COMMON CAUSES FOR SUDDEN SHIFTS: LINKING PHASE TRANSITIONS IN SINUSOIDAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Different phases of learning dynamics exist when training deep neural networks. These can be characterised by statistics called order parameters. In this work we identify a shared, underlying mechanism connecting three seemingly distinct phase transitions in the training of a class of deep regression models, specifically Implicit Neural Representations (INRs) of image data. These transitions include: the emergence of wave patterns in residuals (a novel observation), the transition from fast to slow learning, and Neural Tangent Kernel (NTK) alignment. We relate the order parameters for each phenomenon to a common set of variables derived from a local approximation of the structure of the NTK. Furthermore, we present experimental evidence demonstrating these transitions coincide. Our results enable new insights on the inductive biases of sinusoidal INRs.

1 INTRODUCTION

Implicit neural representations (INRs) are increasingly employed as differentiable alternatives to traditional, discretized signal representations with expressive models such as SIREN(1). By capitalising on the expressive power of Deep Neural Networks (DNNs), INRs provide a compact framework for capturing high-frequency details and spatial derivatives of signals. However, despite their growing popularity in practical applications, our theoretical understanding of their behaviour remains nascent. A crucial aspect that requires deeper insight is the nature of their inductive biases: what features of the data are learned by these models, and through what mechanisms does this learning occur?

An increasing body of research emphasizes the pivotal role played by the optimization algorithm in learning representations. Indeed, it is known that the full expressivity of DNNs - parameter count notwithstanding - is constrained in practice by the limitations of Gradient Descent (GD) in exploring the loss landscape (2; 3). Furthermore, it is known that neural networks learn patterns of different complexity at different rates, resulting in distinct learning phases (4; 5). These phases may be identified by examining changes in the collective evolution of the model's weights, as quantified by summary statistics known as order parameters (6; 7; 8; 9). In statistical mechanics, these parameters quantify symmetries of a system, and change suddenly at a phase transition (10). Although various statistics have been independently identified in the DNN literature (11; 12; 13; 14), no underlying symmetry connects them, and their interrelationships remain unclear. What's more, while order parameters can indicate the timing of phase transitions, they offer limited insight into what DNNs learn during these phases.

By contrast, the Neural Tangent Kernel (15) (NTK) provides a complimentary perspective that manifestly describes how datapoints influence one another during training. Critically, in a phenomenon known as Neural Tangent Kernel Alignment (NTKA), the NTK undergoes significant changes early in training as it engages in feature learning, aligning with the target function. NTKA has been widely documented and is suggested as a reason why real-world DNNs often outperform their infinite-width limit counterparts (16; 17; 18; 19; 20; 21). Despite repeated empirical demonstrations of NTKA, theoretical exploration of the phenomenon has been largely restricted to toy models and classification problems, leaving a gap in understanding the transition in complex regression tasks.

In this work, we explore the phenomenon of NTKA within SIRENs, to determine when this alignment occurs, the driving factors behind it, and its implications for feature learning. Our study is structured around four primary contributions:

1. We study the dynamics of learning in DNNs through three different lenses (training curve, residual evolution, and NTKA), and in each, observe a phase transition. The phase transitions are related to a common underlying phenomenon, as evidenced by their coincidence in time (as quantified by order parameters).
2. We identify that one phase of learning on this task is characterised by diffusion-like wave-crests, and demonstrate analytically that this transition is related to the evolution of the NTK during training.
3. We construct a local approximation of the NTK that allows us to relate the order parameters to spatial variations in the parameter gradients - a hallmark of translational symmetry breaking. We introduce a new order parameter (MAG-Ma) based on these spatial variations. Finally, we connect the symmetry-breaking perspective with insights from traditional computer vision.
4. We investigate the effects of model hyperprameter choices and demonstrate how these order parameters can be used to introspect the learning process.

2 PRELIMINARIES

In this work, we consider the class of INRs that model 2D grayscale images, where pixel coordinates and their intensity form a dataset $\mathcal{D}$ of N samples indexed with i, $(x_i, f_{true}(x_i))$, where $x_i \in \mathbb{R}^2$ and $f_{true} : \mathbb{R}^2 \mapsto \mathbb{R}$. On this dataset, we fit SIREN models (1) f with parameters θ, using sinusoidal activation functions. In the continuum limit, we identify two fields: the local residual field $r(x;\theta(t)) = f_{true}(x) - f(x;\theta(t))$, and gradient field $\nabla_\theta f(x;\theta(t))$. Their time evolution is induced by gradient flow $\dot{\theta} = -\nabla_\theta L$ on the mean square error:

$$L(\theta) = \frac{1}{2}\int dx\ P_{data}(x)\ r(x;\theta)^2 \tag{1}$$

We assume the data is distributed uniformly according to $P_{data}(x) = \text{Vol}(\mathcal{D})^{-1}$. Accordingly, leveraging gradient flow and the chain rule, the residuals evolve as follows:

$$\begin{aligned}
\dot{r}(x;\theta(t)) &= \nabla_\theta r(x;\theta(t)) \cdot \dot{\theta} && (2)\\
&= -\frac{1}{\text{Vol}(\mathcal{D})}\int dx'\ r(x')\nabla_\theta r(x;\theta(t)) \cdot \nabla_\theta r(x';\theta(t)) && (3)\\
&= -\int dx'\ r(x')\Big(\frac{1}{\text{Vol}(\mathcal{D})}\nabla_\theta f(x;\theta(t)) \cdot \nabla_\theta f(x';\theta(t))\Big) && (4)\\
&:= \int dx'\ r(x')K_{NTK}(x,x';\theta(t)) && (5)
\end{aligned}$$

In the last line, we defined K_{NTK}, the Neural Tangent Kernel. Going forward, for notational brevity, we will drop the explicit dependence on θ, and write $x' = x + u$. We also define a kernel closely related to the NTK, the Cos NTK:

$$C_{NTK}(x, x+u) = \frac{1}{\text{Vol}(\mathcal{D})}\frac{\nabla_\theta f(x) \cdot \nabla_\theta f(x+u)}{||\nabla_\theta f(x)||\ ||\nabla_\theta f(x+u)||} \tag{6}$$

3 DERIVING ORDER PARAMETERS FROM THE NTK

We illustrate the different phases of learning with a motivating example. In Figure 1, we train a five-layer deep, 256-unit wide SIREN model on a 128×128 grayscale image of a camera-man, using full-batch GD with a learning rate of 10^{-3} for 2000 epochs. Our validation task is super-resolution reconstruction of the original image. During training, we examine the model's behaviour through three different lenses, with a sudden, qualitative shift revealed in each.

