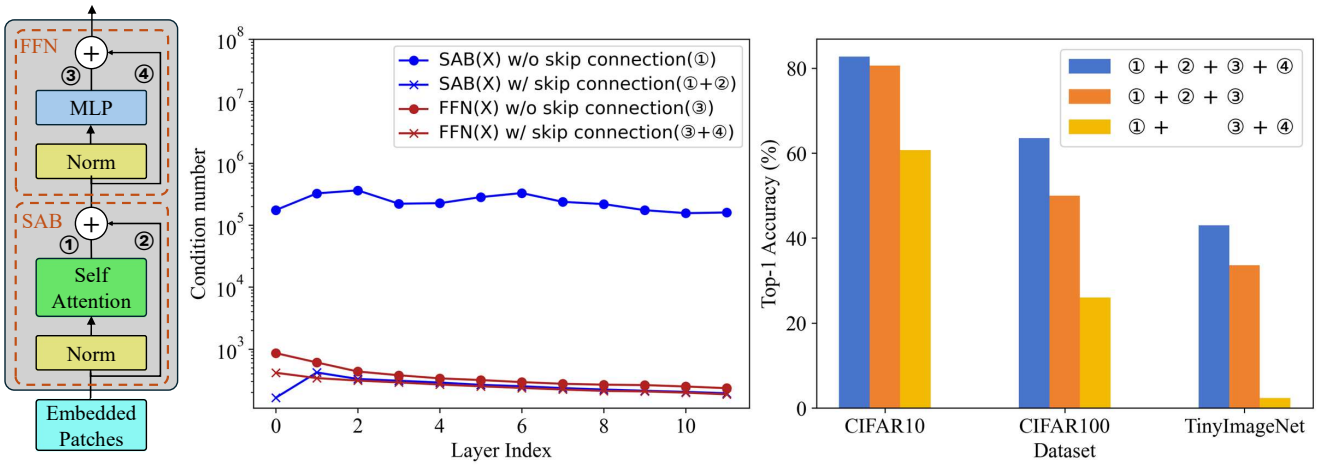
Always Skip Attention

Yiping Ji[1,2], Hemanth Saratchandran[1], Peyman Moghadam[2,3], Simon Lucey[1]
[1]Adelaide University, [2]CSIRO, [3]Queensland University of Technology
{yiping.ji, hemanth.saratchandran, simon.lucy}@adelaide.edu.au,
{yiping.ji, peyman.moghadam}@csiro.au



(a) Architecture

(b) FFN and SAB output embedding condition number using different configurations

(c) Performance after removing skip connections

Figure 1. Removing skip connections from the Self-Attention Block (SAB) versus the Feedforward Network (FFN) in Vision Transformer (ViT) models results in different performance drop. **(a)** Model architecture of ViT consisting of a SAB, and a FFN in each layer. ① represents the output of the self-attention transformation, while ③ represents the output of the MLP transformation. ② and ④ represent identity mappings, referred to as skip connections, in the SAB and FFN, respectively. **(b)** The condition number of the different output embeddings for each hidden layer. A ViT-Tiny model is trained in a typical manner with skip connections on both SAB and FFN blocks (*i.e.*, ① + ② + ③ + ④). It shows that the SAB output embedding without skip connection is poorly conditioned (①), with a condition number significantly higher than that of the other three configurations involving ViT blocks. **(c)** Classification results on three different datasets using the ViT-Tiny model under three different architectures at train time. The result shows the effects of removing skip connections at train time on the SAB and FFN blocks (either ② or ④) in all layers of ViT-Tiny. We observe that removing ② results in a catastrophic performance drop, whereas removing ④ (as predicted) has a modest impact on performance. This difference becomes more pronounced as the scale of the dataset increases.

