Extended Abstract Track

Not So Intrinsic: Rethinking Intrinsic Dimension Estimation in Neural Representations

Editors: List of editors' names

Abstract

The analysis of neural representation has become an integral part of research aiming to better understand the inner workings of neural networks. While there are many different approaches to investigate neural representations, an important line of research has focused on doing so through the lens of intrinsic dimensions (IDs). While this perspective has provided valuable insights and stimulated substantial follow-up research, important limitations of this approach have remained largely unaddressed. In this paper, we highlight an important discrepancy between theory and practice of IDs in neural representations, showing that common ID estimators are, in fact, not tracking the true underlying ID of the representation. We contrast this negative result with an investigation of the underlying factors that may drive commonly reported ID-related results on neural representation in the literature.

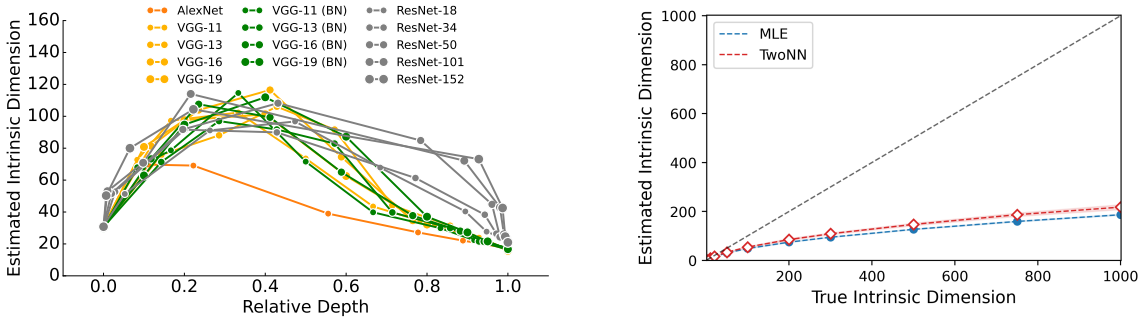
Keywords: Intrinsic Dimensions, Neural Representations

1. Introduction

Intrinsic dimensions (IDs) play a central role in deep learning and have been the focus of research across a broad range of related studies. IDs are often encountered in the context of the so-called *manifold hypothesis* (Tenenbaum et al., 2000; Fefferman et al., 2016). The hypothesis postulates that many high-dimensional datasets frequently encountered in deep learning, such as image and text data, lie on a low-dimensional manifold despite the high-dimensional ambient space d, e.g., the number of pixels of an image. The hypothesis implies that a small number of dimensions $d_{\mathcal{M}} \ll d$ would theoretically suffice to fully characterize such datasets. This manifold dimension $d_{\mathcal{M}}$ is commonly referred to as the *intrinsic dimension*.

The manifold hypothesis has been explored both empirically and theoretically in numerous studies. Although the validity of the hypothesis remains debated, many researchers attribute at least part of the success of deep learning to this phenomenon. In other words, the fact that deep learning models are able to *learn* in the context of high-dimensional image and text data, and thereby escape the so-called *curse of dimensionality* (Bellman, 1961; Bishop and Nasrabadi, 2006), is said to be enabled by the presence of low IDs of the data that neural networks can adapt to (Chen et al., 2019; Schmidt-Hieber, 2019; Nakada and Imaizumi, 2020; Kohler et al., 2023; Schulte et al., 2025).

Related Literature In recent years, there has been a rise in research investigating deep neural networks through the lens of IDs. Besides investigating IDs of frequently encountered datasets in deep learning (Li et al., 2018; Aghajanyan et al., 2020; Pope et al., 2021; Konz and Mazurowski, 2024a), researchers have also aimed to understand the inner workings of these models by examining IDs of neural representations from different layers of the neural network (Gong et al., 2019; Ansuini et al., 2019; Valeriani et al., 2022, 2023; Konz and Mazurowski, 2024b; Doimo et al., 2024; Aljaafari et al., 2025; Cheng et al., 2025).



(a) (b)

Figure 1: a) Layer-wise ID patterns for various architectures (adapted from Ansuini et al., 2019). b) Estimated vs. true ID using TwoNN and MLE. Details in Appendix C.

2. Intrinsic Dimension (Estimators) of Neural Representations

Investigating IDs of neural representations in a broad range of neural architectures, ranging from vision to text-based models, and on various datasets, all previous studies found ID estimates to vary over different network layers. Most strikingly, almost all of these studies find estimated IDs to increase in the early layers and to decrease in later layers (cf. Fig. 1). Such patterns are then often interpreted as the emergence of abstractions or phase transitions (Cheng et al., 2025). In the following, we demonstrate that commonly used ID estimators are not only heavily biased in high dimensions (Section 2.1), but more importantly, are provably not tracking the true underlying IDs of layer-wise neural representations (Section 2.2). This clearly raises the question of what is actually being estimated by the ID estimators, and thus, what drives the commonly found neural ID patterns. This is discussed in Section 2.3.

