte.

Une piste déterminante réside dans la *géométrie* de l'espace de représentation : la manière dont les représentations se distribuent et s'organisent. Capturer cette géométrie n'est pas uniquement descriptif. La perte d'entraînement mesure l'optimisation de l'objectif prétexte, non la qualité représentationnelle, et les deux peuvent diverger. Certains modes de défaillance, au premier rang desquels l'effondrement dimensionnel, restent invisibles aux métriques aval. Enfin, si des propriétés géométriques corrèlent avec le transfert, elles offrent des proxies sans étiquette pour l'évaluation, préservant la promesse du SSL. Diverses propriétés ont été étudiées, sans qu'un consensus sur la manière de les caractériser n'ait été atteint.

Cette thèse aborde cette lacune à travers l'arbre couvrant de poids minimal (MST), dont la longueur constitue une approche rigoureuse permettant de capturer la *dimension* d'une mesure. Les MSTs sont exploités selon deux directions complémen-

taires : *façonner* et *évaluer* les représentations auto-supervisées.

Notre premier axe de contributions s'attaque à l'effondrement dimensionnel dans les architectures à plongement conjoint, où les représentations dégénèrent vers un sous-espace de faible dimension, via l'imposition de l'uniformité. L'alignement rapproche les paires positives ; l'uniformité répartit les descripteurs sur l'hypersphère unitaire et prévient l'effondrement. Les formulations usuelles, toutefois, ne capturent pas toute l'étendue géométrique des plongements. Nous proposons T-REGS, un régulariseur qui maximise la longueur du MST sur l'hypersphère unitaire. Allonger l'arbre revient à écarter les points reliés par ses arêtes ; une contrainte de sphère prévient la divergence triviale. Nous établissons des garanties reliant cette maximisation à l'entropie et à l'uniformité sur les variétés riemanniennes compactes. Validé sur des jeux de données de référence en SSL, T-REGS atténue l'effondrement et améliore la qualité des représentations, seul ou comme terme auxiliaire. Nous étendons ce principe au cadre multimodal, où l'effondrement se manifeste sous la forme de l'écart intermodal : les plongements de modalités distinctes occupent des régions disjointes plutôt qu'un espace partagé. En application biomédicale, nous introduisons ADAPT, qui détecte les changements physiologiques à partir de capteurs portables, en présence de modalités manquantes. ADAPT s'ancre sur la modalité la plus forte, comble l'écart via T-REGS, et fusionne les descripteurs par un transformeur à attention masquée, robuste à l'entraînement comme à l'inférence.

Notre second axe se tourne vers l'évaluation. Les protocoles usuels, *linear probing* ou *fine-tuning*, sont coûteux, sensibles aux hyperparamètres et exigent des étiquettes, alors que la perte auto-supervisée renseigne peu sur le transfert. Nous proposons `IdEst`, un critère non supervisé et agnostique au paradigme : la dimension intrinsèque des représentations, estimée via l'estimateur basé sur le MST ($\dim_{\mathrm{MST}}$), est un proxy fiable et sans étiquette pour les performances sur les tâches en aval. Plus adapté au régime du SSL que les estimateurs par plus proches voisins, `IdEst` corrèle avec le transfert pour un large éventail d'objectifs et d'architectures, et permet le réglage des hyperparamètres sans annotation.

Ensemble, nos contributions soulignent que la géométrie de l'espace de représentation, telle que capturée par les MSTs, n'est pas uniquement un prisme descriptif mais aussi un prisme actionnable : la géométrie qui signale de bonnes représentations peut à la fois être une cible d'optimisation et un outil pour l'analyse non supervisée de modèles.