Abstract

*We highlight a curious empirical result within modern Vision Transformers (ViTs). Specifically, self-attention catastrophically fails to train unless it is used in conjunction with a skip connection. This is in contrast to other elements of a ViT that continue to exhibit good performance (albeit suboptimal) when skip connections are removed. Further, we show that this critical dependence on skip connections is a relatively new phenomenon, with previous deep architectures (*e.g.*, CNNs) exhibiting good performance in their absence. In this paper, we theoretically characterize that the self-attention mechanism is fundamentally ill-conditioned and is, therefore, uniquely dependent on skip connections for regularization. Additionally, we propose **T**oken **G**raying (**TG**), a simple yet effective complement (to skip connections) that further improves the condition of input tokens. We validate our approach in both supervised and self-supervised training methods.*

1. Introduction

Vision Transformers (ViTs) have demonstrated impressive performance on various computer vision and robotic tasks, such as object detection, semantic segmentation, video understanding, visual-language learning, and many oth-

ers [8, 12, 17, 19, 33]. Much of this success is attributed to the self-attention mechanism, commonly referred to as Self-Attention Block (SAB), which allows ViTs to selectively attend to relevant parts of the input sequence when generating each element of the output sequence. Another essential component is the MLP, or Feedforward Networks (FFN), which facilitates intra-token communication across channel dimensions (see Fig. 1). Finally, identity mappings, commonly referred to as skip connections, are incorporated into both the SAB and FFN to enhance model performance.

We argue that the Jacobian of the SAB is disproportionately ill-conditioned compared to other components, notably the FFN block. A poorly conditioned [1] Jacobian is fundamentally detrimental to gradient descent training [1, 13, 20, 24, 25], impeding both convergence and stability. A natural strategy to address this issue is to explicitly model the Jacobian of the entire network. However, this approach is impractical due to the sheer scale of modern transformer architectures. To proceed, we introduce the following simplifying assumptions:

Assumptions
1. The condition of the network Jacobian is bound by the most ill-conditioned sub-block Jacobian.
2. A proxy for the condition of the sub-block Jacobian is the condition of the output embedding of that block.

Although we do not provide formal proofs for these assumptions, we offer indirect validation through extensive empirical analysis. First, we show that better conditioning of a sub-block's output embeddings consistently correlates with improved empirical performance across multiple ViT benchmarks. Second, we demonstrate that this improvement in embedding condition is accompanied by a measurable enhancement in the sub-block's Jacobian condition

Skip in Transformers. Skip connections have become an indispensable component in modern transformer architectures [29] and have been empirically proven to be a *de facto* component. In Fig. 1 (b), we analyze a pre-trained ViT-Tiny model by measuring the condition numbers of the SAB and FFN output embeddings, both with and without skip connections. Our observations reveal that the SAB output embedding becomes highly ill-conditioned in the absence of skip connections, with a condition number approaching e^6. In contrast, the other three configurations have relatively low condition numbers, around e^3. In other words, the SAB output embedding without skip connections has significantly worse condition than the other configurations. To further investigate the impact of skip connections, we train the ViT-Tiny models from scratch under three configurations: the standard SAB and FFN with skip connections, SAB without skip connections, and FFN without skip connections.

[1]The condition number of a full-rank matrix is defined by the ratio of the largest singular value to the smallest singular value. A lower condition number implies better condition.

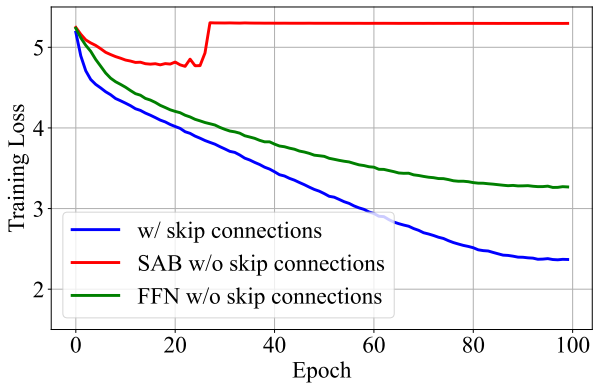


Figure 2. Training loss for three different configurations of ViT-Tiny models trained on the Tiny-ImageNet dataset.