2.1. ID Estimators in High Dimensions

The ID estimators most commonly applied to neural representations are the MLE (Levina and Bickel, 2004) and TwoNN (Facco et al., 2017) estimators, as well as variants thereof. These and other ID estimators are known to be sensitive to their underlying assumptions and to underestimate the true ID in high dimensions. While this has already been demonstrated for relatively small numbers of dimensions by Levina and Bickel (2004), their bias becomes particularly drastic with growing true ID, as shown in Fig. 1(b). The strong negative bias is particularly concerning given that dimensions like those shown in Fig. 1(b) are highly relevant in modern deep learning. For example, the latest DINO embeddings increased from 1,535 in version 2 to 4,096 in version 3 (Siméoni et al., 2025).

Given this underestimation, related papers usually state that ID estimates should be treated as a *lower bound* of the actual ID (Ansuini et al., 2019). For meaningful insights into ID patterns, the key assumption is that the estimates are at least consistently biased. This would render *relative comparisons* of estimates meaningful by looking at the estimator's *patterns*. However, as the next section will demonstrate, even these ID patterns fail to track a (biased) version of the underlying ID and are therefore not at all indicative of it.

2.2. Estimated ID Patterns Are Not Indicative

In the following, we prove that ID estimators do not estimate the true IDs in neural representations and can also not be considered indicative of those. We demonstrate this result using two popular types of IDs, namely the *Hausdorff dimension* and the *pointwise dimension*. Both are formally defined in Appendices B.1 and B.2. Before stating our results, we briefly introduce some relevant notation.

Notation All spaces considered are Euclidean $(\mathbb{R}^d, \|\cdot\|)$. For $f : \mathbb{R}^n \to \mathbb{R}^m$, we say f is *L-Lipschitz* if $\|f(x) - f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^n$, and *(L, α)-Hölder*, $\alpha \in (0, 1]$, if $\|f(x) - f(y)\| \leq L\|x - y\|^\alpha$. We denote the Hausdorff dimension by $\dim_{\mathrm{H}}$. For a probability measure μ on $\mathbb{R}^n$, the pushforward by f is $\nu = f_{\#}\mu$, i.e. $\nu(\cdot) = \mu(f^{-1}(\cdot))$. For a feedforward network $(f_\ell)_{\ell=1}^L$ and an input law μ_0, define $\mu_\ell := (f_\ell)_{\#}\mu_{\ell-1}$.

Main Result Our main finding rests on the observation that almost all neural network architectures are a composition of layer-wise Lipschitz mappings (cf. Appendix A), and that common notions of IDs cannot increase under Lipschitz mappings. The following Lemma 1 is a classic result from fractal geometry and demonstrates that the Hausdorff dimension cannot grow under Hölder/Lipschitz mappings (see, e.g., Falconer, 2013, Prop. 2.3 & Cor. 2.4 (a)). We provide a proof for completeness in the Appendix.

Lemma 1 (Hausdorff dimension under Hölder/Lipschitz mappings)
Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be (L, α)-Hölder and $E \subset \mathbb{R}^n$. Then

$$\dim_{\mathrm{H}}\big(f(E)\big) \;\leq\; \frac{1}{\alpha}\dim_{\mathrm{H}}(E).$$

In particular, if f is Lipschitz ($\alpha = 1$), then $\dim_{\mathrm{H}}(f(E)) \leq \dim_{\mathrm{H}}(E)$.

A similar result for the so-called *Minkowski dimension* can be found in Hochman (2014, Prop. 2.4). Lemma 1 can be used to show the layer-wise monotonicity of the Hausdorff dimension. Hence, $\dim_{\mathrm{H}}$ cannot increase over the layers of any Lipschitz neural network.

Theorem 2 (Layer-wise monotonicity of Hausdorff dimension)
Let $f_1, \ldots, f_L$ be Lipschitz maps and set $\mu_\ell = (f_\ell)_{\#}\mu_{\ell-1}$. Then, for each $\ell \in \{1, \ldots, L\}$,

$$\dim_{\mathrm{H}}\big(\operatorname{supp}\mu_\ell\big) \;\leq\; \dim_{\mathrm{H}}\big(\operatorname{supp}\mu_{\ell-1}\big).$$

If f_ℓ is only (L_ℓ, α_ℓ)-Hölder, then $\dim_{\mathrm{H}}(\operatorname{supp}\mu_\ell) \leq \alpha_\ell^{-1}\dim_{\mathrm{H}}(\operatorname{supp}\mu_{\ell-1})$.

Proof Consider Lemma 1 with $E = \operatorname{supp}\mu_{\ell-1}$ and $f = f_\ell$, and note $f_\ell(\operatorname{supp}\mu_{\ell-1}) = \operatorname{supp}\mu_\ell$. ∎

In Lemma 4 and Theorem 5 in the Appendix, we show an analogous result for the *pointwise dimension*. The latter is of particular interest given that it forms the basis for and is targeted by the previously introduced MLE and TwoNN estimators (see Appendix B.4 for a corresponding derivation).

A Contradiction Theorem 2 and Theorem 5 imply that the ID cannot increase over the layers of any Lipschitz neural network. This raises serious concerns about increasing ID patterns observed throughout all empirical studies investigating layer-wise ID in neural networks. Given that the actual layer-wise ID cannot increase, the layer-wise ID patterns found by these studies cannot correspond to the true IDs of the neural representations, not even in a relative sense (the bias cannot be consistent). Hence, estimated neural ID-patterns are not only strongly biased but also *not at all indicative* of the underlying IDs of neural representations.

