Appendix A

In addition to the TNS results on the self-attention mechanism given in the main text, we additionally provide the TNS with results for other training methods that can improve the model performance. From Fig.1, we again verify that the high-performance models have a strong representational ability to measure the stiffness phenomenon, which is consistent with the results shown in our main text.

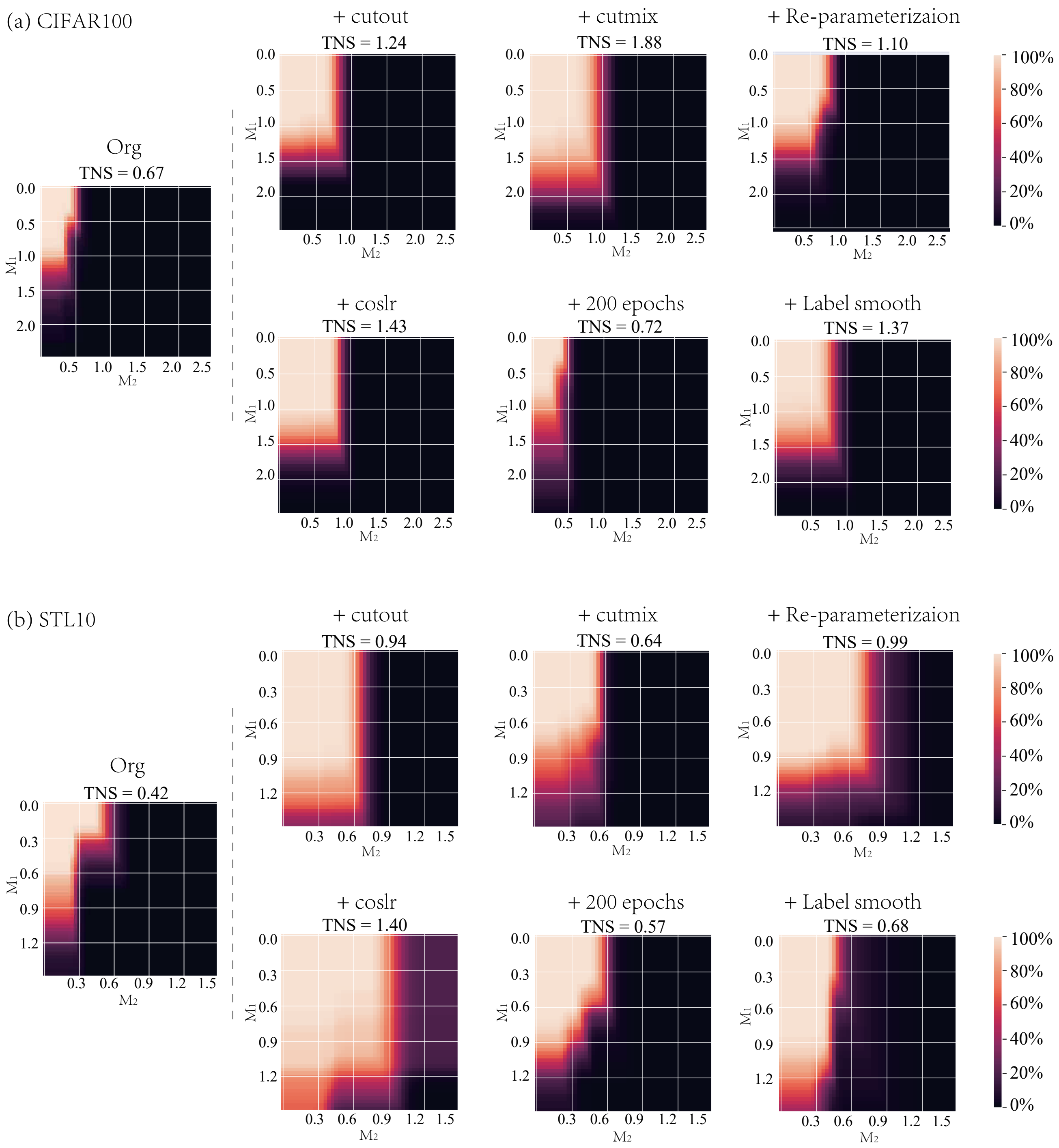

Figure 1. The TNS results for other high-performance training tricks.

Appendix B

Lemma 1. Given the feature trajectories $x_1, x_2, x_3..., x_L$ generated by a neural network with L residual blocks, *i.e.*, $x_{t+1} = x_t + f(x_t;\theta_t)\cdot\Delta_t, t = 0, 1, .., L-1$, where the norm $\|x_t\|_2$ and step size Δt are bounded. For $\delta(\mathbf{M})$ defined as $\delta(\mathbf{M}) = \mathbb{E}_{x_0\sim P(x_0)}\mathbf{I}_{\exists t, s.t. \zeta_{\text{NSI}}(x_t)\geq\max(\mu(1+M_1),M_2)}$, $\exists\tilde{M}\in\mathbb{R}+$, s.t. if $\min(M_1, M_2) > \tilde{M}$, $\delta(\mathbf{M}) = 0$.