In Fig. 1 (c), we observe a modest drop in classification accuracy (2%) a ViT-Tiny model trained on the CIFAR-10 dataset when skip connections are removed from the FFN while retained in the SAB. However, removing skip connections from the SAB, while keeping them in the FFN, leads to a near-catastrophic performance drop of approximately 22%. Moreover, as dataset size increases, models without SAB skip connections degrade substantially in performance; on larger-scale datasets such as Tiny-ImageNet, these models fail to train altogether. To the best of our knowledge, this empirical finding has not been reported in prior literature. Additionally, the training loss curves for the three configurations are shown in Fig. 2. The results demonstrate that the SAB without skip connections, which is significantly ill-conditioned, exhibits much slower convergence and diverges after 30 epochs. Interestingly, this critical dependence on skip connections is unique to transformer architectures.

Skip in ConvMixer. To further highlight the poor conditioning of the SAB, we conducted a similar experiment on a modern CNN ConvMixer [28]. ConvMixer has a similar architecture and competitive performance to ViT with the exception that self-attention is replaced by a convolution block to spatially mix tokens. Unlike ViT, training ConvMixer-Tiny with and without skip connections, we observed that the performance remains effectively unchanged, with variations of less than ±0.2%. Further details are shown in Fig. 6 in Appendix. This contrast further suggests that the SAB introduces an extreme conditioning issue, making skip connections indispensable for stable training in ViTs.

Our main contributions are summarized as follows.

1. We present a proposition that characterizes why SAB output embedding without skip connection is funda-

mentally ill-conditioned, which challenges training convergence and stability (Sec. 4.1).
2. A theoretical analysis on the role of skip connection within the SAB is undertaken. We demonstrate that it significantly improves the condition of the block's output embedding, enhancing stability and performance (Sec. 4.2).
3. Finally, we propose a novel approach—Token Greying (TG)—to better pre-condition the input tokens. This step, when used in conjunction with conventional skip connections—improves ViT in both supervised and self-supervised settings (Sec. 5).

Our central contribution in this paper is not TG itself, but the insight that the SAB within modern ViTs is intrinsically ill-conditioned. Our hope is that this insight opens up brand new lines of inquiry for the vision community for more effective and efficient ViT design.

2. Related work

Self-Attention. Originally proposed by Vaswani *et al*. [29], the self-attention mechanism has become central to transformer architectures, enabling the capture of long-range dependencies by dynamically weighting interactions between tokens because it captures long-range dependencies by dynamically weighting interactions between tokens. Building on this foundation, transformer-based large language models, such as Llama [6] and Deepseek [15], have significantly advanced language understanding. In computer vision, researchers introduced Vision Transformers (ViTs) [5] for image classification, achieving superior results compared to traditional convolutional neural networks. Moreover, researchers have trained Diffusion Transformers [21], replacing the commonly used U-Net backbone with a transformer that operates on latent patches and inherits the excellent scaling properties of the transformer model class. Similarly, self-attention mechanisms have been successfully applied in speech recognition and physics-informed neural networks, further demonstrating their versatility across diverse applications [3, 32].

Skip connections. He *et al*. [11] introduced skip connections in ResNet to make deep networks easier to optimize and to enhance accuracy by enabling significantly increased network depth. Since their introduction, skip connections have been extensively studied and applied across various Convolutional Neural Network (CNN) architectures. Building on these ideas, Veit *et al*. [30] offer a novel interpretation of residual networks, demonstrating that they can be viewed as a collection of multiple paths of varying lengths, which contributes to the network's flexibility and resilience. Furthermore, Hayou *et al*. [9] addresses gradient stability in deep ResNets, noting that while skip connections alleviate the issue of vanishing gradients, they may also introduce gradient explosion at extreme depths. To counter this, they propose scaling the skip connections according to the layer index, thereby stabilizing gradients and ensuring network expressivity even in the limit of infinite depth. Recently, in Table 3 of [16], the authors reported that VGG—a model without skip connections—achieves comparable results to ResNet—a model with skip connections—on modern datasets. To the best of our knowledge, only one paper challenges the skip connection in transformer architectures. In [10], the authors attempted to train deep transformer networks without using skip connections, achieving performance comparable to standard models. However, this approach requires five times more iterations, and the reasons why skip connections are crucial for self-attention-based transformers remain unresolved.