2.3. Connecting Intrinsic Dimensions to Other Metrics

Given the previous conclusion, what are ID estimators then actually measuring? While there are potentially many factors that may impact estimated layer-wise ID patterns, two potential candidates are the sample size and the ambient space dimension d. While larger sample sizes n should make the ID estimates more reliable, larger ambient dimensions may have the opposite effect. However, as can be seen in Fig. 3 and Fig. 5, ID estimates are not directly driven by the ambient space dimension. The same was found by Ansuini et al. (2019) for the effect of ambient space dimension and the sample size. Apart from ID, various other metrics are used to analyze the layer-wise geometry of neural representations. Recently, Skean et al. (2025) studied *matrix-based entropy* and also found distinct layer-wise entropy patterns across different architectures. Inspired by these finding, we study layer-wise von Neumann entropy estimates and indeed find striking similarities in their pattern and ID estimated (cf. Fig. 2) across various CNN architectures.

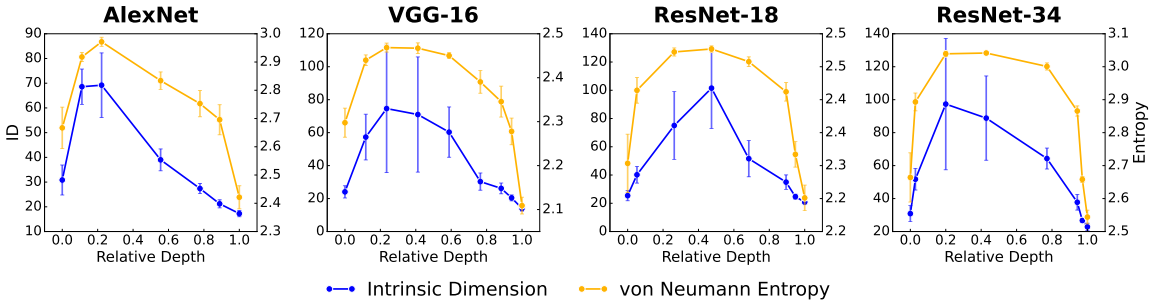


Figure 2: Layer-wise comparison of estimated intrinsic dimensions (left y-axis) vs. von Neumann entropy (right y-axis) of neural representations from different pre-trained models. The x-axis shows relative depth of model layers. Details in Appendix C.

3. Conclusion and Future Outlook

Our investigation showed that commonly found layer-wise ID patterns of neural representations are not indicative of underlying true IDs and should not be interpreted as such. However, a driving factor of these consistently present patterns might be the layer-wise entropy. Based on current evidence, a deeper investigation and formalization of this will be pursued in future work, which we expect to yield promising results.

References

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

Nura Aljaafari, Danilo S Carvalho, and André Freitas. Trace for tracking the emergence of semantic representations in transformers. *arXiv preprint arXiv:2505.17998*, 2025.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Jonathan Bac, Evgeny M Mirkes, Alexander N Gorban, Ivan Tyukin, and Andrei Zinovyev. Scikit-dimension: a python package for intrinsic dimension estimation. *Entropy*, 23(10): 1368, 2021.

Richard E. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, 1961.

Louis Béthune, Thibaut Boissin, Mathieu Serrurier, Franck Mamalet, Corentin Friedrich, and Alberto Gonzalez Sanz. Pay attention to your loss: understanding misconceptions about lipschitz neural networks. *Advances in Neural Information Processing Systems*, 35: 20077–20091, 2022.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient approximation of deep relu networks for functions on low dimensional manifolds. *Advances in Neural Information Processing Systems*, 32, 2019.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Lei Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Diego Doimo, Alessandro Serra, Alessio Ansuini, and Alberto Cazzaniga. The representation landscape of few-shot learning and fine-tuning in large language models. *Advances in Neural Information Processing Systems*, 37:18122–18165, 2024.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1): 12140, 2017.

Kenneth Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2013.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C Principe. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1): 535–548, 2014.

Sixue Gong, Vishnu Naresh Boddeti, and Anil K Jain. On the intrinsic dimensionality of image representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3987–3996, 2019.

Felix Hausdorff. Dimension und äußeres Maß. *Mathematische Annalen*, 79(1):157–179, 1918.

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

Shohei Hidaka and Neeraj Kashyap. On the estimation of pointwise dimension. *arXiv preprint arXiv:1312.2298*, 2013.

Michael Hochman. Lectures on dynamics, fractal geometry, and metric number theory. *J. Mod. Dyn*, 8(3-4):437–497, 2014.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *International Conference on Machine Learning*, pages 5562–5571. PMLR, 2021.

Michael Kohler, Sophie Langer, and Ulrich Reif. Estimation of a regression function on a manifold by fully connected deep neural networks. *Journal of Statistical Planning and Inference*, 222:160–181, 2023. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2022.05.008.

Nicholas Konz and Maciej A Mazurowski. The effect of intrinsic dataset properties on generalization: Unraveling learning differences between natural and medical images. In *International Conference on Learning Representations*, 2024a.