Proof. Let $0 < k_1 \leq \|x_t\|_2 \leq k_2, t = 1, 2, ..., L-1$ and $\Delta_t \in [a, b]$, where k_1, k_2, a and $b \in \mathbb{R}_+$. In fact this condition is practical and mild in neural networks. Next, we prove that $\exists\tilde{M}\in\mathbb{R}_+$ s.t. when $M_1 > \tilde{M}$ and $M_2 > \tilde{M}$, we have (1) $\zeta_{\text{NSI}}(x_t) < M_2$ and (2) $\zeta_{\text{NSI}}(x_t) < \mu(1+M_1)$. For (1) $\zeta_{\text{NSI}}(x_t) < M_2$, since the boundary and triangles inequality,

$$\begin{aligned}\zeta_{\text{NSI}}(x_t) &= \frac{1}{\|x_t\|_2}\|\frac{x_{t+1}-x_t}{\Delta_t}\|_2 \\ &= \frac{1}{\Delta_t}\frac{1}{\|x_t\|_2}\|x_{t+1}-x_t\|_2 \\ &\leq \frac{1}{a}\cdot\frac{1}{\|x_t\|_2}\cdot(\|x_{t+1}\|_2 + \|x_t\|_2) = \frac{1}{a}(1+\frac{\|x_{t+1}\|_2}{\|x_t\|_2}) \leq \frac{1}{a}(1+\frac{k_2}{k_1})\end{aligned}\tag{1}$$

Therefore, if $M_2 > \frac{1}{a}(1+\frac{k_2}{k_1})$, we have $\zeta_{\text{NSI}}(x_t) < M_2$. For (2) $\zeta_{\text{NSI}}(x_t) < \mu(1+M_1)$, we can first estimate the lower bound of $\zeta_{\text{NSI}}(x_t)$.

$$\begin{aligned}\zeta_{\text{NSI}}(x_t) &= \frac{1}{\|x_t\|_2}\|\frac{x_{t+1}-x_t}{\Delta_t}\|_2 \\ &= \frac{1}{\Delta_t}\frac{1}{\|x_t\|_2}\|x_{t+1}-x_t\|_2 \\ &\geq \frac{1}{b}\cdot\frac{1}{\|x_t\|_2}\cdot|\|x_{t+1}\|_2 - \|x_t\|_2| = \frac{1}{b}\cdot\frac{1}{k_2}\cdot|k_1-k_2|.\end{aligned}\tag{2}$$

Moreover, note that

$$\begin{aligned}\zeta_{\text{NSI}}(x_t) < \mu(1+M_1) &\Leftrightarrow \zeta_{\text{NSI}}(x_t) < \frac{1}{L}\sum_{i=1}^{L}\zeta_{\text{NSI}}(x_i)(1+M_1) \\ &\Leftrightarrow L\cdot\zeta_{\text{NSI}}(x_t)/\sum_{i=1}^{L}\zeta_{\text{NSI}}(x_i) - 1 < M_1\end{aligned}\tag{3}$$

From Eq.(1), we have $L\cdot\zeta_{\text{NSI}}(x_t) \leq \frac{L}{a}(1+\frac{k_2}{k_1})$. Furthermore, from Eq.(2), we have

$$\frac{1}{\sum_{i=1}^{L}\zeta_{\text{NSI}}(x_i)} \leq \frac{1}{\sum_{i=1}^{L}\frac{1}{b}\cdot\frac{1}{k_2}\cdot|k_1-k_2|} = \frac{bk_2}{L|k_1-k_2|}.\tag{4}$$

Therefore, when M_1 meets

$$M_1 > \frac{L}{a}(1+\frac{k_2}{k_1})\cdot\frac{bk_2}{L|k_1-k_2|} - 1 = \frac{bk_2(k_1+k_2)}{ak_1\cdot|k_1-k_2|} - 1,\tag{5}$$

we have $M_1 > L\cdot\zeta_{\text{NSI}}(x_t)/\sum_{i=1}^{L}\zeta_{\text{NSI}}(x_i) - 1$ and $\zeta_{\text{NSI}}(x_t) < \mu(1+M_1)$ holds. Let

$$\tilde{M} = \max\big(\frac{bk_2(k_1+k_2)}{ak_1\cdot|k_1-k_2|} - 1, \frac{1}{a}(1+\frac{k_2}{k_1})\big).\tag{6}$$

When $\min(M_1, M_2) > \tilde{M}$, for any t,

$$\zeta_{\text{NSI}}(x_t) < \max(M_2, \mu(1+M_1)),\tag{7}$$

Therefore,

$$\delta(\mathbf{M}) = \mathbb{E}_{x_0\sim P(x_0)}\underbrace{\mathbf{I}_{\exists t, s.t. \zeta_{\text{NSI}}(x_t)\geq\max(\mu(1+M_1),M_2)}}_{\text{equal to 0}} = 0.\tag{8}$$

□

Theorem 1. For the $\delta(\mathbf{M})$ defined as $\delta(\mathbf{M}) = \mathbb{E}_{x_0\sim P(x_0)}\mathbf{I}_{\exists t,s.t.\zeta_{\text{NSI}}(x_t)\geq\max(\mu(1+M_1),M_2)}$, the Total Neural Stiffness (TNS) $\iint_{\mathbf{M}}\delta(\mathbf{M})d\mathbf{M}$ is convergent.