While these shifts are visually striking, in this section, we take a more quantitative approach based on the identification of order parameters. We then demonstrate why these phase transitions occur simultaneously, by relating the order parameters to a common set of features, which control the local structure of the NTK. The three lenses are as follows:

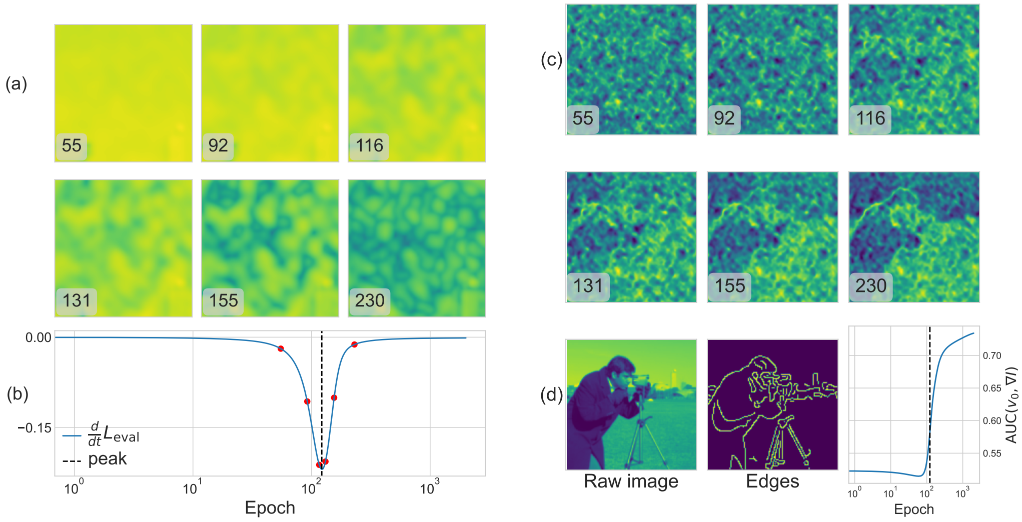


Figure 1: **A Single Phase Transition Through Three Lenses**: (a) The spatio-temporal evolution of the loss, as revealed through the magnitude of the residuals. Near the critical point we see the formation of wavecrests. (b) Evolution of the loss rate during training. The rate of change of the validation loss reaches a peak at the critical point. (c) Evolution of the principle eigenvector of the NTK reveals a sudden shift from disorder to learned features. (d) Quantification of NTKA in terms of alignment between edges and the principal eigenvector.

- **Spatial Distribution of Residuals**: Early in training, the loss decreases uniformly over the dataset (Drift Phase). However, at a critical point, we observe the formation of "wave-crests" corresponding to regions of low-loss, which propagate across the dataset (Diffusion Phase). To the best of our knowledge, we are the first to report this behaviour in SIREN models. In Section 3.1, we attribute this behaviour to changes in the equal-time correlation functions of the gradient field $\nabla_\theta f(x)$, whose parameters we derive in Section 3.2.
- **Principal Eigenvectors of the NTK**: Early in the training, the principal eigenvector v_0 is static and appears as a highly-disordered, structureless image (Disordered Phase). However, at a critical point, v_0 rapidly aligns with the edges of the image (Aligned Phase), after which it becomes static again. Though NTK alignment has been previously studied in the context of classification problems (22; 23; 24; 25), there are additional subtleties to consider for a regression task like INR training. To this end, we introduce a metric, $\text{AUC}(v_0, \nabla I)$ in Section 3.3 to identify when alignment occurs. We also derive an approximation of v_0 based on the local structure of the NTK, as outlined in Sections 3.1 and 3.2.
- **Training Curve Analysis**: There is a rapid shift in the slope of the training curve, which we call the loss rate $\dot{L}$. Initially, $\dot{L}$ is large, indicating the Fast Phase. After a critical point, the loss rate collapses, and learning slows (Slow Phase). Several works have studied this transition using order parameters, but in this work, we focus on the concept of gradient confusion, as described in (12), (13), (14). In Section 3.4, we derive an approximation of this parameter based on the local structure of the NTK outlined in Section 3.2.

Having united the different order parameters, we are in a better position to speculate on the common origin of the underlying phase transition. Motivated by the dependence of each parameter on spatial variations in the magnitude field $||\nabla_\theta f(x)||$, we introduce a new parameter, termed MAG-Ma, in Section 3.5. MAG-Ma explicitly tracks violations in the translational symmetry of the NTK.