Nicholas Konz and Maciej A Mazurowski. Pre-processing and compression: Understanding hidden representation refinement across imaging domains via intrinsic dimension. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024b.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in Neural Information Processing Systems*, 17, 2004.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21 (174):1–38, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Mathew D. Penrose and J. E. Yukich. Limit theory for point processes in manifolds. *The Annals of Applied Probability*, 23(6):2161–2211, 2013.

Mathew D Penrose and Joseph E Yukich. Weak laws of large numbers in geometric probability. *The Annals of Applied Probability*, 13(1):277–303, 2003.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Johannes Schmidt-Hieber. Deep ReLu network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*, 2019.

Rickmer Schulte, David Rügamer, and Thomas Nagler. Adjustment for confounding using pre-trained representations. In *Forty-second International Conference on Machine Learning*, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*, 2025.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in neural information processing systems*, 31, 2018.

Lucrezia Valeriani, Francesca Cuturello, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of protein language models. *bioRxiv*, pages 2022–10, 2022.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36:51234–51252, 2023.

Lai-Sang Young. Dimension, entropy and lyapunov exponents. *Ergodic theory and dynamical systems*, 2(1):109–124, 1982.

Appendix A. Neural Networks are Lipschitz Mappings

In this section, we discuss in more detail the Lipschitz assumption used in our main results. We consider deterministic neural network layers at inference time and assume all weights and scalars are finite. Then, the following holds:

- Standard linear or convolutional layers are Lipschitz mappings (Kim et al., 2021, Cor. 2.1);
- Pointwise activations such as ReLU, leaky-ReLU, tanh, sigmoid, softplus, GELU / SiLU are Lipschitz (Tsuzuku et al., 2018);
- Pooling operators and residual additions preserve Lipschitzness of the composition (Tsuzuku et al., 2018; Béthune et al., 2022);
- Softmax is Lipschitz on $\mathbb{R}^d$ (Gao and Pavel, 2017, Prop. 4);
- Normalization layers such as LayerNorm, BatchNorm, and RMSNorm are Lipschitz (Tsuzuku et al., 2018).

Given that most neural networks are compositions of Lipschitz mappings (compositions of the above components), and given that compositions of Lipschitz mappings remain Lipschitz, Theorem 2 and Theorem 5 apply for such networks.

Not Globally Lipschitz Mappings Operations that are discontinuous or not globally Lipschitz are, for example, hard quantization or sign/argmax/top-k gating. Another subtle exception is self-attention. While there are Lipschitz variants such as L_2 self-attention, standard (scaled) dot-product self-attention is not globally Lipschitz on unbounded domains (Kim et al., 2021). Clearly, the same holds for multi-head attention given that it is just a linear map of single self-attention outputs. Nevertheless, if the input space is compact (e.g. for bounded inputs), self-attention is Lipschitz on that set (Kim et al., 2021). Hence, in this and the other special cases, one could alternatively state the results of Theorem 2 and Theorem 5 on a compact subset of the data domain on which each layer is Lipschitz. The conclusions then hold relative to that subset.

Appendix B. Omitted Proofs and Derivations

B.1. Hausdorff Dimension

Definition 3 (Hausdorff measure and dimension (Hausdorff, 1918)) *Let $s \geq 0$ and $\delta > 0$ and $E \subset \mathbb{R}^d$. Define*

$$\mathcal{H}^s_\delta(E) \;:=\; \inf\Big\{ \sum_{i=1}^{\infty} (\operatorname{diam} U_i)^s : \; E \subset \bigcup_{i=1}^{\infty} U_i, \; \operatorname{diam} U_i \leq \delta \Big\},$$

where the infimum is considered with respect to all countable δ-covers $\{U_i\}$ of E, and $\operatorname{diam} U := \sup\{\|x - y\| : x, y \in U\}$. Further, the s-dimensional Hausdorff measure is defined by

$$\mathcal{H}^s(E) \;:=\; \lim_{\delta \downarrow 0} \mathcal{H}^s_\delta(E) \;=\; \sup_{\delta > 0} \mathcal{H}^s_\delta(E).$$

The Hausdorff dimension of the set E is then defined by

$$\dim_H(E) \;:=\; \inf\{s : \; \mathcal{H}^s(E) = 0\} \;=\; \sup\{s : \; \mathcal{H}^s(E) = \infty\}.$$

B.1.1. Proof on Lemma 1

Proof Fix $s > \dim_H(E)$. Then by definition $\mathcal{H}^s(E) = 0$, hence for each $\eta > 0$ there exists $\delta > 0$ with $\mathcal{H}^s_\delta(E) < \eta$. This means there is a cover $E \subset \bigcup_i U_i$ with $\operatorname{diam}(U_i) \leq \delta$ and $\sum_i (\operatorname{diam} U_i)^s < \eta$. Set $t > s/\alpha$ and fix an arbitrary $\delta' > 0$. Then choose $\delta \leq (\,\delta'/L\,)^{1/\alpha}$. For the cover above we then have by Hölder continuity, $\operatorname{diam}\big(f(U_i)\big) \;\leq\; L \operatorname{diam}(U_i)^\alpha \;\leq\; \delta'$, so $\{f(U_i)\}_i$ is a δ'–cover of $f(E)$. Hence