3. Preliminary

In this section, we define the self-attention blocks used in the Vision Transformer (ViT) architecture and establish the mathematical notation for key quantities referenced in subsequent sections. For a comprehensive overview of transformers, we refer readers to [5].

A Vision Transformer architecture consists of L stacked self-attention blocks (SABs) and feedforward networks (FFNs). An identity mapping, commonly referred to as a skip connection, is applied to connect the inputs and outputs of the transformation in both SAB and FFN.

Self-Attention Blocks in ViT. Formally, given an input sequence $\mathbf{X}_{\text{in}} \in \mathbb{R}^{n\times d}$, with n tokens of dimension d, a SAB is defined as:

$$\mathbf{X}_{\text{out}} := \text{SA}(\mathbf{X}_{\text{in}}) + \mathbf{X}_{\text{in}}, \tag{1}$$

where the self-attention output embedding is

$$\text{SA}(\mathbf{X}_{\text{in}}) = \eta(\mathbf{Q}\mathbf{K}^{\mathbf{T}})\mathbf{V}, \tag{2}$$

where $\mathbf{Q} = \mathbf{X}_{\text{in}}\mathbf{W}_{\text{Q}}$, $\mathbf{K} = \mathbf{X}_{\text{in}}\mathbf{W}_{\text{K}}$ and $\mathbf{V} = \mathbf{X}_{\text{in}}\mathbf{W}_{\text{V}}$ with learnable parameters $\mathbf{W}^{\text{Q}}, \mathbf{W}^{\text{K}}$, and $\mathbf{W}^{\text{V}} \in \mathbb{R}^{d\times d}$. $\eta(\cdot)$ is an activation function and is set by default as $\mathrm{softmax}$ function in [5]. Additionally, some studies [22, 23, 26] have explored the possibilities of using alternative activation functions. In particular, good performance is achieved by using linear attention with an appropriate scale. It allows us to characterize why the SAB output embedding without a skip connection is ill-conditioned.

For general vision transformers, multiple heads $\text{SA}(\mathbf{X}_{\text{in}})_i$, where $1 \le i \le h$, are commonly used. Learnable matrices are divided into head numbers h, such that $\mathbf{W}_{\text{Q}}, \mathbf{W}_{\text{K}}$, and $\mathbf{W}_{\text{V}} \in \mathbb{R}^{h\times d\times d_h}$, where $d_h = \frac{d}{h}$. Then all outputs of each attention head are concatenated together before the skip connection.

Feed Forward Networks in ViT. Formally, given an input sequence $\mathbf{X}_{\text{in}} \in \mathbb{R}^{n \times d}$, with n tokens of dimension d, an FFN is defined as:

$$\mathbf{X}_{\text{out}} := \mathbf{W}_{\text{down}}(g(\mathbf{W}_{\text{up}}(\mathbf{X}_{\text{in}}))) + \mathbf{X}_{\text{in}}, \tag{3}$$

where $\mathbf{W}_{\text{up}} \in \mathbb{R}^{d \times 4d}$ is up projection, $\mathbf{W}_{\text{down}} \in \mathbb{R}^{4d \times d}$ is down projection, and $g(\cdot)$ is an activation function.

Conditioning of matrices. Formally, for a rectangular full rank matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$, the condition number of $\mathbf{A}$ is defined below:

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|_2 \|\mathbf{A}^{\dagger}\|_2 = \frac{\sigma_{\max}(\mathbf{A})}{\sigma_{\min}(\mathbf{A})}, \tag{4}$$

where $\|\cdot\|_2$ is the matrix operator norm induced by the Euclidean norm. We use the Euclidean norm in this paper. $\mathbf{A}^{\dagger}$ is the pseudo inverse of the matrix $\mathbf{A}$. $\sigma_{\max}$ and $\sigma_{\min}$ are the maximal and minimal singular values of $\mathbf{A}$ respectively. We will use Eq. (4) as a metric for analysis in future sections.

