Exploring Model Transferability through the Lens of Potential Energy

Xiaotong Li[1,2] Zixuan Hu[1] Yixiao Ge[3] Ying Shan[3] Ling-Yu Duan[1,2*]
[1] School of Computer Science, Peking University, Beijing, China,
[2] Peng Cheng Laboratory, Shenzhen, China, [3] ARC Lab, Tencent PCG, Beijing, China
lixiaotong@stu.pku.edu.cn, {yixiaoge, yingsshan}@tencent.com,
{hzxuan, lingyu}@pku.edu.cn

Abstract

Transfer learning has become crucial in computer vision tasks due to the vast availability of pre-trained deep learning models. However, selecting the optimal pre-trained model from a diverse pool for a specific downstream task remains a challenge. Existing methods for measuring the transferability of pre-trained models rely on statistical correlations between encoded static features and task labels, but they overlook the impact of underlying representation dynamics during fine-tuning, leading to unreliable results, especially for self-supervised models. In this paper, we present an insightful physics-inspired approach named PED to address these challenges. We reframe the challenge of model selection through the lens of potential energy and directly model the interaction forces that influence fine-tuning dynamics. By capturing the motion of dynamic representations to decline the potential energy within a force-driven physical model, we can acquire an enhanced and more stable observation for estimating transferability. The experimental results on 10 downstream tasks and 12 self-supervised models demonstrate that our approach can seamlessly integrate into existing ranking techniques and enhance their performances, revealing its effectiveness for the model selection task and its potential for understanding the mechanism in transfer learning. Code is available at https://github.com/lixiaotong97/PED.

1. Introduction

Transfer learning has achieved remarkable success in computer vision by fine-tuning models pre-trained on large-scale datasets (*e.g.,* ImageNet [16]) for downstream tasks. However, the proliferation of various network designs and training strategies presents a challenge in selecting an optimal model from the extensive range of options for a particular downstream task. While fine-tuning each potential model in a brute-force manner is a direct approach for model selection, it is computationally infeasible due to the growing number of model candidates.

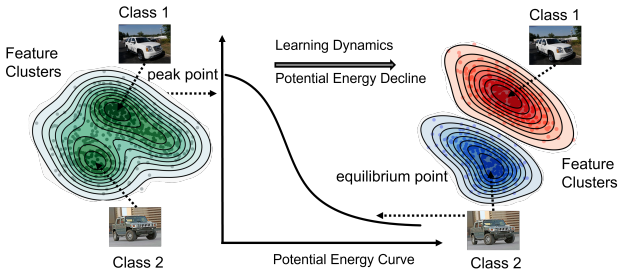

Figure 1. We analogy the physical concept and consider the transfer learning dynamics in the perspective of potential energy. The objective to push apart different classes can be viewed as an interaction "force" to decline the system "potential energy" and the dynamics can be seen as a process from unstable to stable point of the energy plane.

To address this challenge, prior studies [31, 46, 40, 36] have endeavored to efficiently measure the transferability of pre-trained models related to the separability of encoded representations. The principle underlying these approaches is to select a pre-trained model that can effectively segregate its initial features using the ground-truth labels (*i.e.*, image classes) in the downstream task.

While the aforementioned methodology is effective for ranking supervised pre-trained models, which are originally optimized toward class separability, it is not always reliable for ranking un/self-supervised pre-trained models [20]. These models have emerged as dominant in transfer learning and have exhibited superior performance compared to supervised learning models. Nevertheless, self-supervised models exhibit different properties due to the discrepancy between pre-training target and downstream classification objective [20]. We argue that the limitations of the existing separability-based methodology stem from its inability to consider the underlying representation dynamics during

*Corresponding Author.

the fine-tuning process of transfer learning and encounter challenges for ranking self-supervised models.

Modeling the representation dynamics for model ranking is a crucial yet challenging task. The present study focuses on image classification tasks without loss of generality. To understand the nature of model evolution in transfer learning, we examine the process of backward-propagating gradients measured by classification cross-entropy loss. The process aims to cluster features out of the same classes, which can be viewed as creating a force that separates the clusters and system potential energy gets decreased driven by the force from a physical perspective [18, 48]. Reframing model evolution through the lens of potential energy reveals that the pre-trained model attains a state of equilibrium after pre-training, with low interaction forces and stable sample relationships. However, this stable state is disrupted when the model is transferred to a downstream task, leading to changes in the potential energy plane. Intuitively, predicting model transferability based on an unstable observation will hinder its predicting performance. Drawing from the principles of physics [10], the present unstable state is inclined to move towards a reduction in potential energy and results in a more stable state. To properly predict a model's transferability, it is essential to model the force that determines the system's tendency.

We therefore formulate the representation dynamics in terms of potential energy[1] and propose the approach to tackle these challenges named Potential Energy Decline (PED), as demonstrated in Fig.1. To quantify the interaction force acting on each class cluster and measure its corresponding movement on the potential energy landscape implicitly defined by the optimization objective, we consider each class's representations in the downstream task as a ball in the latent space, with the class center indicating the coordinate and the variation representing the radius. The interaction force between different classes is formulated by the overlap radius of the two balls. We can simulate the positions of dynamic representations without backward-propagation by unfreezing the system and observing the moving tendency that leads to a new state with lower potential energy. Our force-directed dynamic representations provide a better observation and can be readily integrated into existing ranking algorithms, such as LogME [46], to achieve better model transferability measurement.