$$\mathcal{H}^t_{\delta'}\big(f(E)\big) \;\leq\; \sum_i \big(\operatorname{diam} f(U_i)\big)^t \;\leq\; L^t \sum_i \big(\operatorname{diam} U_i\big)^{\alpha t} \;\leq\; L^t \sum_i \big(\operatorname{diam} U_i\big)^s \;<\; L^t\,\eta,$$

where we used $\alpha t > s$ in the second-to-last inequality. Since $\eta > 0$ was arbitrary, $\mathcal{H}^t_{\delta'}(f(E)) = 0$ for every $\delta' > 0$, and therefore $\mathcal{H}^t\big(f(E)\big) \;=\; \lim_{\delta' \downarrow 0} \mathcal{H}^t_{\delta'}\big(f(E)\big) \;=\; 0$. As this holds for all $t > s/\alpha$, we obtain $\dim_H(f(E)) \leq s/\alpha$. Letting $s \downarrow \dim_{\mathrm{H}}(E)$ concludes the proof. ■

B.2. Pointwise Dimension

We begin by defining the concept of *pointwise dimension* that was first introduced by Young (1982). For a measure μ, its *upper* and *lower pointwise dimension* at point x are

$$\overline{d}_\mu(x) = \limsup_{r \downarrow 0} \frac{\log \mu(B(x,r))}{\log r}, \quad \underline{d}_\mu(x) = \liminf_{r \downarrow 0} \frac{\log \mu(B(x,r))}{\log r},$$

where $B(x,r)$ corresponds to a ball with radius r that is centered around the point x. When the upper and lower limit agree, i.e. $\overline{d}_\mu(x) = \underline{d}_\mu(x)$, it is also called *pointwise* (or *local Hausdorff*) *dimension* and is denoted by $d_\mu(x)$. Note that the pointwise dimension is defined for a single point instead of the entire dataset, which is why it is sometimes considered a local instead of a global dimension. Nonetheless, it can be considered at multiple points x. In particular, $d_\mu(x)$ is said to be *exact dimensional* in case it exists and is μ-a.s. independent of the point x (hence, $d_\mu(x)$ equals the same constant for all x μ-a.s.). In this case, it is sometimes denoted by d_μ (Hochman, 2014, Def. 3.9).

Lemma 4 (Pointwise dimensions under Hölder/Lipschitz mappings) *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be (L, α)-Hölder, μ a Borel probability measure on $\mathbb{R}^n$, and $\nu = f_{\#}\mu$. For μ-a.e. x with $y = f(x)$,*

$$\overline{d}_\nu(y) \;\le\; \frac{1}{\alpha}\,\overline{d}_\mu(x), \qquad \underline{d}_\nu(y) \;\le\; \frac{1}{\alpha}\,\underline{d}_\mu(x).$$

In particular, if f is Lipschitz ($\alpha = 1$), then $\overline{d}_\nu(y) \le \overline{d}_\mu(x)$ and $\underline{d}_\nu(y) \le \underline{d}_\mu(x)$.

Proof For small $\rho > 0$, (L, α)-Hölder gives $f(B(x, \rho)) \subseteq B(y, L\rho^\alpha)$. Setting $r = L\rho^\alpha$,

$$\nu(B(y, r)) = \mu\big(f^{-1}(B(y, r))\big) \;\ge\; \mu\big(B(x, \rho)\big).$$

Taking logs and dividing by $\log r < 0$ (fulfilled for a sufficiently small r) reverses the inequality:

$$\frac{\log \nu(B(y, r))}{\log r} \;\le\; \frac{\log \mu(B(x, \rho))}{\log r} = \frac{\log \mu(B(x, \rho))}{\alpha \log \rho + \log L}.$$

As $r \downarrow 0$, we have $\rho \downarrow 0$ and $\log \rho \to -\infty$, so the additive constant $\log L$ is negligible in lim sup/lim inf. Hence, for the upper pointwise dimension, we get

$$\overline{d}_\nu(y) = \limsup_{r \downarrow 0} \frac{\log \nu(B(y, r))}{\log r} \;\le\; \limsup_{\rho \downarrow 0} \frac{\log \mu(B(x, \rho))}{\alpha \log \rho + \log L} = \frac{1}{\alpha} \limsup_{\rho \downarrow 0} \frac{\log \mu(B(x, \rho))}{\log \rho} = \frac{1}{\alpha}\,\overline{d}_\mu(x).$$

The same logic applies to the lower pointwise dimension, which concludes the proof. ■

A related result to Lemma 4, studying the dimension under linear maps, can be found in Hochman (2014, Lemma 4.5). Lemma 4 is more general in the sense that any linear map is Lipschitz, but not vice versa.

Theorem 5 (Layer-wise monotonicity of pointwise dimensions) *Consider $\ell \in \{1, \dots, L\}$ and $\mu_\ell = (f_\ell)_{\#}\mu_{\ell-1}$ as in Theorem 2. Then for each ℓ and for $\mu_{\ell-1}$-a.e. x with $y = f_\ell(x)$,*

$$\overline{d}_{\mu_\ell}(y) \;\le\; \overline{d}_{\mu_{\ell-1}}(x), \qquad \underline{d}_{\mu_\ell}(y) \;\le\; \underline{d}_{\mu_{\ell-1}}(x).$$

Consequently, if every μ_ℓ is exact dimensional (i.e., $d_{\mu_\ell}(x)$ exists and equals a constant d_ℓ for μ_ℓ-a.e. all x), then $d_\ell \le d_{\ell-1}$ for all ℓ.