4. Theoretical analysis

In this section, we provide a theoretical analysis showing that the output embedding of the SAB without skip connections is highly ill-conditioned. **This matters as the condition of the output embedding is a proxy for the condition of the sub-block's Jacobian**. We then demonstrate that an essential function of the skip connection is to improve this conditioning.

4.1. Self-attention is ill-conditioned

We attempt to validate our claim for the simpler case of linear attention.

Proposition 4.1. *Assume* $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{W_Q}$, *and* $\mathbf{W_K}$ *and* $\mathbf{W_V} \in \mathbb{R}^{d \times d}$ *have entries that are independently drawn from a distribution with zero mean. The condition number of the SAB output embedding (without skip-connection) can be expressed as:*

$$\kappa(\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^T\mathbf{X}^T\mathbf{X}\mathbf{W}_V) \leq C \cdot \left(\frac{\sigma_{max}}{\sigma_{min}}\right)^3, \tag{5}$$

where $\sigma_{\max}$ and $\sigma_{\min}$ are the maximal and minimal singular values of $\mathbf{X}$. C is a fixed constant that is statistically close to unity.

Proposition 4.1 assumes the use of linear attention (*i.e.*, no softmax) as advocated for in [26]. This was done to offer a simplified theoretical exposition on how the SAB without skip connection can affect the condition of the output embedding. The proposition shows the condition number of the output embedding without skip connection is bounded by the cube of the condition number of the input matrix $\mathbf{X}$. Unless the condition of $\mathbf{X}$ is unity, the output of the SAB without skip connection will therefore result in a poorly conditioned embedding. Applying this same process across multiple layers within a transformer results in a highly ill-conditioned outcome.

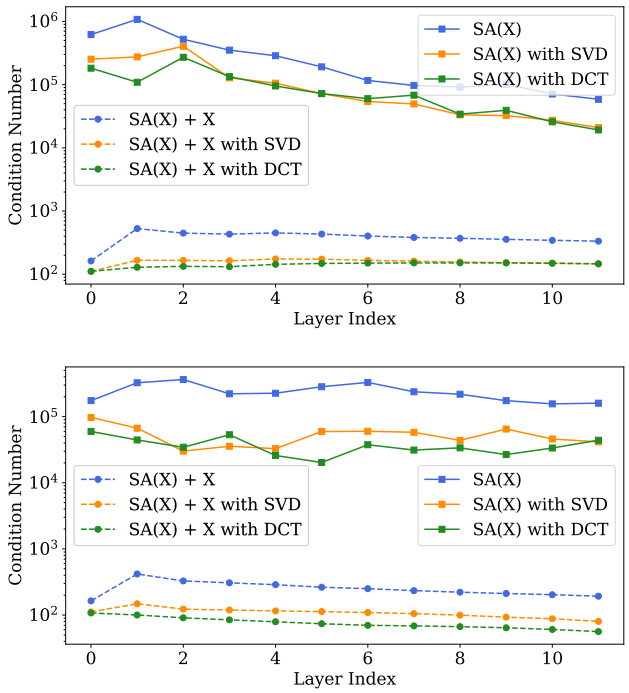


Figure 3. Condition numbers across all layers of trained ViT-Tiny on Tiny-ImageNet datasets. We use 10% validation sets to plot the above figures. **Upper:** fixed scale linear self-attention. **Lower:** softmax self-attention

Fig. 3 illustrates our theoretical analysis by showing that the linear self-attention operation increases the condition number of embeddings from e^3 to e^6, and we can empirically extend this analysis to *softmax* self-attention. The output embedding of the *softmax* SAB without skip connections is also significantly ill-conditioned.