To the best of our knowledge, we are the first to explore model transferabilty through the lens of potential energy and simulate the underlying representation dynamics during transfer learning in a physics-driven approach. To evaluate our proposed method, we conduct extensive experiments on a variety of self-supervised pre-trained models. Our method can be easily integrated into existing approaches with negligible time consumption. The experimental results on 10 downstream tasks and 12 self-supervised pre-trained models demonstrate our method can boost various metrics for more accurate prediction. Our findings might have implications beyond the realm of image classification, as our approach is generic and can be extended to more pre-trained models and other downstream tasks. We hope that our work will inspire future studies and have a broader impact in the field of transfer learning.

[1]Throughout this paper, the "energy" is a quantitative property held by features because of relative class positions in its latent space.

2. Related work

2.1. Transferability Metric

In the field of computer vision, transfer learning has become a significant milestone due to the availability of a model zoo of pre-trained deep learning models [20, 44]. As a result, selecting the most appropriate model from the model zoo for a particular downstream task has become an important challenge and model selection is therefore proposed to tackle this problem with a low budget estimation.

Transferabilty Metric. Model transferability is a fundamental aspect in the field of transfer learning, and it has received much attention from researchers in recent years for designing various transferability metrics [34, 42, 46, 31, 40, 5, 36, 6, 19, 1, 47, 17, 3, 15]. For example, LEEP [34] estimates the joint probability of the source and target label space, while NLEEP [31] predicts the label by fitting a Mixture of Gaussian model. LogMe [46] propose to estimate the maximum value of label evidence given the encoded features. PARC [6] uses pairwise pearson product-moment correlation between the features of each pair of images. GBC [36] measures the pairwise class overlaps in distribution density with a Bhattacharyya coefficient. SFDA [40] measures model transferability using the class discrimination in a Fisher space and proposes a self-challenging approach named ConfMix to simulate the hard negative samples in fine-tuning. These methods have made significant contributions to the field of transfer learning, but there is still challenge to deal with un/self-supervised pre-trained models [22, 9, 11, 12, 30, 8], because the models be useful as a starting point for many downstream tasks, they are not sufficient on their own to separate different classes of samples [20].

2.2. Energy-based Methods in Deep Learning

Energy-based methods have a long-standing history in the field of machine learning and have been commonly employed to model interactions between objects [2, 39]. Early works in this area can be traced back to Restricted Boltzmann Machines (RBM) [2] and DeepBM [39], that use a series of layers of stochastic binary units to represent data and is trained to minimize an energy function measured by

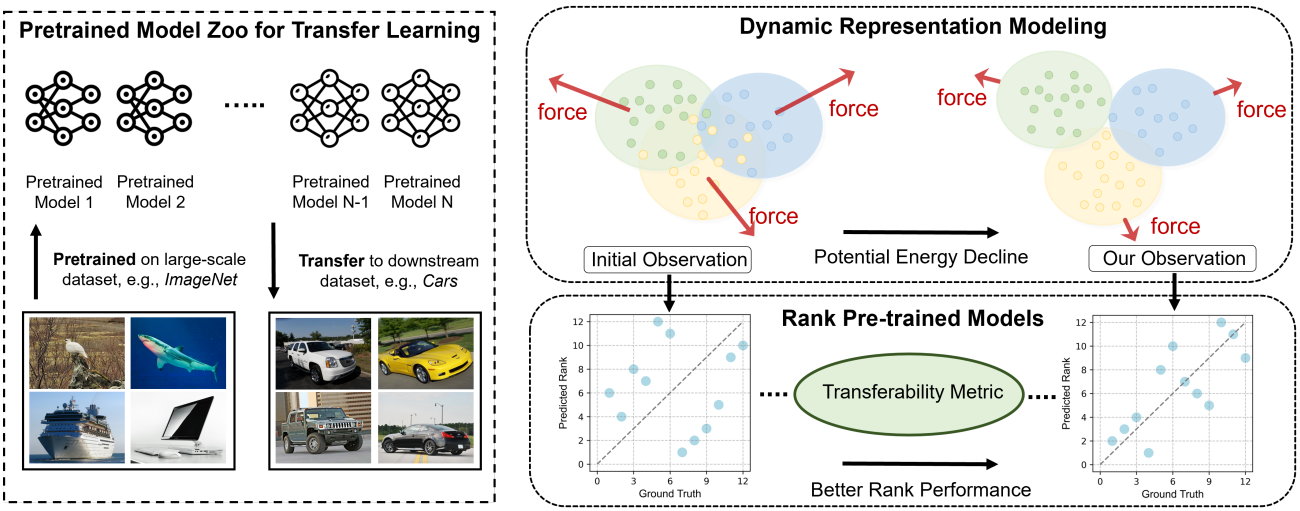

Figure 2. Pipeline of the proposed Potential Energy Decline (PED) approach for model selection. The models are first trained on a large-scale dataset and then transferred to a given downstream task. We propose a novel approach through the perspective of potential energy to alleviate the limitations of the initial observations. By treating the learning dynamics as minimizing the potential energy and considering the system's tendency to change, we model the interaction force of different clusters using a repulsion-based force to capture the moving tendency. Subsequently, we unfreeze the state of the start point and apply the force to push each class away, leading to a decline of the potential energy. This approach leads to a more stable observation of features, resulting in more accurate transferability score predictions.