3.1 CORRELATION FUNCTIONS AND THE ONSET OF DIFFUSION

The form of equation 5 is reminiscent of the linear response functions in statistical field theory (10; 26): to find the rate of change of the residual field at a point x, the kernel K aggregates information about the residual at points $x + u$. To quantify the range of these interactions, we may examine the rate

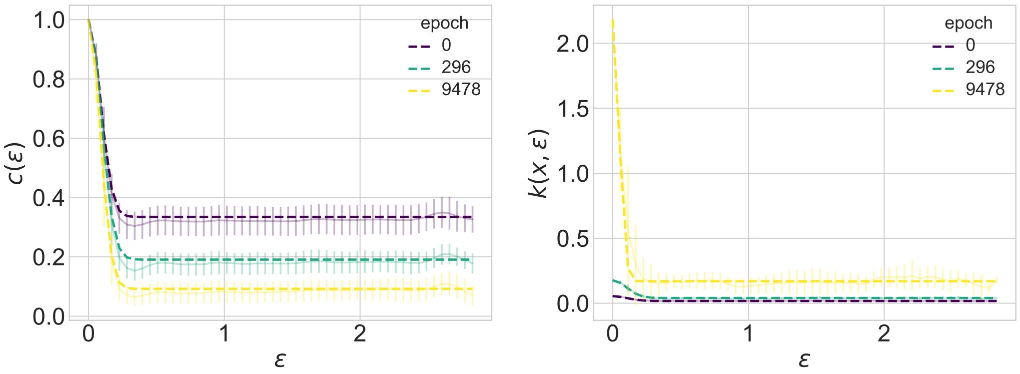


Figure 2: Visualization showing the empirical correlation function for the normalized parameter gradients. On the left-hand side is the global correlation-function for the C_{NTK}. On the right is the local-correlation function for the K_{NTK} around a test point x. Dashed lines show fitted Gaussian approximation, and error bars show variance across dataset. Over the course of training, both the global correlation lengthscale ξ_{corr}, and the terminal value c_∞, evolve.

at which correlations decay with distance. For the field $\nabla_\theta f(x)$, the local, equal-time correlation function measures the average alignment between the gradients at points separated by a distance ϵ:

$$k(x,\epsilon) = \mathbb{E}_\phi\left[\nabla_\theta f(x)\cdot\nabla_\theta f(x+\epsilon\hat{e}_\phi)\right] \tag{7}$$

$$= \mathbb{E}_\phi\left[K_{NTK}(x, x+\epsilon\hat{e}_\phi)\right] \tag{8}$$

Here, $\hat{e}_\phi$ denotes a unit vector pointing in the direction ϕ. Similarly, the global, equal-time correlation is given by:

$$k(\epsilon) = \mathbb{E}_x\left[k(x,\epsilon)\right] \tag{9}$$

Here, the expectation is taken uniformly over the unit vectors $\hat{u}$. We may define similar quantities for the C_{NTk}, which we denote by $c(x,\epsilon)$ and $c(\epsilon)$. To estimate the correlation functions empirically, we group pairs of datapoints based on their distance, and then compute the mean value of the C_{NTk} for the group. For SIREN models, we observe that the equal-time correlation functions are well-approximated by Gaussians of the form:

$$c(\epsilon) \approx (1-c_\infty)e^{-\epsilon^2/2\xi_{corr}^2} + c_\infty \tag{10}$$

$$k(x,\epsilon) \approx ||\nabla_\theta f(x)||^2(1-c_\infty(x))e^{-\epsilon^2/2\xi(x)^2} + ||\nabla_\theta f(x)||^2 c_\infty(x) \tag{11}$$

This is illustrated in Figure 2. This approximation introduces two important order parameters: the first, the correlation length-scale ξ_{corr}, controls the rate at which correlations decay with distance, defining the range of interactions. The second, the asymptotic value c_∞, describes the interactions between points at separations ϵ much greater than ξ, where the gradient field vectors become uncorrelated. We have:

$$\lim_{\epsilon\to\infty} c(\epsilon) = c_\infty = \left|\left|\mathbb{E}_x\left[\frac{\nabla_\theta f(x)}{||\nabla_\theta f(x)||}\right]\right|\right|^2 \tag{12}$$

Dynamically, we see from Figure 2 that both ξ and c_∞ evolve during training, and we shall demonstrate that changes in these values account for the onset of diffusion. When c_∞ decays to zero, we have, as a very simple approximation of the NTK:

$$K(x, x+u) \approx ||\nabla_\theta f(x)||^2 \exp(-||u||^2/\xi^2(x)) \tag{13}$$

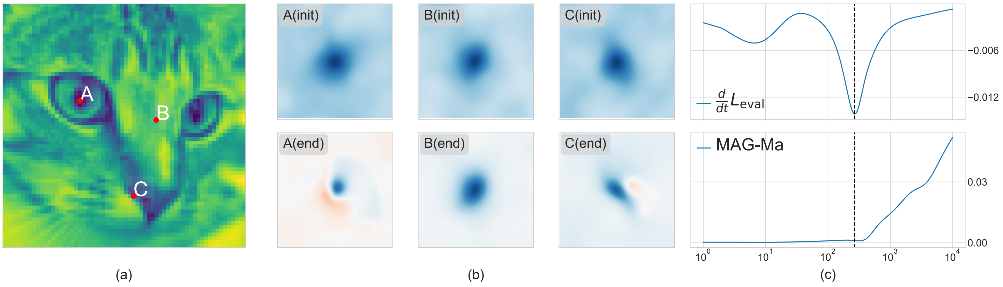
When $\xi(x)$ is small, the NTK will suppress all contributions to the residual $\dot{r}(x)$ except from the immediate vicinity of x. As such, performing a Taylor expansion to second order in u, we obtain:

$$r(x+u;\theta) \approx r(x;\theta) + u^\top \nabla_x r(x;\theta) + \frac{1}{2}u^\top \nabla_x^2 r(x;\theta)u \tag{14}$$

Inserting this, along with the NTK approximation, into equation 5, the full integral may be solved using Gaussian integration (full details in Appendix A.2). We obtain:

$$\frac{d}{dt}r(x;\theta) = -2\pi\xi^2(x)||\nabla_\theta f(x)||^2 r(x) - \pi\xi^4(x)||\nabla_\theta f(x)||^2\Delta^2 r, \tag{15}$$

which resembles a standard diffusion equation.



(a) (b) (c)

Figure 3: **Evolution of the Cosine NTK**: We visualize $C_{NTK}(x, x + u)$ around three points $x \in \{A, B, C\}$ for small separations u. At initialization, C_{NTK} locally resembles an isotropic, translation-invariant RBF. However, as training progresses, these symmetries are broken. MAG-Ma (described in Section 3.5) is an order-parameter that monitors the original symmetry, and changes at the critical point.