Proof Apply Lemma 4 to each layer. In case of exact dimensionality, the $\mu_{\ell-1}$-a.e. restriction can be dropped, making the a.e. inequalities constant and giving $d_\ell \le d_{\ell-1}$. ■

B.3. Invariance of Hausdorff and Pointwise Dimensions Under Bi-Lipschitz Mappings

The first part of the following result is another classic result from fractal geometry that can be found in Falconer (2013, Cor. 2.4 (b)). The second part that is about the (upper/lower) pointwise dimension is discussed in Hidaka and Kashyap (2013).

Proposition 6 (Invariance under Bi-Lipschitz mappings) *If f is bi-Lipschitz on $E \subset \mathbb{R}^n$, then $\dim_{\mathrm{H}}(f(E)) = \dim_{\mathrm{H}}(E)$. If moreover $\nu = f_{\#}\mu$ and f is bi-Lipschitz on $\operatorname{supp}\mu$, then $\overline{d}_\nu(f(x)) = \overline{d}_\mu(x)$ and $\underline{d}_\nu(f(x)) = \underline{d}_\mu(x)$ for μ-a.e. x.*

Proof Applying Lemma 1 to the Lipschitz function $f : E \to \mathbb{R}^m$ yields $\dim_{\mathrm{H}}(f(E)) \leq \dim_{\mathrm{H}}(E)$. Due to the bi-Lipschitzness, $f^{-1} : f(E) \to E$ is also Lipschitz. Hence, by Lemma 1 we get that $\dim_{\mathrm{H}}(E) \leq \dim_{\mathrm{H}}(f(E))$. This proofs $\dim_{\mathrm{H}}(E) = \dim_{\mathrm{H}}(f(E))$ for bi-Lipschitz f. For the (upper/lower) pointwise dimensions, apply Lemma 4 to both f and f^{-1} using the same logic from above. ■

B.4. MLE and TwoNN target the pointwise dimension

Let M be a d-dimensional manifold and let $Y_1, \ldots, Y_n \in M$ be i.i.d. with probability measure μ. For simplicity, we use a slightly different notation in the following derivation (compared to other sections), with d and D ($d \ll D$) denoting the dimension of the manifold and the ambient space, respectively. The observed sample $\{X_j\}_{j=1}^n$ is its (smooth) embedding $X_j := g(Y_j) \in \mathbb{R}^D$, with a continuous and sufficiently smooth mapping g as in Levina and Bickel (2004).

Local model (homogeneous PPP) Fix a point $x \in M$ for which the pointwise dimension exists, $d_\mu(x) = d$. Assume that in a neighbourhood of x, μ has a density κ with respect to the Riemannian volume on M, with κ continuous at x and $0 < \kappa(x) < \infty$. Under these conditions, the standard Binomial-to-Poisson coupling (Penrose and Yukich, 2003, 2013) implies that, at sufficiently small scales around x, the sample $\{X_j\}$ can be approximated by a *homogeneous Poisson point process* (PPP) in the tangent space $\mathbb{R}^d$ with *intensity* $\lambda_n = n\,\kappa(x)$, meaning the expected number of points in a set equals λ_n times its d-dimensional volume. In particular, for small $r > 0$, the count $N(r, x) := \sum_{j=1}^n \mathbf{1}\{X_j \in B(x, r)\}$ is well-approximated by $N(r, x) \sim \mathrm{Poisson}\big(\lambda_n\, \omega_d\, r^d\big)$, where ω_d is the volume of the unit ball in $\mathbb{R}^d$.

Levina-Bickel MLE Let $T_i(x)$ be the distance from x to its i-th nearest neighbor. Under the local PPP model, conditional on the distance $T_k(x)$ to the k-th neighbor, the ratios $U_i = T_i(x)/T_k(x)$ for $i = 1, \ldots, k-1$ are the order statistics of $k-1$ i.i.d. random variables drawn from a distribution with CDF $F(u) = u^d$ and corresponding PDF $f(u|d) = du^{d-1}$ for $u \in [0, 1]$. The joint log-likelihood of these order statistics is $\ell(d) = C + \sum_{i=1}^{k-1} \log(du_i{}^{d-1})$, where C is a constant independent of d. Maximizing $\ell(d)$ w.r.t. d, yields the MLE from Levina and Bickel (2004):

$$\widehat{d}_{\mathrm{MLE}}(x) \;=\; \left[\frac{1}{k-1}\sum_{i=1}^{k-1} \log \frac{T_k(x)}{T_i(x)}\right]^{-1}.$$

As $n \to \infty$, $k \to \infty$ and $k/n \to 0$ so that $r = T_k(x) \downarrow 0$, the PPP small-scale approximation becomes exact in the limit and $\widehat{d}_{\mathrm{MLE}}(x) \to d_\mu(x)$ in probability.