the compatibility between the input data and the internal representation. The Hopfield network [27] is also designed to find a state of minimum energy, which corresponds to a stable solution or equilibrium and is used for segmentation [38]. In face recognition task, Uniformface[18] and Regularface [48] also borrow the force in potential energy to model the inter-class regularization to design a optimized loss. Inspired by these works, we reframe the challenge of ranking self-supervised pre-trained model through the lens of potential energy. To the best of our knowledge, we are the first to consider transfer learning from an energy-based perspective and propose a physical approach to model the dynamic representation in model selection task.

3. Methodology

In this section, we first present the problem setup, ranking metrics, and evaluation protocol of the model selection problem. Then we state the inspiration from a physical view and illustrate how we efficiently model the representation dynamics in terms of potential energy. Without loss of generality, we take classification as an example throughout our paper.

3.1. Preliminaries

Problem Setup. Consider a model zoo $\{\Phi_i\}_{i=1}^N$ from which the selected pre-trained model can be transferred to a downstream dataset $\mathcal{T} = \{X, Y\}$. The purpose of model selection is to predict model transferability with minor computational costs without fine-tuning.

Ranking Metric. Given a model Φ_i, we encode the features Z for the downstream dataset X, then feed the features and labels to a metric $\mathcal{M}(Z, Y)$ to estimate a transferability score P_i. Intuitively, the metric measures the transferability based on the separability of the encoded features.

Evaluation Protocol. To evaluate different model selection algorithms, we follow previous arts to estimate the weighted Kendalls' τ_w [46] between ground-truth model rankings and the predicted rankings. Specifically, we obtain the ground-truth rankings $\{G_i\}_{i=1}^N$ via fully fine-tuning. Then τ_w can be formulated as

$$\tau_w = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \text{sign}(G_i - G_j) \cdot \text{sign}(P_i - P_j). \tag{1}$$

Although existing ranking methods [46, 31, 40, 36, 6] are effective for supervised pre-trained models, they are not always reliable for un/self-supervised pre-trained models which are not trained toward class separability and need to be fine-tuned for downstream tasks. We argue that it is actually due to the fact that they mostly ignore the underlying representation dynamics during the fine-tuning process of transfer learning. Properly and efficiently modeling such dynamics is called for.

3.2. Understanding Transfer Learning from a Dual View

It is of great significance to rethink the fine-tuning optimization process and dive into the dynamics of learning representations. In this section, we provide a novel dual view to reframe gradient-based optimization into a physics perspective.

3.2.1 A Gradient-based View

When transferring a pre-trained model Φ parameterized by θ to a particular downstream task $\{X, Y\}$, the model is generally optimized toward minimizing a loss function $\mathcal{L}$ (*e.g.*, cross-entropy loss) through gradient backpropagation, such as

$$\begin{aligned} \theta^{t+1} &= \theta^t - \frac{\partial \mathcal{L}(Z^t, Y)}{\partial \theta^t}, \\ Z^t &= \Phi(X|\theta^t). \end{aligned} \tag{2}$$

As a result of the iterative optimization in Eq. (2), the ability of Φ to discriminate between different classes has been improved. The separability of different classes in the latent space has also been enhanced. The underlying representation dynamics can be thought of as the state evolution from Z^0 to Z^T, where T is the total number of iterations. Obviously, optimizing the network and updating its encoded features is non-trivial for the model selection process, and what we need is to simulate the dynamics representation without resorting to fine-tuning.

3.2.2 An Energy-based View

Every coin has two sides, we discover that the learning objective in optimization shows resemblances to the concept of potential energy in physics. Intuitively, the loss function $\mathcal{L}$ and the gradient $\frac{\partial \mathcal{L}}{\partial \theta}$ show similarities in form with potential energy U and force F, *i.e.*, $-\frac{\partial U}{\partial s}$, respectively. The loss gradient minimizes the loss and distinguishes between different class features by adjusting the network parameters, in a similar way to how an object's position affects the force acting on it to decrease potential energy. Building upon this insight, we reformulate the optimization process during transfer learning from the lens of potential energy in physics.

From the physics perspective, the direction of object movement under the influence of an interaction force F can be denoted by the path n. As the object moves in the path of n, the potential energy of the system decreases, as expressed in the following equation:

$$\begin{aligned} U(Z^{t+1}) &= U(Z^t) - \int_n F ds, \\ Z^{t+1} &= Z^t + \int_n ds, \end{aligned} \tag{3}$$

where $\int_n ds$ is the relative position movement along the path of n. By viewing optimization in terms of physics, we shed light on the behavior of the loss function and the optimization process. The concepts of potential energy and network gradients can be seen as two sides of the same coin, which can help us understand the nature of the optimization process and model the representation dynamics without loss backpropagation.