Proof. Let

$$\tilde{M} = \max\big(\frac{bk_2(k_1+k_2)}{ak_1\cdot|k_1-k_2|} - 1, \frac{1}{a}(1+\frac{k_2}{k_1})\big). \tag{9}$$

Note that $0\leq M_1, M_2$,

$$\begin{aligned}\iint_{\mathbf{M}}\delta(\mathbf{M})d\mathbf{M} &= \int_0^{+\infty}\int_0^{+\infty}\delta(M_1,M_2)dM_1dM_2\\ &= \int_0^{\tilde{M}}\int_0^{\tilde{M}}\delta(M_1,M_2)dM_1dM_2 + \underbrace{\int_{\tilde{M}}^{+\infty}\int_{\tilde{M}}^{+\infty}\delta(M_1,M_2)dM_1dM_2}_{\text{equal to 0 since Lemma 1}}.\end{aligned} \tag{10}$$

Since $0\leq\mathbf{I}_{\exists t,s.t.\zeta_{\text{NSI}}(x_t)\geq\max(\mu(1+M_1),M_2)}\leq 1$, we have

$$0\leq\delta(\mathbf{M}) = \mathbb{E}_{x_0\sim P(x_0)}\mathbf{I}_{\exists t,s.t.\zeta_{\text{NSI}}(x_t)\geq\max(\mu(1+M_1),M_2)}\leq 1. \tag{11}$$

Therefore,

$$\begin{aligned}\iint_{\mathbf{M}}\delta(\mathbf{M})d\mathbf{M} &= \int_0^{\tilde{M}}\int_0^{\tilde{M}}\delta(M_1,M_2)dM_1dM_2 + 0 && \text{from Eq. (10)}\\ &\leq \int_0^{\tilde{M}}\int_0^{\tilde{M}}dM_1dM_2 < +\infty && \text{from Eq. (11)}\end{aligned}$$

The $\delta(\mathbf{M})$ is positive and $\iint_{\mathbf{M}}\delta(\mathbf{M})d\mathbf{M}$ is bounded, thus the Total Neural Stiffness $\iint_{\mathbf{M}}\delta(\mathbf{M})d\mathbf{M}$ is convergent.

□

Appendix C

Theorem 2. For an ordinary differential equation $\mathrm{d}\mathbf{u}/\mathrm{d}t = \mathbf{f}(\mathbf{u})$, if the Jacobian matrix $\mathbf{J}_{u^t}$ at u^t is a $n \times n$ symmetric real matrix and $\{\lambda_i\}_{i=1}^n$ are its n distinct eigenvalues, and $\mathbf{Re}(\lambda_i) < 0, i = 1, 2, ..., n$, then

$$\zeta_{\text{SAI}}(u^t) \approx \zeta_{\text{SI}}(u^t) \cdot \sqrt{c + Q[\zeta_{\text{SI}}(u^t)]}, \tag{12}$$

where c is a constant and $Q(\cdot)$ is a function with respect to $\zeta_{\text{SI}}(u^t)$ and when $\zeta_{\text{SI}}(u^t)$ is large enough, $Q[\zeta_{\text{SI}}(u^t)]$ converges to a 0.

Proof. Note that $\zeta_{\text{SAI}}(\cdot)$ is computed by adjacent states with small step size, and the adjacent states are closed to a linearized ODE. Therefore, we use Taylor expansion to provide a reasonable approximation for the right-hand side of the equation. Specifically, we consider the Taylor expansion at u^t for $\mathbf{f}(\mathbf{u})$, we have

$$\begin{aligned} \mathbf{f}(\mathbf{u}) &= \mathbf{f}(u^t) + \mathbf{J}_{u^t}(\mathbf{u} - u^t) + o(\|\mathbf{u} - u^t\|) \\ &= \mathbf{J}_{u^t}\mathbf{u} + \mathbf{f}(u^t) - \mathbf{J}_{u^t}u^t + o(\|\mathbf{u} - u^t\|) \approx \mathbf{J}_{u^t}\mathbf{u} + \mathbf{h}(t). \end{aligned} \tag{13}$$

Let $\{\mathbf{v}_i\}_{i=1}^n$ be the eigenvectors corresponding to the eigenvalues $\{\lambda_i\}_{i=1}^n$. Since the Jacobian matrix $\mathbf{J}_{u^t}$ at u^t is a $n \times n$ symmetric real matrix, $\{\mathbf{v}_i\}_{i=1}^n$ form a set of orthogonal vectors. Without loss of generalization, we assume $\{\mathbf{v}_i\}_{i=1}^n$ is standard orthogonal basis. Therefore, u^t can be represented by the basis $\{\mathbf{v}_i\}_{i=1}^n$. Let

$$\mathbf{u}(0) = u^t = \sum_{i=1}^n c_i\mathbf{v}_i, \tag{14}$$

where $c_i \in \mathbb{R}$. Moreover, Eq.(13) is a linear constant coefficient inhomogeneous equation. The solution of this equation is