Further, for condition numbers of FFN output embeddings without skip connection using linear activation g we have:

$$\kappa(\text{MLP}(\mathbf{X})) = \kappa(\mathbf{W}_{\text{down}}\mathbf{W}_{\text{up}}\mathbf{X}) \leq C_{\text{down}} \cdot C_{\text{up}} \cdot \frac{\sigma_{\max}}{\sigma_{\min}}, \tag{6}$$

where $\sigma_{\max}$ and $\sigma_{\min}$ are maximal singular values of $\mathbf{X}$ computed using SVD. Since $\mathbf{W}_{\text{down}}$ and $\mathbf{W}_{\text{up}}$ do not depend on data $\mathbf{X}$, we denote $\kappa(\mathbf{W}_{\text{down}}) = C_{\text{down}}$ and $\kappa(\mathbf{W}_{\text{up}}) = C_{\text{up}}$.

Compared to the self-attention mechanism, the MLP transformation does not worsen conditioning as severely, since it has a much lower bound on the condition number. In Fig. 1 (b), the output embeddings of the FFN without

skip connection is observed to have a significantly better condition than the output embedding of the SAB without skip connection. Therefore, in (c), the performance drop after removing the FFN's skip connections is less severe than after removing the SAB's skip connections.

4.2. Skip connection improves condition

In this subsection, we theoretically demonstrate that the skip connection improves the condition of the self-attention output embeddings.

Previous works [4, 10, 18] have demonstrated that transformers, without skip connections, converge very slowly or may even become un-trainable. In [10], the authors show that without skip connections, output embeddings of the SAB quickly converge to rank 1. They hypothesize that this is due to the inability of the transformer to propagate signals through the self-attention mechanism. This rank-1 output embedding implies an infinite condition number. The authors introduce an inductive bias into the self-attention mechanism to enable what they refer to as "better signal propagation". While their approach matches the performance of standard (vanilla) models, it incurs a substantial computational cost—running approximately five times slower. We hypothesize that the success of this method could be directly attributed to its ability to improve the condition of the self-attention output embeddings.

This observation further motivates our view that the primary role of skip connections may be to regularize the condition of the self-attention output embeddings

Proposition 4.2. *Let* $\mathbf{X} \in \mathbb{R}^{n\times d}$ *and* $\mathbf{W_Q}, \mathbf{W_K}, \mathbf{W_V} \in \mathbb{R}^{d\times d}$ *have entries that are independently drawn from a distribution with zero mean. We define* $\mathbf{M} = \mathbf{W_Q}\mathbf{W_K^T}\mathbf{X^T}\mathbf{X}\mathbf{W_V}$ *and assume* $\mathbf{M}$ *is the positive semi-definite matrix. We have the following bounds on the condition numbers* κ*.*

$$\kappa(\mathbf{XM} + \mathbf{X}) \ll \kappa(\mathbf{XM}). \tag{7}$$

Proposition 4.2 shows that during training, the condition number of the SAB output embedding has a much lower bound than that of the SAB output embedding without skip connections. Furthermore, with other conditions being equal, better condition of SAB can contribute to faster convergence and more stable gradient updates, reducing the risk of gradient explosion or vanishing. This ensures the signal can propagate well during the training process in the Vision Transformers.

In other words, the self-attention mechanism enables tokens to communicate by focusing on different parts of the input, thereby capturing correlations between elements within the same sequence. However, one critical drawback is that it is ill-conditioned and challenging to train. In contrast, although skip connections do not facilitate communication between tokens, they help regularize the condition of output embeddings and aid in the optimization process.

As shown in Fig. 3, empirical results demonstrate that in both the linear and $\mathrm{softmax}$ self-attention cases, adding skip connections significantly improves the condition of output embeddings.