Rethinking transfer learning from the energy view. Based on the above findings, we propose to revisit transfer learning through the lens of potential energy. When the pre-trained model converges, the model "system" defined by the training objective reaches a state of relative stability with equilibrium potential energy. When the model is transferred to downstream tasks, this initial state becomes unstable due to changes in the potential energy landscape, resulting in an unreliable observation (representation). Therefore, it is inappropriate to predict the model transferability solely based on the current observations (static representations).

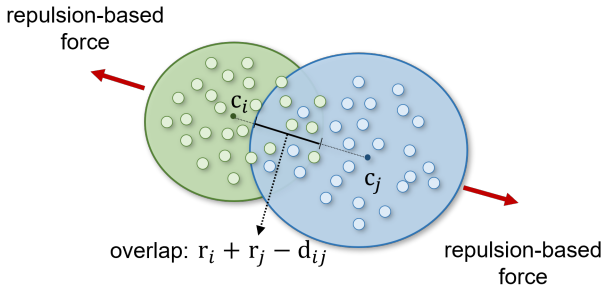

Figure 3. By treating the problem as minimizing the potential energy and viewing each cluster as a ball with overlap $(r_i + r_j - d_{ij})$ to another, we can simulate the system's dynamics by releasing the start point and observing how each ball is pushed away under the resulting forces.

During fine-tuning, the objective is to separate different classes in the latent space, which can be seen as a force separating clusters, similar to objects interacting in physics. In this context, the loss function $\mathcal{L}$ creates a potential energy plane U based on the relative positions of training sample features Z, with the current state being unstable and favoring a decrease in potential energy [10]. Therefore, it is crucial to capture the movement of the system to model the changing tendency of the current observation. To achieve this, we propose a physical scene for representation dynamics based on the concept of elastic potential energy that arises when an object is deformed under tension or stress. We use a repulsive force to simulate the interactions between different classes, improving the effectiveness and accuracy of the learning dynamics.

3.3. Modeling Representation Dynamics with Mechanical Motion

We propose a physical modeling approach named Potential Energy Decline (PED), that leverages potential energy to develop a mechanical motion process for representation dynamics. Specifically, each class of feature points Z_i in the embedding space is modeled as a multi-variate Gaussian distribution $\mathcal{N}(c_i, {\sigma_i}^2)$ with mean feature c_i and variance ${\sigma_i}^2$. Then we simplify it as a ball with c_i as its cen-

troid, $\lambda\|\sigma_i\|_2$ as its radius r_i, and a unit mass m_i. The force that repels different features is modeled as an elastic deformation force between the balls, similar to *Hooke's law* [26] (i.e., $\boldsymbol{F} = k\boldsymbol{x}$). Whenever two balls overlap, a force proportional to the deformation $\boldsymbol{x_e}$ is exerted in the direction of the vector $\boldsymbol{n}$ connecting the centers of the two balls (as shown in Fig. 4):

$$\begin{aligned} \boldsymbol{x_e} &= \max(\mathrm{r}_i + \mathrm{r}_j - \mathrm{d}_{ij}, 0), \\ \boldsymbol{F}_{ij} &= k\boldsymbol{x_e} \cdot \boldsymbol{n}, \end{aligned} \tag{4}$$

where k is a hyper-parameter of elasticity resisting coefficient, and d_{ij} denotes the distance $\|\mathrm{c}_i - \mathrm{c}_j\|_2$ between centers of the two balls.

As shown in Eq. (4), the force between two clusters becomes larger as the overlap becomes larger, and becomes zero when they move apart and no longer overlap. We further model the force from each ball to every other ball and sum up the forces to obtain the joint force $\boldsymbol{F}_i$ acting on ball i. To model the moving tendency, it is revealed by the acceleration $\boldsymbol{a}_i$ by *Newton's* second law of motion [33]:

$$\boldsymbol{a}_i = \frac{\boldsymbol{F}_i}{m_i} = \frac{\sum_{j \neq i} \boldsymbol{F}_{ij}}{m_i}. \tag{5}$$

In the field of physics, it is often assumed that force remains constant over a very short period of time. Following such a philosophy, we propose a method for simulating the phase position or relative position changes of a system by releasing it within a brief time interval Δt. The motion equation is then used to compute the position changes.

$$\overline{Z_i} = Z_i + \Delta Z_i = Z_i + \frac{1}{2}\boldsymbol{a}_i \cdot \Delta t^2. \tag{6}$$

By applying force to the samples in the system, they are effectively driven towards the direction of decreasing potential energy. By repeating this process multiple times, we obtain an even better system state $\overline{Z}$ with lower potential energy.

Discuss the feasibility of physical modeling. We can view our physics-inspired approach back to the other side of the coin, *i.e.*, conventional gradient-based perspective. We model the elastic potential of the system following *Hooke's law* [26] (i.e., $U = \frac{1}{2}k\boldsymbol{x}^2$) and the formulation is as follows,

$$\begin{aligned} U(Z) &= \sum\nolimits_i \sum\nolimits_{j \neq i} \frac{1}{2} k \boldsymbol{x}_{ij}^2 \\ &= \sum\nolimits_i \sum\nolimits_{j \neq i} \frac{1}{2} k \max(r_i + r_j - d_{ij}, 0)^2, \end{aligned} \tag{7}$$

where x_{ij} describes the overlap of the feature clusters between Z_i and Z_j. It is found that the form in Eq. (7) is analogous to gradient-based optimization methods that aim to minimize the overlaps of different clusters, which can be viewed as a pairwise loss to enhance class prototype separation in metric learning. In contrast, our approach offers a more efficient alternative to the optimization-based method by using a physical modeling approach to decrease energy potential, which can be easily integrated into existing methods for transfer learning dynamics.