TwoNN In the special case $k = 2$, the ratio $\rho(x) := T_2(x)/T_1(x) \in [1, \infty)$ satisfies $\log \rho(x) \sim \mathrm{Exp}(d)$ and hence ρ is Pareto(d):

$$f(\rho \,|\, d) = d\,\rho^{-(d+1)}, \qquad F(\rho \,|\, d) = 1 - \rho^{-d} \quad (\rho \geq 1),$$

see Facco et al. (2017, Eqs. (5) & (6)). Treating $\{\rho(x_j)\}$ as approximately independent gives the pseudo-log-likelihood $\ell(d) = n \log d - d \sum_{j=1}^{n} \log \rho(x_j)$, whose maximizer is

$$\widehat{d}_{\text{TwoNN}} = \left[\frac{1}{n} \sum_{j=1}^{n} \log \frac{T_2(x_j)}{T_1(x_j)} \right]^{-1}.$$

Alternatively, the estimation can be based on linear regression (both approaches are asymptotically equivalent). The regression form used in TwoNN follows from the Pareto distribution-based identity

$$-\log\big(1 - F(\rho \,|\, d)\big) = d \log \rho,$$

see Facco et al. (2017, Eqs. (7)). So fitting a straight line (passing through the origin) to $\{\big(\log \rho_j,\ -\log(1 - \widehat{F}_n(\rho_j))\big)\}_{j=1}^{n}$ estimates the slope d (Facco et al., 2017). Under exact dimensionality we have that $d_\mu(x) = d$ for μ-a.e. interior x. In that case, as $n \to \infty$ and $r = T_2(x) \downarrow 0$, the PPP small-scale approximation becomes exact and $\widehat{d}_{\text{TwoNN}} \to d$ in probability.

Appendix C. Additional Experiments and Experimental Details

C.1. Layer-wise Analysis

For the layer-wise neural representation analysis, we follow along the investigation of Ansuini et al. (2019). We consider several classic model architectures such as *Alexnet* (Krizhevsky et al., 2012), *VGG* (Simonyan and Zisserman, 2014), *ResNets* (He et al., 2016) with pretrained weights (pretrained on ImageNet (Deng et al., 2009)) obtained from the PyTorch library *torchvision* (Paszke et al., 2019). The addition of (BN) for the VGGs of Fig. 1 a) indicates that models incorporated Batch Normalization layers.

The layer-wise neural representations are obtained by passing a set of images through each pretrained model and extracting the corresponding representations from the layers. Each dataset consists of 500 samples from the seven largest categories of ImageNet. Further details can be found in Ansuini et al. (2019). In Fig. 1a), Fig. 2, Fig. 3, the point estimates correspond to the mean of the respective estimates on the seven datasets, and the error bars to the corresponding standard deviations. The estimated intrinsic dimensions are obtained via the TwoNN ID estimator (Facco et al., 2017). In each of these plots, the x-axis denotes the relative depth of each model layer to foster visual comparison between models with varying numbers of layers.

Interestingly, although our investigation in Section 2.3 focused on classic vision-based architectures, the striking connection between the estimated layer-wise ID and entropy patterns can also be found in LLMs, when comparing the ID patterns in Fig. 1 of Valeriani et al. (2022) or Fig. 1 of Cheng et al. (2025) to the entropy patterns (called dataset entropy) in Fig. 11 of Skean et al. (2025).

Von Neumann Entropy In Section 2.3 we considered *von Neumann entropy*. This is a special type of so-called *matrix-based entropy* (Giraldo et al., 2014). For this, consider the matrix of layer-wise representation $Z \in \mathbb{R}^{n \times d}$ (hence n observations of d-dimensional neural

representations) and define the corresponding Gram matrix $Q = ZZ^\top \in \mathbb{R}^{n\times n}$. Then the matrix-based entropy of the matrix Z, as defined in Skean et al. (2025), corresponds to:

$$S_\alpha(Z) = \frac{1}{1-\alpha} \log\left(\sum_{i=1}^{r} \left(\frac{\lambda_i(Q)}{\mathrm{tr}(Q)}\right)^\alpha\right), \tag{1}$$

where α can be any $\alpha > 0$, $\lambda_i(Q)$ corresponds to the i-th (nonnegative) eigenvalues of Q, and $r = \mathrm{rank}(Q) \leq \min(n, d)$. The von Neumann entropy can be obtained by letting $\alpha \to 1$. Similarly to Skean et al. (2025), we therefore chose $\alpha = 1$ to obtain our von Neumann entropy estimates in Fig. 2. Point estimates correspond to the mean of the entropy estimates obtained from the seven above mentioned datasets, and the error bars to the corresponding standard deviations.

ID vs. Ambient Space Dimension In Section 2.3, we explored what might be underlying factors that drive the commonly found layer-wise ID patterns. Along with this analysis, we also conducted an experiment comparing the layer-wise ID patterns with the layer-wise embedding dimension in different pretrained models. The result in Fig. 3 shows that the two patterns are relatively different, indicating that ambient space dimension itself does not seem to be the core driving factor for the commonly found ID patterns.