3.2 BEYOND THE ISOTROPIC GAUSSIAN APPROXIMATION

Though the isotropic Gaussian approximation of the NTK can explain the appearance of the diffusion wavecrests, empirically, the NTK is anisotropic (see Figure 3). What's more, the isotropic Gaussian approximation is positive definite, whereas the real NTK takes on negative values. In this section, we develop a better local approximation that overcomes these limitations. Our approach has the additional benefit that we may predict the correlation length-scale, along with other order parameters.

Our starting point is the local structure of the Cos NTK. The full details of our derivation are found in Appendix A.3, but the main strategy is to leverage the law of cosines to express the C_{NTK} as:

$$C_{NTK}(x, x+u) = \frac{||\nabla_\theta f(x)||^2 + ||\nabla_\theta f(x+u)||^2 - ||\nabla_\theta f(x+u) - \nabla_\theta f(x)||^2}{2||\nabla_\theta f(x)||\,||\nabla_\theta f(x+u)||} \tag{16}$$

Performing a Taylor expansion in u and retaining terms only up the second order, we find the Cosine NTK locally takes the form of a Cauchy Distribution:

$$C_{NTK}(x, x+u) \approx \frac{2a_x^2 + u^\top D_x}{2a_x^2 + u^\top D_x + u^\top H_x u}, \tag{17}$$

where we have:

$$a_x = ||\nabla_\theta f(x)|| \tag{18}$$

$$D_x = \nabla_x ||\nabla_\theta f(x)||^2 \tag{19}$$

$$H_x = (\nabla_x \nabla_\theta f(x))(\nabla_x \nabla_\theta f(x))^\top \tag{20}$$

To obtain a correlation length-scale from this anisotropic model, we note that the level sets of equation 17 correspond to ellipses. For a given value c, the area of the level set can be shown to be (see Appendix A.4):

$$A_{ellipse}(x; c) = \frac{\pi}{\sqrt{\det H}} \left(\frac{2(1-c)}{c} a_x^2 + \frac{(1-c)^2}{4c^2} D^\top H^{-1} D \right) \tag{21}$$

To take into account the asymptotic value of C_{NTK}, we choose $c = 1/2 + c_\infty/2$. Then:

$$\xi(x) \approx \sqrt{\frac{A_{ellipse}(x; 1/2 + c_\infty/2)}{\pi}} \tag{22}$$

3.3 ORDER PARAMETERS FOR THE ONSET OF NTK ALIGNMENT

In the classification problems typically studied in the NTKA literature, the principle eigenvector v_0 is seen to learn class-separating boundaries (22; 23). Similarly, for our 2D image reconstruction task, we see the NTK learns information about the distribution of edges in the image (Figure 4).

To quantify this alignment, we use a Canny Edge Detector (27) to estimate connected image edges. We then quantify the utility of v_0 in predicting edges in terms of average recall, as measured by the area under the Receiver Operating Characteristic Curve (ROC AUC). We denote this measure $\text{AUC}(v_0, \nabla I)$, and it has the advantage of being insensitive to monotonic transformations of v_0.

Another hallmark of NTKA is early anisotropic growth of the spectrum of the NTK (23), as the NTK becomes stretched along a relatively small number of directions that are correlated with the task. This is especially the case for the principal eigenvalue λ_0, which becomes orders of magnitude larger than the next leading eigenvalue. In subsequent sections, we will demonstrate, empirically, that this also holds during INR training.

The divergence of λ_0 enables a particularly simple approximation of the princpal eigenvector v_0. Namely, because the principal eigenvalue is so dominant, K_{NTK} becomes effectively low-rank, and so power iterations converge quickly. Thus, choosing a vector of ones $v = 1$ as our initial vector, we expect $K1/1^\top 1$ to have strong cosine alignment with the principal eigenvalue. In the continuum limit, this is simply given by:

$$\begin{aligned} K1/N &\to \mathbb{E}_u[K(x, x+u)] && (23) \\ &= \mathbb{E}_\epsilon[\mathbb{E}_u[K(x, x+u)|\ ||u|| = \epsilon]] && (24) \\ &= \int_0^{\epsilon_{max}} d\epsilon\ k(x, \epsilon) P(x, \epsilon) && (25) \end{aligned}$$

Here, $P(x, \epsilon)$ denotes the density of points that are located a distance ϵ from the point x, and ϵ_{max} is an upper bound on the distance that we assume is much greater than ξ_{corr}. Close to this x[1], $P(x, \epsilon)$ grows like $2\pi\epsilon$. Thus, leveraging equations 11 and 18, we have:

$$\begin{aligned} \mathbb{E}_u[K(x, x+u)] &= 2\pi a_x^2 \int_0^{\epsilon_{max}} d\epsilon\ \epsilon \left[c_\infty(x) + (1 - c_\infty(x)) e^{-\epsilon^2/2\xi^2(x)} \right] && (26) \\ &= 2\pi a_x^2 \left[c_\infty(x)\epsilon_{max}^2 + \xi^2(x)(1 - c_\infty(x))(1 - e^{-\epsilon_{max}^2/2\xi^2(x)}) \right] && (27) \\ &\approx a_x^2 \left[c_\infty(x)\text{Vol}(\mathcal{D}) + 2\pi\xi^2(x)(1 - c_\infty(x)) \right] && (28) \\ &\approx v_0(x) && (29) \end{aligned}$$

As we approach the phase transition, the asymptotic values tend towards 0, and the second term dominates. Considering the approximation in equation 22 for the correlation length-scale ξ, we note that $v_0(x)$ grows as $\mathcal{O}(||\nabla_\theta f(x)||^4)$. This implies particular sensitivity to pixels in regions with substantial high-frequency information, such as edges and corners. As natural images tend to be piecewise smooth, pixels on boundaries have the strongest spatial gradients within their neighbourhood, and are therefore the greatest source of information, being poorly compressible due to the lack of smoothness/redundancy, and accordingly disagreement in paramater gradients. Given the inability of models to accurately describe sharp discontinuities these edge pixels can be considered as influential datapoints, which accounts for their prominence within the principal eigenvector. We consider parallels between these observations of the NTK principal eigenvector and traditional approaches from the image processing literature concerning corners and edges in Appendix E. The fidelity of our approximation is evaluated in Appendix F.