$$\mathbf{u}(t) = \sum_{i=1}^n c_i\mathbf{v}_i e^{\lambda_i t} + \mathbf{g}(t), \tag{15}$$

where $\mathbf{g}(t)$ is steady-state solution and since Eq.(14), $\mathbf{g}(0) = \mathbf{0}$. Let $u^{t'} = \mathbf{u}(\Delta t)$, we have

$$\begin{aligned} \lim_{\Delta t \to 0} \frac{1}{\|u^t\|_2} \Big\| \frac{u^{t'} - u^t}{\Delta t} \Big\|_2. &= \lim_{\Delta t \to 0} \frac{1}{\|u^t\|_2} \Big\| \frac{\sum_{i=1}^n c_i\mathbf{v}_i e^{\lambda_i \Delta t} + \mathbf{g}(\Delta t) - \sum_{i=1}^n c_i\mathbf{v}_i}{\Delta t} \Big\|_2 \\ &= \lim_{\Delta t \to 0} \frac{1}{\|u^t\|_2} \Big\| \frac{\sum_{i=1}^n c_i\mathbf{v}_i(e^{\lambda_i \Delta t} - 1)}{\Delta t} + \frac{\mathbf{g}(\Delta t) - \mathbf{g}(0)}{\Delta t} \Big\|_2 \\ &= \frac{1}{\|u^t\|_2} \Big\| \sum_{i=1}^n \lambda_i c_i \mathbf{v}_i + \nabla\mathbf{g}(0) \Big\|_2 && \text{Since } e^x - 1 \sim x \end{aligned}$$

And $\nabla\mathbf{g}(0)$ can be linear combination by the standard orthogonal basis $\{\mathbf{v}_i\}_{i=1}^n$. Let $\nabla\mathbf{g}(0) = \sum_{i=1}^n a_i\mathbf{v}_i$, where $a_i \in R$. Note that

$$\|u^t\|_2 = \| \sum_{i=1}^n c_i\mathbf{v}_i \|_2 = (\sum_{i=1}^n c_i^2)^{1/2}, \tag{16}$$

we can find that

$$\begin{aligned} \lim_{\Delta t \to 0} \frac{1}{\|u^t\|_2} \Big\| \frac{u^{t'} - u^t}{\Delta t} \Big\|_2. &= \frac{1}{\|u^t\|_2} \Big\| \sum_{i=1}^n \lambda_i c_i \mathbf{v}_i + \nabla\mathbf{g}(0) \Big\|_2 \\ &= \frac{1}{\|u^t\|_2} \Big\| \sum_{i=1}^n (\lambda_i c_i + a_i)\mathbf{v}_i \Big\|_2 && \text{Since } \nabla\mathbf{g}(0) = \textstyle\sum_{i=1}^n a_i\mathbf{v}_i \\ &= \Big[\frac{\sum_{i=1}^n (\lambda_i c_i + a_i)^2}{\sum_{i=1}^n c_i^2} \Big]^{\frac{1}{2}}. && \text{Since Eq.(16)} \end{aligned}$$

Without loss of generalization, we assume $|\mathbf{Re}(\lambda_1)| \geq |\mathbf{Re}(\lambda_2)| \geq ... \geq |\mathbf{Re}(\lambda_n)|$. Moreover, since the matrix J_{u^t} is symmetric real matrix, the eigenvalues are real number, *i.e.*, $\mathbf{Re}(\lambda_i) = \lambda_i, i = 1, 2, ..., n$. Therefore,

$$\zeta_{\text{SI}}(u^t) = \max(|\mathbf{Re}(\lambda_i)|) = |\lambda_1|. \tag{17}$$

Moreover,

$$\begin{aligned}
\lim_{\Delta t \to 0} \frac{1}{\|u^t\|_2} \Big\| \frac{u^{t'} - u^t}{\Delta t} \Big\|_2 &= \Big[\frac{\sum_{i=1}^n (\lambda_i c_i + a_i)^2}{\sum_{i=1}^n c_i^2}\Big]^{\frac{1}{2}}. \\
&= \Big[\sum_{i=1}^n \Big(\frac{c_i^2}{\sum_{i=1}^n c_i^2}\lambda_i^2 + \frac{2a_i}{\sum_{i=1}^n c_i^2}\lambda_i + \frac{a_i^2}{\sum_{i=1}^n c_i^2}\Big)\Big]^{\frac{1}{2}} \\
&= \Big[\frac{c_1^2}{\sum_{i=1}^n c_i^2}\lambda_1^2 + \sum_{i=2}^n \frac{c_i^2}{\sum_{i=1}^n c_i^2}\lambda_i^2 + \sum_{i=1}^n \Big(\frac{2a_i}{\sum_{i=1}^n c_i^2}\lambda_i + \frac{a_i^2}{\sum_{i=1}^n c_i^2}\Big)\Big]^{\frac{1}{2}} \\
&= \zeta_{\text{SI}}(u^t)\Big[\underbrace{\frac{c_1^2}{\sum_{i=1}^n c_i^2}}_{\text{Constant}} + \underbrace{\sum_{i=2}^n \frac{c_i^2}{\sum_{i=1}^n c_i^2}\frac{\lambda_i^2}{\lambda_1^2} + \sum_{i=1}^n \Big(\frac{2a_i}{\sum_{i=1}^n c_i^2}\frac{\lambda_i}{\lambda_1^2} + \frac{a_i^2}{\lambda_1^2 \sum_{i=1}^n c_i^2}\Big)}_{\text{The term with respect to SI}}\Big]^{\frac{1}{2}} \qquad \text{Since Eq.(17)} \\
&\triangleq \zeta_{\text{SI}}(u^t) \cdot \sqrt{(c + Q[\zeta_{\text{SI}}(u^t)])}
\end{aligned}$$