ECOLE DOCTORALE

Titre: Capturer la Géométrie des Représentations Auto-Supervisées

Mots clés: Apprentissage profond, Apprentissage auto-supervisé, Vision par ordinateur, Géometrie, Analyse topologique des données

Résumé: L'apprentissage auto-supervisé (SSL) s'est imposé comme un paradigme incontournable pour apprendre des représentations significatives sans annotations. Si son succès est indéniable, comprendre *pourquoi* ces représentations sont efficaces au-delà des résultats empiriques reste une question ouverte. Diverses propriétés géométriques reflétant la qualité des représentations ont été étudiées, sans qu'un consensus sur la manière de les caractériser n'ait été atteint. Cette thèse aborde cette lacune à travers l'arbre couvrant de poids minimal (MST), dont la longueur constitue une approche rigoureuse permettant de capturer la *dimension* d'une mesure. Les MSTs sont exploités selon deux directions complémentaires : *façonner* et *évaluer* les représentations auto-supervisées.

Notre premier axe de contributions s'attaque à l'effondrement dimensionnel dans les architectures auto-supervisées à plongement conjoint, où les représentations dégénèrent vers un sous-espace de faible dimension, via l'imposition de l'uniformité. Premièrement, nous proposons T-REGS, un régulariseur qui maximise la longueur du MST sur l'hypersphère unitaire, favorisant de manière prouvée l'uniformité, validé sur des jeux de données de référence en SSL. Deuxièmement, nous étendons ce principe au cadre multimodal, où l'effondrement se manifeste sous la forme de l'écart intermodal. En guise d'application biomédicale concrète, nous introduisons ADAPT, qui détecte les changements physiologiques à partir de capteurs portables, en présence de modalités manquantes, en s'ancrant sur la modalité la plus forte et en comblant l'écart via T-REGS. Notre second axe de contributions se tourne vers l'évaluation des représentations. Nous proposons IdEst, un critère non supervisé et agnostique au paradigme : la dimension intrinsèque des représentations auto-supervisées, estimée via l'estimateur basé sur le MST, est un proxy fiable et sans étiquette pour les performances sur les tâches en aval, permettant le réglage des hyperparamètres sans annotation.

Ensemble, nos contributions soulignent que la géométrie de l'espace de représentation, telle que capturée par les MSTs, n'est pas uniquement un prisme descriptif mais aussi un prisme actionnable : la géométrie qui signale de bonnes représentations peut à la fois être une cible d'optimisation et un outil pour l'analyse non supervisée de modèles.

Title: Capturing the Geometry of Self-Supervised Representations

Keywords: Deep Learning, Self-supervised learning, Computer Vision, Geometry, Topological Data Analysis

Abstract: Self-supervised learning (SSL) has emerged as a powerful paradigm for learning meaningful representations without human annotations. While its success is undeniable, understanding *why* these representations are effective beyond empirical validation remains an open question. Various geometric properties reflecting the quality of representations have been studied, yet no consensus has emerged on how to characterize them effectively. This thesis addresses this gap through the Minimum Spanning Tree (MST), whose length provides a principled, sample-based way to capture the *dimension* of a measure. MSTs are leveraged along two complementary directions: *shaping* and *evaluating* SSL representations.

Our first line of contributions tackles dimensional collapse in joint-embedding SSL, where representations degenerate onto a low-dimensional subspace, through uniformity enforcement. First, we propose T-REGS, a regularizer that maximizes embeddings' MST length on the unit hypersphere, provably driving uniformity, validated on classical SSL benchmarks. Second, we extend this principle to the multimodal setting, where collapse manifests as the modality gap. As a concrete biomedical application, we introduce ADAPT, which detects physiological changes from wearable sensors, under missing modalities, by anchoring to the strongest modality and closing the gap via T-REGS. Our second line of contributions turns to evaluating representations. We propose IdEst, a paradigm-agnostic, unsupervised criterion: the intrinsic dimension of SSL representations, estimated via the MST-based estimator, is a reliable, label-free proxy for downstream performance, enabling annotation-free hyperparameter tuning.

Together, our contributions highlight that the geometry of the representation space, as captured by MSTs, is not merely a descriptive lens but an actionable one: geometry that signals good representations can be both a target for optimization and a tool for unsupervised model analysis.

Institut Polytechnique de Paris

91120 Palaiseau, France