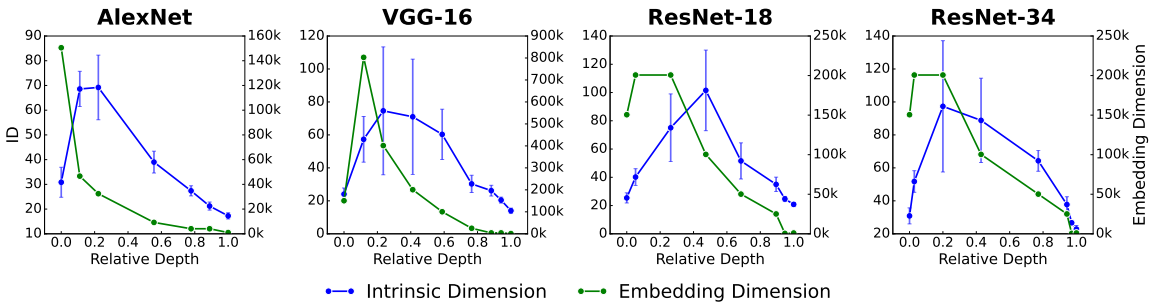


Figure 3: Layer-wise comparison of estimated intrinsic dimensions (left y-axis) vs. embedding dimensions (right y-axis) of neural representations from different pre-trained CNN architectures. The x-axis shows the relative depth of the model layers. The embedding dimension corresponds to the ambient space dimension of each layer.

C.2. Intrinsic Dimension Estimator Analysis

In Fig. 1 b), we have investigated the accuracy of the two ID estimators, Two Nearest Neighbors (TwoNN) Facco et al. (2017) and the Maximum Likelihood Estimator (MLE) Levina and Bickel (2004). In order to obtain datasets with known intrinsic dimension, we sampled datasets with 5k data points uniformly distributed on a $d_\mathcal{M}$-hyperball with varying true ID $d_\mathcal{M}$. For each true intrinsic dimension, we sampled 20 datasets and estimated the IDs via MLE and TwoNN on each of these.

Fig. 1 b) and Fig. 4 depict both to the average over these 20 ID estimates (lines) and the related 95% CI. For the sampling and subsequent ID estimation, we used the *scikit-dimension* library (Bac et al., 2021). Fig. 1 b) shows a strong negative bias for the two

estimators that is growing with increasing true intrinsic dimension. While Fig. 1 b) uses MLE with $k = 20$, Fig. 4 shows that the negative bias is persistent also for other choices of nearest neighbors k.

Similar to Fig. 3, we also demonstrate in a controlled study setup (with known ID and datasets sampled as described above) that the ambient space dimension does not seem to have strong effects on the TwoNN and MLE estimates. For a fixed true ID of 50, the two estimators still exhibit a negative bias, but are largely invariant to changes in the size of the ambient space dimension, as can be seen in Fig. 5.

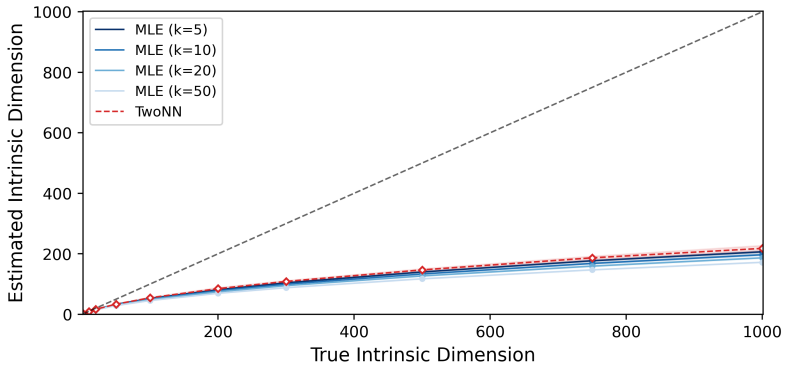


Figure 4: Estimated vs. True ID: Estimated IDs using TwoNN and MLE (different k) of datasets with varying true ID. Each dataset consists of 5k data points uniformly distributed on a $d_{\mathcal{M}}$-hyperball with varying true ID $d_{\mathcal{M}}$. 95% CI are computed based on 20 ID estimates. Both estimators exhibit strong negative bias with increasing $d_{\mathcal{M}}$.

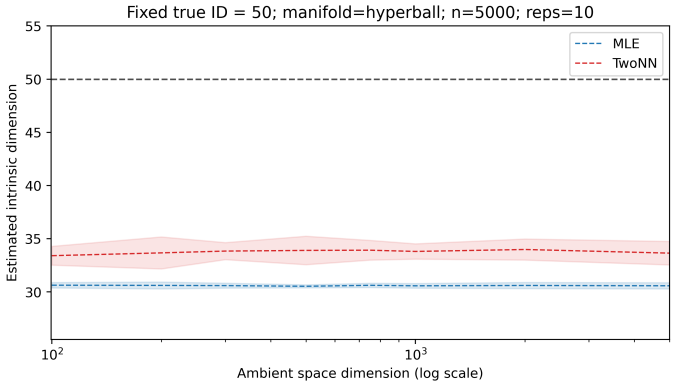


Figure 5: Estimated ID vs. Ambient Space Dim.: Estimated IDs using TwoNN and MLE ($k = 20$) of datasets with varying ambient space dimension d. Each dataset consists of 5k data points uniformly distributed on a $d_{\mathcal{M}}$-hyperball with fixed to $d_{\mathcal{M}} = 50$. 95% CI are computed based on 10 ID estimates. Ambient space dimension does not strongly impact ID estimates.