Thus, $\zeta_{\text{SAI}}(u^t) = \zeta_{\text{SI}}(u^t) \cdot \sqrt{c + Q[\zeta_{\text{SI}}(u^t)]} + \mathcal{O}(\Delta t)$ and according to Eq.(13), $\zeta_{\text{SAI}}(u^t) \approx \zeta_{\text{SI}}(u^t) \cdot \sqrt{(c + Q[\zeta_{\text{SI}}(u^t)])}$ for ordinary differential equation $\mathrm{d}\mathbf{u}/\mathrm{d}t = \mathbf{f}(\mathbf{u})$ holds. Next, since Eq.(17),

$$\lim_{\zeta_{\text{SI}}(u^t) \to +\infty} Q[\zeta_{\text{SI}}(u^t)] = \lim_{|\lambda_1| \to +\infty} \sum_{i=2}^n \frac{c_i^2}{\sum_{i=1}^n c_i^2}\frac{\lambda_i^2}{\lambda_1^2} + \sum_{i=1}^n \Big(\frac{2a_i}{\sum_{i=1}^n c_i^2}\frac{\lambda_i}{\lambda_1^2} + \frac{a_i^2}{\lambda_1^2 \sum_{i=1}^n c_i^2}\Big) = 0. \tag{18}$$

In Eq.(13), if we consider another linear approximation like $\mathbf{f}(\mathbf{u}) \approx \mathbf{J}_{u^t}\mathbf{u}$, we can also obtain the similar conclusion as $\zeta_{\text{SAI}}(u^t) \approx \zeta_{\text{SI}}(u^t) \cdot \sqrt{c + U[\zeta_{\text{SI}}(u^t)]}$, where $U(\cdot)$ is a function with respect to $\zeta_{\text{SI}}(u^t)$ and when $\zeta_{\text{SI}}(u^t)$ is large enough, $U[\zeta_{\text{SI}}(u^t)]$ converges to a 0. Specifically,

$$\begin{aligned}
\lim_{\Delta t \to 0} \frac{1}{\|u^t\|_2} \Big\| \frac{u^{t'} - u^t}{\Delta t} \Big\|_2 &= \frac{1}{\|u^t\|_2} \Big\| \sum_{i=1}^n \lambda_i c_i \mathbf{v}_i \Big\|_2 \\
&= \Big[\sum_{i=1}^n \frac{c_i^2}{\sum_{i=1}^n c_i^2}\lambda_i\Big]^{\frac{1}{2}} \\
&= \Big[\frac{c_1^2}{\sum_{i=1}^n c_i^2}\lambda_1^2 + \sum_{i=2}^n \frac{c_i^2}{\sum_{i=1}^n c_i^2}\lambda_i^2\Big]^{\frac{1}{2}} \\
&= \zeta_{\text{SI}}(u^t)\Big[\underbrace{\frac{c_1^2}{\sum_{i=1}^n c_i^2}}_{\text{Constant}} + \underbrace{\sum_{i=2}^n \frac{c_i^2}{\sum_{i=1}^n c_i^2}\frac{\lambda_i^2}{\lambda_1^2}}_{\text{The term with respect to SI}}\Big]^{\frac{1}{2}} \qquad \text{Since Eq.(17)} \\
&\triangleq \zeta_{\text{SI}}(u^t) \cdot \sqrt{(c + U[\zeta_{\text{SI}}(u^t)])}
\end{aligned}$$

Moreover,

$$\lim_{\zeta_{\text{SI}}(u^t) \to +\infty} U[\zeta_{\text{SI}}(u^t)] = \lim_{|\lambda_1| \to +\infty} \sum_{i=2}^n \frac{c_i^2}{\sum_{i=1}^n c_i^2}\frac{\lambda_i^2}{\lambda_1^2} = 0. \tag{19}$$

□

Appendix D:

Since the attention values are generally less than 1 (the attention values are usually measured by the Sigmoid function or Softmax function), they are more fine-grained than the step size $\Delta t = 1$ of the original residual neural networks. In various backbones and their different stages, these small step sizes are used in different ways. In Fig.2, we show an example to understand how the self-attention module uses these small step sizes. Specifically, we take three blocks in the first stage of SENet as an example, we can find that the NSI and attention values are negatively correlated. In other words, in this case, if the stiffness issues exist in the blocks, the module tends to generate a smaller step size to alleviate the stiff issues, which is consistent with the discussions about the stiffness of ODEs in Section 2.1.2.