3.4 ORDER PARAMETERS FOR THE LOSS RATE COLLAPSE

In (12), (13), (14), and related works, the authors examine the role of gradient alignment statistics in determining the speed of learning under stochastic gradient descent. They note that the emergence of negative alignments between batches correlates with a reduction in learning speed. Intuitively, when gradient alignment becomes negative, the sum of the gradients approaches zero, resulting in a diminished learning signal. The minimum alignment between the gradients is simply given by the minimum value of the Cos NTK, which we may obtain explicitly from equation 17 as follows (full derivation in Appendix A.5):

$$\min_u C_{NTK}(x, x+u) = \frac{D_x^\top H_x^{-1} D_x}{D_x^\top H_x^{-1} D_x - 8a_x^2} \tag{30}$$

---

[1] The true form of $P(x, \epsilon)$ is complicated and varies from point to point, due to edge effects. However, these effects are suppressed as $P(x, \epsilon)$ only appears when multiplied the Gaussian k_x.

$\min C_{NTK}$ is then simply the minimum of 30 across the whole dataset.

3.5 MAG-MA: ORDER PARAMETERS FROM TRANSLATIONAL SYMMETRY BREAKING

While previous sections have focused on a bottom-up construction and analysis of order parameters, this section adopts a top-down approach rooted in symmetry principles. Within the framework of statistical field theory, symmetry plays a crucial role in analyzing phase transitions by enabling the classification of distinct interaction mechanisms in complex systems. Specifically, the alteration of symmetry properties during a phase transition provides key insights into the nature of the transition and informs the construction of appropriate order parameters.

In Sections 3.1-3.4, we expressed several order parameters in terms of the parameters a, D, H, which characterize the local structure of the C_{NTK}. Tellingly, each of these parameters is now a function of the spatial variation of the parameter gradients. This suggests it is a translation symmetry which is broken at the phase transition. Indeed, from Figure 3, we observe that the C_{NTK} is an approximately stationary, isotropic kernel. Phrased another way, the Kernel exhibits no bias for location or direction. Over the course of training, we may monitor the emergence of such a bias with the following metric :

$$||\mathbb{E}_x[\nabla_x \log ||\nabla_\theta f||^2]||^2 = ||\mathbb{E}_x[D_x/a_x^2]||^2 \tag{31}$$

We refer to this statistic as **MAG-Ma**: the **M**agnitude of the **A**verage **G**radient of the Log Gradient-Field **Ma**gnitudes. Intuitively, this order parameter captures the statistical preference for a spatial direction in the dataset. The evolution of this quantity is plotted in Figure 3, and its alignment with the other order parameters is shown in Figure 4. We see that throughout the Fast Phase of training (before the peak in the loss rate $\dot{L}_e val$), the local structure of the C_{NTK} is statistically translation invariant, and MAG-Ma is close to zero. However, just after the critical point, it grows rapidly - coinciding with the structure learning described in Section 3.3.

4 EXPERIMENTAL RESULTS

4.1 EXAMINING THE DISTRIBUTION OF CRITICAL POINTS

In this section, we demonstrate that the critical points defined in Section 3 all cluster around a common time. We train a range of SIREN models (1) on fifteen images (Figure 9), varying seed, depth, and width (full details can be found in Appendix B). We also vary ω_0 (which we term the bandwidth), an important hyperparameter which multiplies the pre-activations before non-linearity in SIRENs. We illustrate the results of this sweep in Figure 4. In addition to the order parameters described in Section 3, we consider three more order parameters from the literature, which may be defined in terms of the NTK:

- We track the principal eigenvalue λ_0 of the NTK.
- In (11) and others, the authors consider the impact of the norm and standard deviations of model parameters on the loss rate. During the fast learning phase, the means of model gradients are large and the variances are small, with the converse true in the slower phase. To this end, we monitor spikes in the variance of the weight gradients, as measured by the trace of the covariance matrix. In terms of the NTK and the residual r, this corresponds to (see Appendix A.6 for derivation):

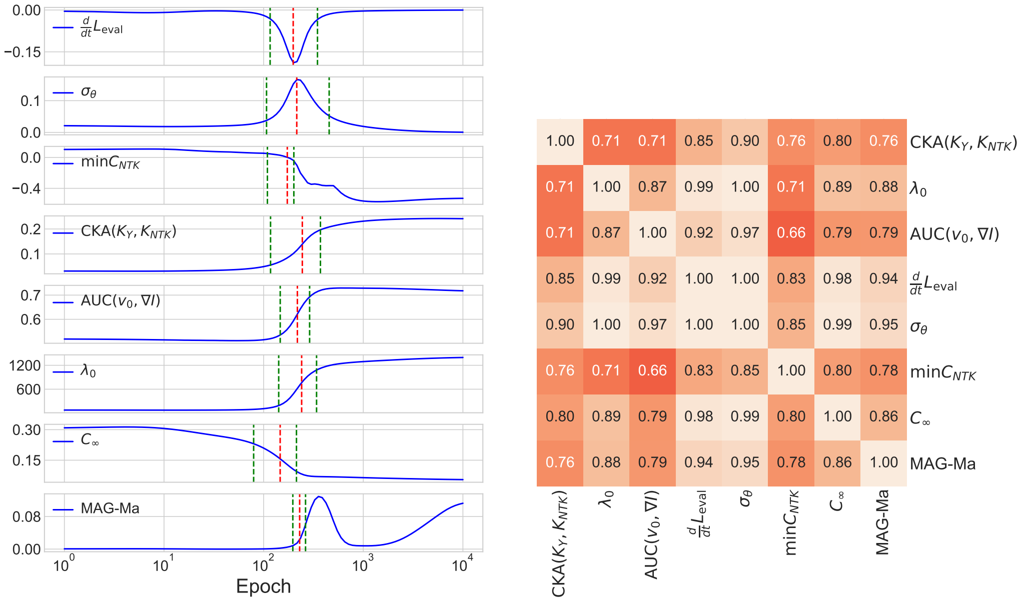
  $$\sigma_\theta^2 = \frac{1}{N}\text{Tr}(\text{diag}(r)^2 K_{NTK}) - \frac{1}{N^2} r^\top K_{NTK} r \tag{32}$$