3.4. Overall

Our physical modeling approach provides a refined observation $\overline{Z}$ of the system to take over the initial observation without performing updating the network. The dynamic representation is achieved by mechanical motion and more details of the proposed physics-driven approach can be found in Alg. 1. An arbitrary model selection metric, such as LogME[46], $\mathcal{M}(\overline{Z}, Y)$ can be adopted subsequently to rank the models, *i.e.*, obtaining $\{P_i\}_{i=1}^N$. This approach allows us to gain a better understanding of the system's dynamics and boost existing model ranking algorithms.

Algorithm 1: Algorithm of the proposed Potential Energy Decline (PED)

Input: Model zoo $\{\Phi_i\}_{i=1}^N$; Downstream labeled dataset $\mathcal{T} = \{X, Y\}$ including C classes; The hyper-parameter λ, Δ_t, k; Maximum iteration steps M, early termination condition ϵ; The model selection metric $\mathcal{M}$;
Output: The transferability score P_i for each model in the model zoo.

for Φ_i *in Model zoo* **do**
  Encode images X to feature embeddings $Z = \Phi_i(X)$ and normalize features with mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ of ImageNet features: $Z \leftarrow (Z - \hat{\mu})/\hat{\sigma}$;
  while *step* $\leq M$ **do**
    Compute the mean feature as ball center c_j and standard deviation σ_j of each class cluster Z_j
    $\mathrm{c}_j = \mathrm{E}[Z_j], \mathrm{r}_j = \lambda||\sigma_j||_2$;
    Compute distances of the feature clusters $\mathrm{d}_{jl} = ||\mathrm{c}_j - \mathrm{c}_l||_2, j \neq l \in \{1, ..., C\}$;
    Compute the force and the ac- celeration of each cluster $F_j = \sum_{l \neq j} k(r_j + r_l - d_{jl}) \cdot \frac{\hat{\mu}_j - \hat{\mu}_l}{||\hat{\mu}_j - \hat{\mu}_l||_2}, a_j = \frac{F_j}{m_j}$;
    Simulate the moving process and obtain a more stable state of features Z $z \leftarrow z + \frac{1}{2}a \cdot \Delta_t^2, z \in Z_j, j \in \{1, ..., C\}$;
    Calculate the ternminal condition $\omega[step] \leftarrow ||\sum_{j=1}^{C} \frac{1}{2}a \cdot \Delta_t^2||_1$;
    if $\omega[step] \leq \epsilon \cdot \omega[0]$ **then**
      break;
    end
    step $\leftarrow$ step+1;
  end
  Features revert back to the original space $Z \leftarrow Z \cdot \hat{\sigma} + \hat{\mu}$;
  Feed Z and Y into transferability predicting metric $\mathcal{M}(Z, Y)$ to obtain a score P_i;
end
Rank models in $\{\Phi_i\}_{i=1}^N$ according to their scores $\{P_i\}_{i=1}^N$.

Table 1. The experiment results of different transferability metrics on various self-supervised learning models, with the weighted Kendall's τ_w employed as the ranking correlation protocol. A larger τ_w represents a better prediction rank order to the ground truth rank. The best results are denoted in bold.

Method	Reference	Aircraft	Caltech101	Cars	Cifar10	Cifar100	Flowers	VOC	Pets	Food	DTD
$\mathcal{N}$LEEP [31]	CVPR'21	-0.029	0.525	0.486	-0.044	0.276	0.534	-0.101	0.792	0574	0.641
PARC [6]	NIPS'21	-0.03	0.196	0.424	0.147	-0.136	0.622	0.618	0.496	0.359	0.447
LogME [46]	ICML'21	0.223	0.051	0.375	0.295	-0.008	0.604	0.158	0.684	0.570	0.627
LogME+Ours	this paper	**0.509**	0.505	0.516	0.511	**0.667**	0.715	**0.620**	**0.795**	0.650	0.780
SFDA [40]	ECCV'22	0.254	0.523	0.515	0.619	0.548	0.773	0.568	0.586	**0.685**	0.749
SFDA+Ours	this paper	0.464	**0.614**	**0.647**	**0.673**	0.568	**0.777**	0.583	0.462	0.581	**0.907**
GBC [36]	CVPR'22	0.048	-0.18	0.424	0.008	-0.249	0.532	-0.041	0.655	0.268	0.05
GBC+Ours	this paper	0.462	0.285	0.547	0.017	0.359	0.768	-0.035	0.684	0.402	0.576

4. Experiment

In recent years, self-supervised learning has emerged as a dominant approach in vision pre-training, showing superior transferability compared to supervised learning methods. However, the potential learning dynamics can significantly impact the performance of traditional model transferability prediction metrics for self-supervised pre-trained models. Therefore, in this paper, we analyze the performance on self-supervised learning models to evaluate our proposed approach.