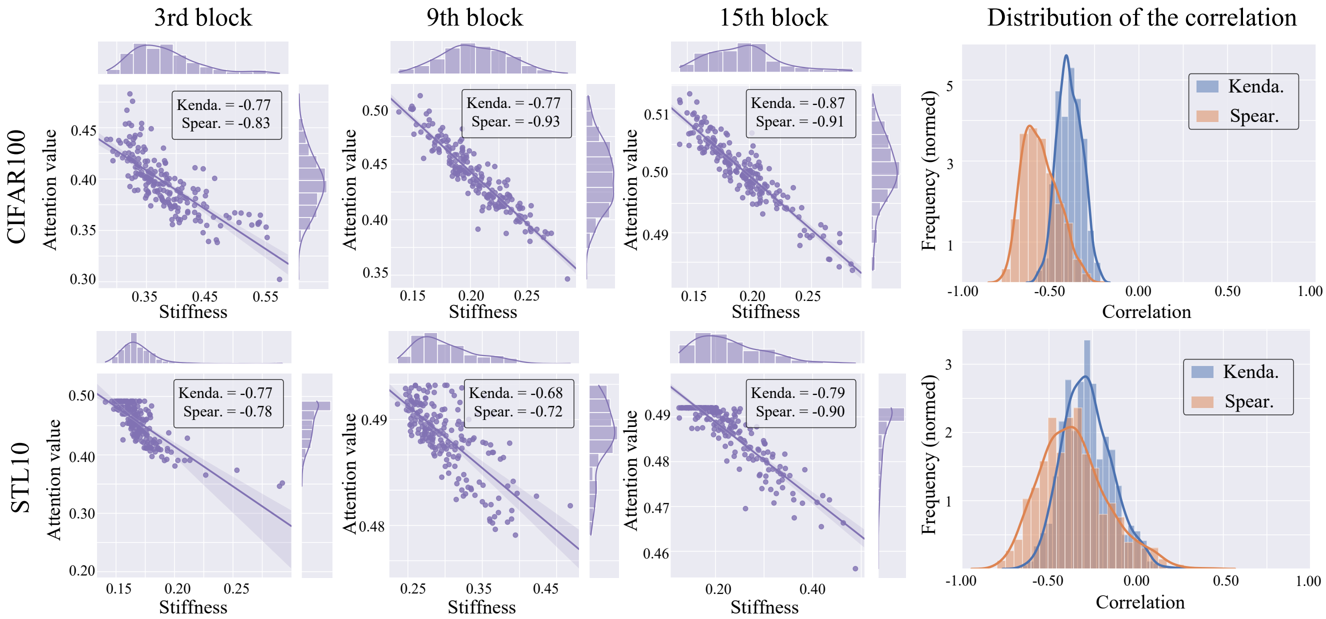

Figure 2. The relationship between the stiffness and attention values (step size). We take the feature trajectories from the first stage of SENet164 as an example. **a,** the correlations on three specific blocks. **b,** the distribution of correlations for all trajectories.

Appendix E:

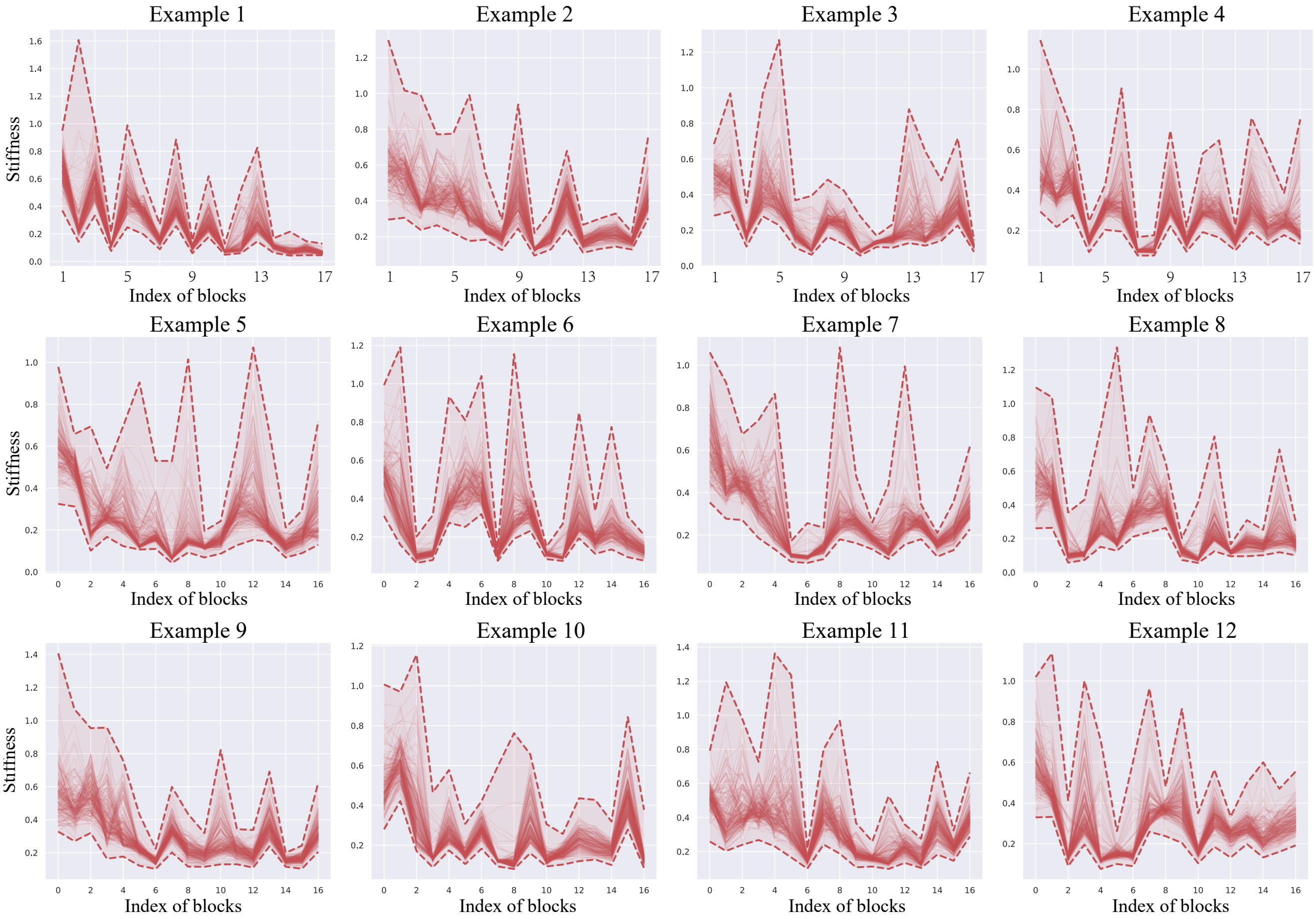

Figure 3. The visualization of NSI in SENet164 (first stage) on CIFAR100 with different random seeds.

Appendix F:

We first introduce the structure of Adaptor in proposed StepNet in main paper. σ is Sigmoid activation function and "Pooling" is global average pooling. "Conv" is group convolution with kernel size $k = 1$. "IEBN" [5] is the combination of batch normalization and a linear transformation "IE" from [3]. Specifically, for input $x \in R^n$, "IE" can be written as $\mathbf{W}_{\text{IE}} \otimes x + \mathbf{b}_{\text{IE}}$, where $\mathbf{W}_{\text{IE}} \in R^n$ and $\mathbf{b}_{\text{IE}} \in R^n$ are learnable parameters. The elements in $\mathbf{W}_{\text{IE}}$ and $\mathbf{b}_{\text{IE}}$ can be initialized as 0.0 and -1.0, respectively. All experiments in this paper are verified 5 times with random seeds on RTX 3080 GPUs. We will release our source codes after peer review.

Dataset	**#class**	**#training**	**#test**	**#Image size**
CIFAR10	10	50,000	10,000	32 x 32
CIFAR100	100	50,000	10,000	32 x 32
STL-10	10	5,000	8,000	96 x 96
ImageNet	1000	1,281,123	50,000	224 x 224

Table 1. The summary of the datasets for image classification experiments.

For image classification experiments, the details of the datasets are shown in Table 1. The hyper-parameter settings of CIFAR and ImageNet are shown in Table 2 respectively. For object detection tasks, we consider MS COCO dataset on the same setting as [9]. We use Faster R-CNN as detectors, which are implemented by the open-source MMDetection toolkit. The MS COCO dataset contains 80 classes with 118,287 training images and 40,670 test images. "AP", "AP$_S$", "AP$_M$", and "AP$_L$" are averaged AP for overall, small, medium, and large scale objects, respectively, at [50%, 95%] IoU interval with step as 5%, "AP$_{50}$" and "AP$_{75}$": AP at IoU as 50% and 75%, respectively.

	ResNet34	ResNet50	ResNet164
Batch size	256	256	128
Epoch	120	120	164
Optimizer	SGD(0.9)	SGD(0.9)	SGD(0.9)
depth	34	50	164
schedule	30/60/90	30/60/90	81/122
wd	1.00E-04	1.00E-04	1.00E-04
gamma	0.1	0.1	0.1
lr	0.1	0.1	0.1
Pooling	GAP	GAP	GAP

Table 2. Implementation detail for ImageNet 2012/CIFAR10/CIFAR100 image classification. Normalization and standard data augmentation (random cropping and horizontal flipping) are applied to the training data. For ImageNet dataset, the random cropping of size 224 by 224 is used in corresponding experiments. For CIFAR datasets, we use 32 by 32 size for the random cropping.

Appendix G:

The transformer-based self-attention neural networks are a recently popular residual neural network structure in various artificial intelligence fields. We try to give a preliminary analysis of transformer-based self-attention neural networks by the idea in our main paper. Although there are many variants of this kind of model, we consider the simplest structure shown in Fig.4a, *i.e*,

$$\hat{x}_{t+1} = x_t + \underbrace{f_3(x_t;\theta_{3t})}_{\text{Feature map}} \otimes \underbrace{\mathbf{F}(f_1(x_t;\theta_{1t}), f_2(x_t;\theta_{2t});\phi_t)}_{\text{Attention value}}, \tag{20}$$

where σ usually is Softmax activation function, ϕ_t is the learnable parameter of the self-attention module, $\mathbf{F}(f_1(x_t;\theta_{1t}), f_2(x_t;\theta_{2t});\phi_t)$ is the attention value. From Eq.(20) and the analysis in main paper, the transformer-based self-attention can also be regarded as the adaptive step size. However, is this kind of step size also stiffness-aware?