  The critical point corresponds to the point where the variance reaches its maximum.
- **Centred Kernel Alignment (CKA)**: Empirically, as a DNN learns features that support the prediction of a target, its NTK begins to resemble the task kernel K_Y. For classification problems, if Y denotes a onehot encoding of the class labels, then K_Y is simply $YY^\top$. For INR regression, we opt to use:

  $$K_Y(x, x+u) = \exp\left(-\frac{||f(x) - f(x+u)||^2}{2\kappa^2}\right) \tag{33}$$

> Here, κ is a bandwidth parameter. We also compare the alignment of K_{NTK} with the RBF $K_X(x, x+u) = \exp(-||u||^2/\kappa)$. To monitor the similarity between kernels, we employ the centred, normalized Hilbert-Schmidt Information Criterion (HSIC) used in (23; 24; 25). See Appendix B.2 for additional details.

The left side of Figure 4 illustrates our procedure for identifying critical points on the astro dataset[2]. We use a simple peak detector to identify the region of interest for the loss rate $\dot{L}_{eval}$ and the gradient variance σ_θ, using the $FWHM$ to define a confidence region. For the $\min C_{NTK}$, we look for zero-crossings, with a confidence region constructed from the cumulative variance. For every other order parameter, we fit a sigmoid, where the inflection point marks the critical point, and the slope defines the confidence region (refer to Appendix B.2 for full details). The right side of Figure 4 demonstrates how frequently these confidence regions overlap across the different architectures and images studied[3]. Remarkably, the phase transitions described by the order parameters - despite being derived to measure different phenomenon in the literature - consistently occur at the same time during training.



	$\text{CKA}(K_Y, K_{NTK})$	λ_0	$\text{AUC}(v_0, \nabla I)$	$\frac{d}{dt}L_{\text{eval}}$	σ_θ	$\min C_{NTK}$	C_∞	MAG-Ma
$\text{CKA}(K_Y, K_{NTK})$	1.00	0.71	0.71	0.85	0.90	0.76	0.80	0.76
λ_0	0.71	1.00	0.87	0.99	1.00	0.71	0.89	0.88
$\text{AUC}(v_0, \nabla I)$	0.71	0.87	1.00	0.92	0.97	0.66	0.79	0.79
$\frac{d}{dt}L_{\text{eval}}$	0.85	0.99	0.92	1.00	1.00	0.83	0.98	0.94
σ_θ	0.90	1.00	0.97	1.00	1.00	0.85	0.99	0.95
$\min C_{NTK}$	0.76	0.71	0.66	0.83	0.85	1.00	0.80	0.78
C_∞	0.80	0.89	0.79	0.98	0.99	0.80	1.00	0.86
MAG-Ma	0.76	0.88	0.79	0.94	0.95	0.78	0.86	1.00

Figure 4: **Alignment of Order Parameters**. Left: Order parameter evolution and critical points during training of a SIREN model on the astro image. The red vertical lines denote the location of the critical points, and the green vertical lines denote confidence regions. Right: Heatmap showing the frequency of intersections between the confidence regions.

4.2 DYNAMICAL CONSEQUENCES OF HYPERPARAMETERS

In this section, we perform an ablation study to understand the impact of different hyperparameters on the phase transitions. The baseline model is a 5-layer 128-unit wide SIREN with $\omega_0 = 60$, which on average was the best performing model on the cameraman dataset[4]. We visualize the evolution of the $\min C_{NTK}$, the global correlation length ξ_{corr}, and the CKA between the NTK and an RBF Kernel K_X in Figures 6 and 5. Error bars are obtained by averaging the runs over five random seeds.

---

[2] Additional figures for other datasets may be found in Section G.1 of the Supplementary materials.

[3] In computing the coincidence matrix on the right side of Figure 4, we only included experimental runs in which our critical point detection succeeded for both pairs of order parameters. In Section C of the Supplementary Materials, we investigate the specific failure modes, and tie their failure rates (between 21% and 51%) to image properties

[4] Additional figures for other datasets may be found in Section G.2 of the supplementary materials.

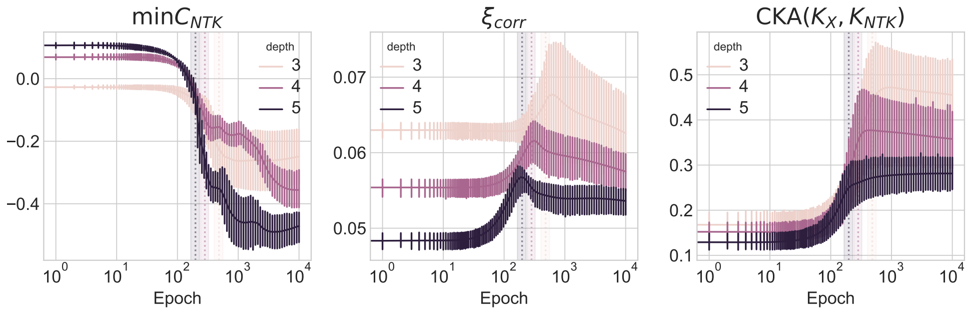


Figure 5: **Effect of depth on Critical Behaviour**: Average MSEs, in order of ascending depth: $7.742e^{-3} \pm 1.580e^{-4}$, $6.819e^{-3} \pm 2.696e^{-5}$, $6.571e^{-3} \pm 2.705e^{-5}$. Dashed vertical lines denote the location of the peak of the loss rate $\dot{L}_{eval}$, marking the phase transition.