Downstream Dataset. In this study, we utilize a variety of widely-used datasets for transfer learning in downstream classification tasks, including FGVC Aircraft [32], Caltech-101 [21], Standford Cars [28], Cifar-10 [29], Cifar-100 [29], DTD [14], Oxford102 Flowers [35], Food-101 [7], and Oxford-IIIT Pets [37]. These datasets include diverse and comprehensive characteristics, such as street view, texture, and coarse/fine-grained scenes are suitable for our setting with diversity.

Pre-trained Model Zoo. To assess the generality of our method for self-supervised learning models, we consider 12 different types of pre-trained models with ResNet-50 [25], which have been developed using state-of-the-art self-supervised learning methods, including BYOL [22], Infomin [41], PCLv1 [30], PCLv2 [30], Selav2 [4], InsDis [45], SimCLRv1 [11], SimCLRv2 [12], MoCov1 [24], MoCov2 [13], DeepClusterv2 [8], and SWAV [9] [23]. Due to the limited space, some detailed information are provided in Appendix.

Ground Truth Model Rank. The construction of the ground truth rank $\{G_i\}_{i=1}^N$ for model zoo follows the implementation in [40, 46], where a grid search strategy is employed to compute the ground truth performance of each model in downstream task. Specifically, the grid search strategy includes a range of learning rates from the set $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ and weight decay values from the set $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$. To ensure robustness, each experiment was run with an average of 5 seeds. Obviously, it is observed from the above process that selecting the most suitable model through fine-tuning incurs a significant computational cost in terms of time and GPU resources.

Results of Existing Methods. We show the experiment results in the Table 1 and it is evident that the state-of-the-art methods encounter significant challenge in predicting transferability of self-supervised models, and even show inability to provide recommendations for some particular datasets, *e.g.,* the Kendall' weights are less than 0. For example, GBC, which achieves impressive performance in supervised learning scenarios by directly measuring the overlap degree of different clusters, exhibits a significant decline in prediction accuracy when tasked with an unstable initial state. Among the techniques evaluated, $\mathcal{N}$LEEP and SFDA exhibit better performances due to their implicit inclusion of a learning process aimed at adapting to downstream tasks. Nevertheless, the limited initial observations still hinder their performance, *e.g.,* the relative low performance in Aircraft. The above experiment results show that predicting the model performances of self-supervised models are not reliable with solely the initial observation and it is of great significance to take the representation dynamics into consideration.

Results of Our method. To evaluate the efficacy of our approach, we integrate our method upon different state-of-the-art transferability metrics, including evidence-based LogME [46], discrimination-based SFDA [40], separation-based GBC [36]. Through taking the dynamics representations into consideration, the performance combined with ours approach show obvious gains in many downstream scenes. For instance, our approach yields a significant performance gain of +0.675 and +0.608 compared to LogME and GBC on Cifar100, respectively. Even though SFDA has specifically designed Confix to alleviate the dynamics of fine-tuning by augmenting hard examples, it remains or-

thogonal to our method. For example, it achieves a gain of +0.210 and +0.158 on Aircraft and DTD, respectively. Although existing methods have achieved remarkable performance in experimental results (B), achieving above 0.6 in Kendall weight, our method can still provide diverse benefits upon different metrics. Our experiments confirm the effectiveness of our physics-inspired modeling approach and highlight the significance of considering representations in self-supervised models.

5. Ablation Study

In this section, we perform an ablation study of our method on downstream datasets of Cars, Flowers, and DTD. Specifically, we investigate the impact of the hyperparameters and implementation details of our method. Through these experiments, we aim to gain a deeper understanding of our method's performance and identify key factors that contribute to its effectiveness.

5.1. Period of Time

The period of time Δ_t is a crucial factor that determines the degree of the dynamic process. In Fig. 4, we conduct ablation experiments to investigate the effect of varying Δ_t. The results demonstrate that increasing the time of movement drives the clusters to decline the potential energy, which in turn leads to improved observation and performance gain.

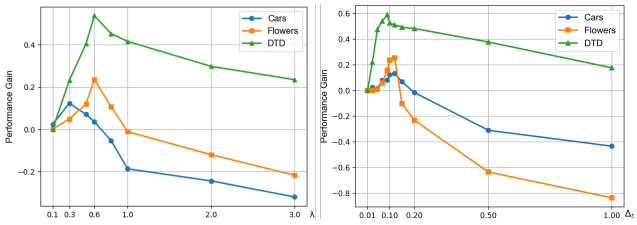

Figure 4. The performance gain in Kendell τ with respect to different λ and Δ_t.

In physics, force conditions are typically approximated as constant over a short period of time. However, as the time period increases, the clusters move further apart, causing the force to change significantly and introducing errors into the physical assumption. Consequently, we set the time period to a small value of 0.1 for all experiments.

5.2. The Radius Coefficient λ

As we represent the features of each cluster as a Gaussian distribution, and simplify it as a ball in a physical view, we adopt the radius coefficient λ controls the degree of modeling the overlaps among different clusters and set hyperparameter k to be 1.0 by default.

Through ablation experiments presented in Fig. 4, we interestingly observed that setting λ to 0.3 yields in better performance in datasets with significant data cluster overlaps, such as Cars, whereas setting λ to 0.6 was more effective for datasets with less cluster differences, such as Cifar10. Consequently, we use the set of $\{0.3, 0.6\}$ as candidate radius coefficients for all downstream tasks. Our findings indicate that controlling the dynamic process can enhance the model's performance and that selecting an appropriate radius coefficient is crucial in achieving optimal results.