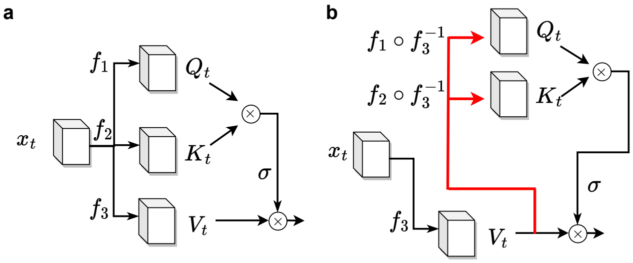

Figure 4. The structure of the transformer-based self-attention mechanism.

In general, $f_3(x_t;\theta_{3t})$,$f_2(x_t;\theta_{2t})$ and $f_1(x_t;\theta_{1t})$ are some learnable matrices. We assume that they are all invertible (they are generally invertible with probability 1 [4, 7]), as shown in Fig.4b, and the attention value in Eq.(20) can be rewritten as

$$\begin{aligned}\mathbf{F}(f_1(x_t;\theta_{1t}), f_2(x_t;\theta_{2t});\phi_t) &= \mathbf{F}(f_1 \circ f_3^{-1}(V_t)), f_2 \circ f_3^{-1}(V_t));\phi_t) \\ &\triangleq \tilde{\mathbf{F}}(V_t;\tilde{\phi}_t) \\ &= \tilde{\mathbf{F}}(f_3(x_t;\theta_{3t});\tilde{\phi}_t).\end{aligned} \tag{21}$$

Eq.(21) is similar to Eq.(10) in the main paper. At this point, the transformer-based model can also be seen as an adaptive step size adaptor with the stiffness information $f_3(x_t;\theta_{3t}) = \frac{1}{\Delta t}(x_{t+1} - x_t)|_{\Delta t=1}$ at x_t as input. Of course, this analysis is not necessarily accurate. In fact, the stiffness information at x_t is first about x_t. For Eq.(20), it would not be surprising if the stiffness information is provided only by x_t. In previous works [1, 2] on channel attention, they also consider x_t as an input to the self-attention module, rather than $f(x_t;\theta_t)$. The experimental results illustrate that the performance of the model can also be improved when x_t is used as input, but the performance of the model with $f(x_t;\theta_t)$ as input is somewhat stronger. This phenomenon may be also attributed to the powerful representational capabilities of neural networks.

Method	CIFAR10	CIFAR100	STL10
ViT	$89.08_{(\pm 0.84)}$	$66.32_{(\pm 1.03)}$	$62.58_{(\pm 1.02)}$
ViT+StepNet	$\mathbf{90.32}_{(\pm 0.48)}$	$\mathbf{68.88}_{(\pm 0.25)}$	$\mathbf{65.58}_{(\pm 0.59)}$

Table 3. The results about ViT with StepNet. All experiments are trained from scratch.

Actually, the channel attention and transformer-based models are two views of the self-attention mechanism, the former considers the self-attention mechanism as an additional module that can be plugged into the backbone, and the latter considers the self-attention mechanism as a part of the backbone. As shown in Table 3, we propose to directly replace the attention modules in ViT with StepNet. Our experimental results demonstrate that the proposed StepNet can indeed be used to enhance the performance of ViT on multiple datasets. Previous works [6, 8] have also shown that the original ViT can be improved by replacing their attention modules. However, since such a replacement is equivalent to changing the core part of the ViT backbone (transformer-based attention module), can the obtained neural network structures still be called transformer-based methods? This is an issue that deserves further discussion and analysis.

References

[1] Jingda Guo, Xu Ma, Andrew Sansom, Mara McGuire, Andrew Kalaani, Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Spanet: Spatial pyramid attention network for enhanced image recognition. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 9

[2] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011–2023, 2020. 9

[3] Zhongzhan Huang, Senwei Liang, Mingfu Liang, Weiling He, and Liang Lin. Layer-wise shared attention network on dynamical system perspective. *arXiv preprint arXiv:2210.16101*, 2022. 8

[4] Zhongzhan Huang, Wenqi Shao, Xinjiang Wang, Liang Lin, and Ping Luo. Convolution-weight-distribution assumption: Rethinking the criteria of channel pruning. *arXiv preprint arXiv:2004.11627*, 2020. 9

[5] Senwei Liang, Zhongzhan Huang, Mingfu Liang, and Haizhao Yang. Instance enhancement batch normalization: An adaptive regulator of batch noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4819–4827, 2020. 8

[6] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 9

[7] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010. 9

[8] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10819–10829, 2022. 9

[9] Jingyu Zhao, Yanwen Fang, and Guodong Li. Recurrence along depth: Deep convolutional neural networks with recurrent layer aggregation. *Advances in Neural Information Processing Systems*, 34:10627–10640, 2021. 8