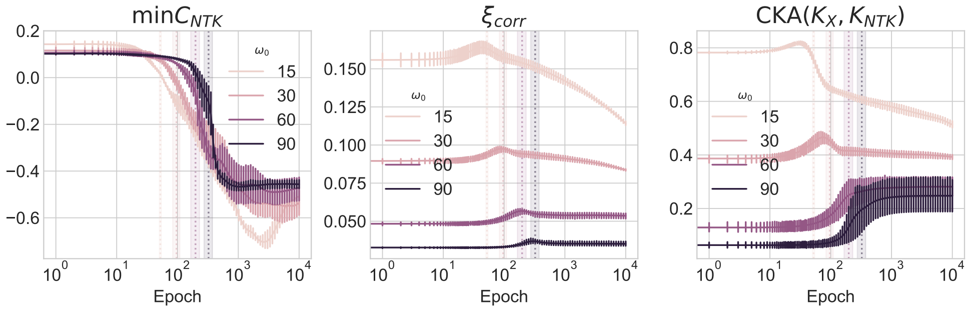


Figure 6: **Effect of ω_0 on Critical Behaviour**: Average MSEs, in order of ascending ω_0: $8.380e^{-3} \pm 9.191e^{-5}$, $7.234e^{-3} \pm 5.591e^{-5}$, $6.571e^{-3} \pm 2.705e^{-5}$, $7.853e^{-3} \pm 5.629e^{-4}$. Dashed vertical lines denote the location of the peak of the loss rate $\dot{L}_{\text{eval}}$, marking the phase transition.

When depth (and therefore model capacity) is decreased (Figure 5), we observe a corresponding increase in the validation error. For such models, the initial $\min C_{NTK}$ is lower, and consequently, learning is slower. Correspondingly, the peak in the loss rate $\dot{L}_{eval}$, occurs later. With increasing depth, there is less variance in order parameters across runs. Importantly, though the location of the phase transition changes, the trajectory shapes remain consistent. By contrast, we see more variance in the runs as we increase ω_0 (Figure 6) accompanying a decrease in validation error. The phase transition is also delayed until later in training.

Modifying ω_0 also leads to a dramatic change in the shape of the trajectories of the order parameters. When ω_0 is low, we see strong alignment between the NTK and a uniform RBF model. This indicates a strong preference for aggregating information from immediate neighbours. CKA(K_X, K_{NTK}) peaks at the critical point, where, accompanying a large decrease in the global correlation length-scale ξ_{corr}, it begins to rapidly decrease. In contrast, when ω_0 is high, the model begins with a very low CKAK_X, K_{NTK}, and rapidly grows, in sigmoidal fashion, at the critical point.

Taken together, our results suggest that hyperparameters such as depth and width are less important than ω_0 in terms of controlling inductive bias.

5 RELATED WORK

Neural Tangent Kernels for Implicit Neural Representations: Previous research has investigated the inductive biases of INRs using the Neural Tangent Kernel (NTK), focusing on aspects such as spectral properties (28) and dependencies on uniformly sampled data (29). Furthermore, studies by (30) and (31) have analyzed the eigenfunctions of the empirical NTK to elucidate the approximation capabilities of INRs. These investigations, however, primarily examine static properties of the NTK at initialization, which do not account for feature learning dynamics. In contrast, our work concentrates on the evolution of the NTK, aiming to deepen our understanding of how INRs learn to model images.

Neural Tangent Kernel Alignment In practical settings, recent studies have shown that during training, the NTK dynamically aligns with a limited number of task-relevant directions (32; 33; 22; 34; 23; 35; 24; 25). Specifically in classification tasks, this alignment results in a block structure within the kernel matrix, where correlations between samples from the same class are notably stronger than those between different classes (35). Concurrently, at the eigenfunction level, the modes increasingly reflect salient features of the dataset, such as class-separating boundaries (22; 23). The widespread occurrence and influence of kernel alignment suggest its critical role in DNN feature learning, contributing to the superior performance of DNNs over models based on infinite-width NTKs (24).

That said, these theoretical discussions often focus on shallow networks (34; 35), toy models (24; 23), and deep linear networks (35). Empirical studies of more complex models primarily analyze centered-kernel alignment and cumulative power (23; 24; 25). In this work, we extend this research to encompass INRs. INRs exhibit the full complexity of DNNs, though their low-dimensional input space makes certain analyses more tractable, and we may leverage expert knowledge from computer vision. Additionally, we adapt metrics from the theory of DNN phase transitions to analyze the conditions, timing, and mechanisms of NTK alignment.

Fast and Slow Phases of Neural Network Training The literature highlights a critical dynamical phase transition in DNN training, marked by a shift from fast to slow learning regimes (11). In the initial Fast Phase where gradient norms are significantly larger than their individual fluctuations, global consensus amongst the datapoints leads to rapid loss reduction. In the subsequent Slow Phase, fluctuations dominate, leading to slow learning. This transition may be quantified by a number of order parameters, such as changes in the signal-to-noise ratios of the gradient norms (11), gradient confusion, (12; 13; 14), and changes in the correlation lengthscale (12).

The transition from fast to slow learning not only delineates changes in learning dynamics, but also aligns with the model's progression from learning easy patterns to memorizing complex ones (4). This shift represents a collision between the dataset and the model's inductive biases, and thus, presents an avenue to understanding feature learning. (36) is similarly motivated to study inductive biases in terms of the dynamics of representations in ReLu networks. In this work, we focus on insights obtained from the dynamics of the NTK.

6 CONCLUSION

We have conducted preliminary investigations into the dynamics of feature learning within INRs for image data. Specifically, we demonstrated that SIREN models typically exhibit pronounced spatial variations in the parameter gradients $||\nabla_\theta f||$. This variation facilitates the alignment of the local structure of the NTK with the edges in the images being modelled. Notably, this alignment predominantly occurs during a critical phase transition, characterized by increased spatial variation and translational symmetry breaking. This phase transition aligns with a shift from the fast to a slow learning phases. By approximating the local structure of the NTK with a Cauchy distribution, we were able to relate various order parameters associated with this dynamic phase transition, such as the correlation length, and gradient confusion.