5.3. Multiple Step Moving

To decrease the potential energy and achieve the desired effect, we employ a multi-step moving process, where the exit ratio ϵ and the maximum number of steps M determine the end condition. As shown in Fig. 5, it is evident that the force decreases rapidly within a few steps, indicating that the clusters are being pushed apart and the interaction force is decreasing.

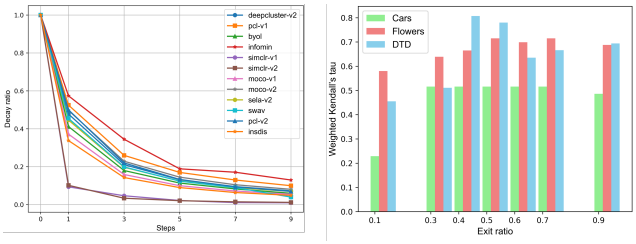

Figure 5. The left picture is the decay curve of force changing with steps M and the right picture shows the influence of different end condition to the performance.

According to the results, we set the attenuation condition and the maximum number of steps to be 5 to terminate the moving process when the current force is less than $\epsilon = 0.5$ of the initial force in the experiments. The multiple step moving approach allows us to model the movement in a short period of time, while taking into account the updated feature positions and re-calculating the interaction force.

6. Visualization and Analysis

In this section, we present visualization results to assess the effectiveness of our proposed method and to gain a deeper understanding of its mechanisms.

6.1. Different Cluster Change

The visualization in Fig. 6 reveals that the initial state of clusters, encoded by self-supervised pre-trained models, is not well-distributed. While the different clusters should be apart from the others when transferring to downstream task. Therefore, this highlights the potential for improvement that our method offers. To further investigate the underlying processes, we take the BYOL model on Cifar100

dataset and employ the t-SNE algorithm [43] to visualize our dynamic modeling process.

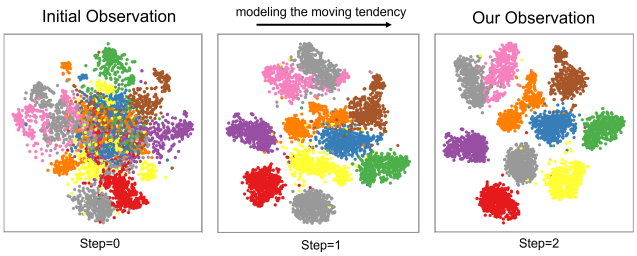

Figure 6. The dynamic representation of our modeling process on self-supervised models.

As shown in Fig. 6, the initial observations are rather chaotic and samples are not clearly separated by class in the t-SNE visualization, which is due to the lack of class information in the self-supervised pre-training. Therefore, predicting the transferability of the initial state is unreliable. However, our dynamic modeling process improves the separability of the feature clusters, achieving a similar effect to fine-tuning without requiring network updates. Overall, the visualization results provide strong evidence in support of the effectiveness of our proposed method, and offer insights into the underlying processes that contribute to its success.

6.2. Model Rank

By utilizing our proposed method, we are able to refine models that have a relatively low transferability score as a result of initial unstable observations. To evaluate the efficacy of our approach, we generated a visualization of the model rank comparisons between the initial observation and our refined observation, as shown in Fig. 7.

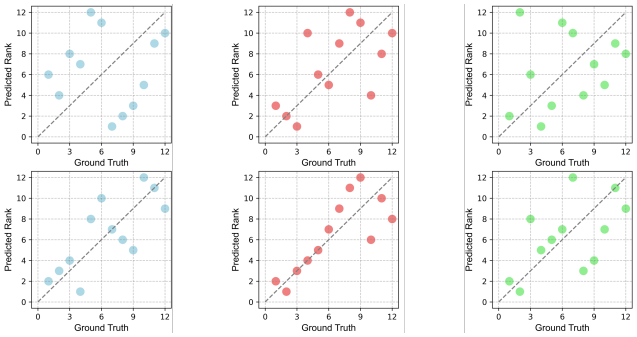

Figure 7. We present a visualization of the variation in model ranking concerning the prediction of transferability scores, utilizing both our own observations and initial observations. The depicted progression spans three datasets: DTD, Flowers, and Cars, arranged from left to right.

The results demonstrate a significant improvement in the calibration of model rankings when utilizing our refined observation, where the optimal model is positioned near the reference line in the figure. Notably, some models that begin with an unfavorable starting point can be rapidly improved to achieve a superior ranking, highlighting the effectiveness of our refinement methodology.

7. Efficiency Analysis

In the results of our experiment, we have demonstrated the effectiveness of our proposed method in enhancing various transferability prediction techniques. Additionally, we highlight in this section that our approach exhibits computational efficiency in terms of algorithmic complexity and practical running time. Our method is computationally efficient due to the simplification in physics that we have adopted. Specifically, we model the same class features as a whole ball and consider the stress of the class center exclusively, resulting in a computational complexity of $\mathcal{O}(C^2D)$, where C represents the number of classes and D denotes the feature dimension. Consequently, our method produces a relatively small time overhead in comparison to the transferability prediction process. We present our experimental findings on the running time of our approach in Table 2. Notably, our method contributes minimal overhead to the transferability prediction process, further attesting to its effectiveness and efficiency in computation.

Table 2. The comparisons of running time on Flowers.

Metrics	PARC	LogME	NLEEP	SFDA	GBC	Ours
Running Time	27s	8s	392s	82s	11s	2s

8. Conclusions and Future Work

This paper presents a fresh insight to reframe transfer learning as a process of decreasing the system potential energy. To this end, we propose physically motivated modeling technique that effectively captures the dynamics of representations. Despite being a simplified physical modeling approach, our method consistently boosts the existing metrics for ranking the self-supervised pre-trained models. In the future work, we intend to enhance the sophistication of our physical model by incorporating adaptive hyperparameters and expanding its applicability to more transfer learning scenes. We hope our work will shed light on the representation dynamics in transfer learning and inspire further research in this field.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant 62088102, and in part by the PKU-NTU Joint Research Institute (JRI) sponsored by a donation from the Ng Teng Fong Charitable Foundation.

References

[1] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *ICCV*, pages 6430–6439, 2019.

[2] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, pages 147–169, 1985.

[3] Andrea Agostinelli, Michal Pándy, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. How stable are transferability metrics evaluations? In *ECCV*, pages 303–321, 2022.

[4] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019.

[5] Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *ICIP*, pages 2309–2313, 2019.

[6] Daniel Bolya, Rohit Mittapalli, and Judy Hoffman. Scalable diverse model selection for accessible transfer learning. *NeurIPS*, pages 19301–19312, 2021.

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014.

[8] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, pages 132–149, 2018.

[9] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 33:9912–9924, 2020.

[10] T.M. Charlton. Energy principles in theory of structures. In *Oxford University Press*, 1973.

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

[12] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *NeurIPS*, pages 22243–22255, 2020.

[13] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[14] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[15] Quan Cui, Bingchen Zhao, Zhao-Min Chen, Borui Zhao, Renjie Song, Boyan Zhou, Jiajun Liang, and Osamu Yoshie. Discriminability-transferability trade-off: an information-theoretic perspective. In *ECCV*, pages 20–37, 2022.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[17] Nan Ding, Xi Chen, Tomer Levinboim, Soravit Changpinyo, and Radu Soricut. Pactran: Pac-bayesian metrics for estimating the transferability of pretrained models to classification tasks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, pages 252–268. Springer, 2022.

[18] Yueqi Duan, Jiwen Lu, and Jie Zhou. Uniformface: Learning deep equidistributed representation for face recognition. In *CVPR*, pages 3415–3424, 2019.

[19] Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In *CVPR*, pages 12387–12396, 2019.

[20] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *CVPR*, pages 5414–5423, 2021.

[21] Li Fei-Fei. Learning generative visual models from few training examples. In *CVPR workshop*, 2004.

[22] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *NeurIPS*, pages 21271–21284, 2020.

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

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[26] Robert Hooke. De potentia restitutiva, or of spring explaining the power of springing bodies. *Royal Society of London Philosophical Transactions*, 1678.

[27] John J Hopfield. Hopfield network. *Scholarpedia*, page 1977, 2007.

[28] Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. 2013.

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

[30] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[31] Yandong Li, Xuhui Jia, Ruoxin Sang, Yukun Zhu, Bradley Green, Liqiang Wang, and Boqing Gong. Ranking neural checkpoints. In *CVPR*, pages 2663–2673, 2021.

[32] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[33] Isaac Newton. *The Principia: Mathematical Principles of Natural Philosophy*. University of California Press, 1999.

[34] Cuong Nguyen, Tal Hassner, Matthias Seeger, and Cedric Archambeau. Leep: A new measure to evaluate transferabil-

ity of learned representations. In *ICML*, pages 7294–7305, 2020.

[35] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[36] Michal Pándy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. Transferability estimation using bhattacharyya class separability. In *CVPR*, pages 9172–9182, 2022.

[37] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505, 2012.

[38] Saroj Rout, Pramod Srivastava, Jharna Majumdar, et al. Multi-modal image segmentation using a modified hopfield neural network. *Pattern Recognition*, pages 743–750, 1998.

[39] Ruslan Salakhutdinov and Geoffrey E. Hinton. Deep boltzmann machines. In *AISTATS*, 2009.

[40] Wenqi Shao, Xun Zhao, Yixiao Ge, Zhaoyang Zhang, Lei Yang, Xiaogang Wang, Ying Shan, and Ping Luo. Not all models are equal: Predicting model transferability in a self-challenging fisher space. In *ECCV*, pages 286–302, 2022.

[41] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *NeurIPS*, 33:6827–6839, 2020.

[42] Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *ICCV*, pages 1395–1405, 2019.

[43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[44] Ross Wightman. Pytorch image models, 2019.

[45] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018.

[46] Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. Logme: Practical assessment of pre-trained models for transfer learning. In *ICML*, pages 12133–12143, 2021.

[47] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, pages 3712–3722, 2018.

[48] Kai Zhao, Jingyi Xu, and Ming-Ming Cheng. Regularface: Deep face recognition via exclusive regularization. In *CVPR*, pages 1136–1144, 2019.