Overall, many promising lines of research remain open. In this work, our focus has been primarily on SIREN models trained using full-batch gradient descent. Future works may investigate whether our observations hold true across different architectures and optimizers. For example, the ADAM optimizer (37), which adaptively adjusts the learning rate throughout the optimization process, could potentially influence the stability and divergence behaviours of the principal eigenvalue - a key element in our study of NTK alignment. Moreover, one could employ a static positional encoding, such as a random Fourier embedding (29), which may prevent the accumulation of spatial gradients.

This work has demonstrated how the NTK provides a rich theoretical tool for deriving and relating order parameters to understand training dynamics. Our approach provides new methodology to rigorously study the influence of inductive biases, such as model architectures and hyper-parameter values, on the underlying learning process and may have practical utility in diagnosing the cause of poor learning outcomes.

REFERENCES

[1] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*, 2020.

[2] Boris Hanin and David Rolnick. Deep relu networks have surprisingly few activation patterns. In *Neural Information Processing Systems*, 2019.

[3] Maxwell Nye and Andrew M. Saxe. Are efficient deep representations learnable? *ArXiv*, abs/1807.06399, 2018.

[4] Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. (arXiv:1706.05394), 2017.

[5] Xiao Zhang and Dongrui Wu. Rethink the connections among generalization, memorization, and the spectral bias of dnns. In *International Joint Conference on Artificial Intelligence*, 2020.

[6] Yu Feng and Yuhai Tu. Phases of learning dynamics in artificial neural networks in the absence or presence of mislabeled data. *Machine Learning: Science and Technology*, 2(4):043001, jul 2021.

[7] Cory Stephenson and Tyler Lee. When and how epochwise double descent happens. *ArXiv*, abs/2108.12006, 2021.

[8] Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. *ArXiv*, abs/2205.10343, 2022.

[9] Liu Ziyin and Masakuni Ueda. Exact phase transitions in deep learning. *ArXiv*, abs/2205.12510, 2022.

[10] James P. Sethna. *Statistical mechanics: Entropy, order parameters, and complexity*. Oxford University Press, 2021.

[11] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. (arXiv:1703.00810), 2017.

[12] Stanislav Fort, Paweł Krzysztof Nowak, Stanislaw Jastrzebski, and Srini Narayanan. Stiffness: A new perspective on generalization in neural networks. (arXiv:1901.09491), 2020.

[13] Karthik A. Sankararaman, Soham De, Zheng Xu, W. Ronny Huang, and Tom Goldstein. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. (arXiv:1904.06963), 2020.

[14] Yu Feng and Yuhai Tu. Phases of learning dynamics in artificial neural networks in the absence or presence of mislabeled data. 2(4):043001, 2021. Publisher: IOP Publishing.

[15] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *CoRR*, abs/1806.07572, 2018.

[16] Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. (arXiv:1909.05989), 2019.

[17] Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. (arXiv:1909.08156), 2019.

[18] Laurence Aitchison. Why bigger is not always better: on finite and infinite neural networks. (arXiv:1910.08013), 2020.

[19] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. (arXiv:1812.07956), 2020.

[20] Jaehoon Lee, Samuel S. Schoenholz, Jeffrey Pennington, Ben Adlam, Lechao Xiao, Roman Novak, and Jascha Sohl-Dickstein. Finite versus infinite neural networks: an empirical study. (arXiv:2007.15801), 2020.

[21] Mariia Seleznova and Gitta Kutyniok. Neural tangent kernel beyond the infinite-width limit: Effects of depth and initialization. (arXiv:2202.00553), 2022.

[22] Dmitry Kopitkov and Vadim Indelman. Neural spectrum alignment: Empirical study. (arXiv:1910.08720), 2020.

[23] Aristide Baratin, Thomas George, César Laurent, R. Devon Hjelm, Guillaume Lajoie, Pascal Vincent, and Simon Lacoste-Julien. Implicit regularization via neural feature alignment. (arXiv:2008.00938), 2021.

[24] Haozhe Shan and Blake Bordelon. A theory of neural tangent kernel alignment and its influence on training. (arXiv:2105.14301), 2022.

[25] Abdulkadir Canatar and Cengiz Pehlevan. A kernel analysis of feature learning in deep neural networks. In *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–8. IEEE.

[26] Joseph W Goodman. Statistical optics. *New York, Wiley-Interscience, 1985, 567 p.*, 1, 1985.

[27] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986.

[28] Zhemin Li, Hongxia Wang, and Deyu Meng. Regularize implicit neural representation by itself. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10280–10288. IEEE.

[29] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. (arXiv:2006.10739), 2020.

[30] Gizem Yüce, Guillermo Ortiz-Jiménez, Beril Besbinar, and Pascal Frossard. A structured dictionary perspective on implicit neural representations. (arXiv:2112.01917), 2022.

[31] Vishwanath Saragadam, Daniel LeJeune, Jasper Tan, Guha Balakrishnan, Ashok Veeraraghavan, and Richard G. Baraniuk. WIRE: Wavelet implicit neural representations. (arXiv:2301.05187), 2023.

[32] Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M. Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. (arXiv:2010.15110), 2020.

[33] Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. 2020(11):113301.

[34] Jonas Paccolat, Leonardo Petrini, Mario Geiger, Kevin Tyloo, and Matthieu Wyart. Geometric compression of invariant manifolds in neural nets. 2021(4):044001.

[35] Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. (arXiv:2111.00034), 2021.

[36] John Lazzari and Xiuwen Liu. Understanding the spectral bias of coordinate based MLPs via training dynamics. (arXiv:2301.05816), 2023.

[37] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[38] Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[39] 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.

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[41] Richard Zou Horace He. functorch: Jax-like composable function transforms for pytorch. 2021.

[42] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[43] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[44] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020.

[45] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the Alvey Vision Conference*, pages 23.1–23.6. Alvety Vision Club, 1988. doi:10.5244/C.2.23.