Algorithm 1 SVD Token Graying

▷ **Input:** training data, amplification coefficient $\epsilon \in (0, 1]$
▷ **Training:**
for each minibatch $\{(\mathbf{x}_i, y_i)\}_{i=1}^{k}$ **do**
  Convert image $\mathbf{x}_i$ into token $\mathbf{X}$
  $\mathbf{U\Sigma V^T} \leftarrow \mathbf{X}$ ▷ Compute SVD
  $\tilde{\mathbf{\Sigma}} \leftarrow \frac{\mathbf{\Sigma}}{\max(\mathbf{\Sigma})}$ ▷ Normalize elements $\in (0, 1]$
  $\tilde{\mathbf{\Sigma}} \leftarrow \tilde{\mathbf{\Sigma}}^{\epsilon}$ ▷ Amplify
  $\tilde{\mathbf{X}} \leftarrow \mathbf{U}\tilde{\mathbf{\Sigma}}\mathbf{V^T}$ ▷ SVD reconstruct
  Patch Embedding: $\mathbf{Z} \leftarrow \mathrm{PatchEmbed}(\tilde{\mathbf{X}})$
  Forward Pass: $\hat{y} \leftarrow \mathrm{Model}(\mathbf{Z})$
  Compute Loss: $\mathcal{L} \leftarrow \mathrm{Loss}(\hat{y}, y_i)$
  Backward Pass: Compute gradients $\nabla_\theta \mathcal{L}$
  Parameter Update: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$
end for

5. Methodology

In this section, we introduce **T**oken **G**raying **(TG)**, a simple yet effective method that employs Singular Value Decomposition and Discrete Cosine Transform techniques to reconstruct better-conditioned input tokens.

5.1. Singular Value Decomposition (SVD)

Singular Value Decomposition of a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the factorization of $\mathbf{X}$ into the product of three matrices $\mathbf{X} = \mathbf{U\Sigma V^T}$ where the columns of $\mathbf{U} \in \mathbb{R}^{n\times n}$ and $\mathbf{V} \in \mathbb{R}^{d\times d}$ are orthonormal and $\mathbf{\Sigma} \in \mathbb{R}^{n\times d}$ is a rectangular diagonal matrix whose diagonal elements are non-negative in descending order and represent the singular values such that $\Sigma_{ii} = \sigma_i$. As observed in Eq. (4), we can reduce the condition number by increasing the minimal singular value while keeping the maximal singular value unchanged. Hence, our goal is to reconstruct $\tilde{\mathbf{X}} = \mathbf{U}\tilde{\mathbf{\Sigma}}\mathbf{V^T}$, where $\tilde{\mathbf{\Sigma}}$ is formed by amplifying non-maximal elements of $\mathbf{\Sigma}$ while keeping the maximal element unchanged. Our SVD token graying pipeline is presented in Algorithm 1.

However, directly applying SVD to matrices is computationally expensive, with a cost of $O(nd \cdot \min(n, d))$ leading to longer training time. Tab. 1 demonstrates that training a ViT is significantly slower when using SVD reconstruction. Therefore, in the next subsection, we introduce a more efficient **T**oken **G**raying (**TG**) method using the discrete cosine transform.

Methods	ViT-B	ViT-B + SVDTG	ViT-B + DCTTG
Time (days)	0.723	4.552	0.732

Table 1. ViT-Base training time w/o and w/ token graying methods on ImageNet-1K dataset.

5.2. Discrete Cosine Transform (DCT)

A Discrete Cosine Transform (DCT) expresses a finite sequence of data points as a sum of cosine functions oscillating at different frequencies. It is real-valued and has an inverse, which is often simply called the Inverse DCT (IDCT).

There are several variants of the DCT and in this work we choose DCT-II which is one of the most commonly used forms. Given a real-valued sequence data $\mathbf{x} \in \mathbb{R}^{N\times 1}$, the 1D DCT sequence $\hat{\mathbf{x}}$ is defined as:

$$\hat{\mathbf{x}}_k = \alpha_k \sum_{i=0}^{N-1} x_i \cos\left[\frac{\pi(2i+1)k}{2N}\right], \text{for } k = 0, 1, ..., N-1$$
$$\text{where } \alpha_k = \begin{cases} \sqrt{\frac{1}{N}}, \text{if } k = 0 \\ \sqrt{\frac{2}{N}}, \text{if } k \neq 0 \end{cases} \tag{8}$$

Simplicitly, 1D DCT is a linear transformation, Eq. (8) can be written with the formula $DCT(\mathbf{x}) = \hat{\mathbf{x}} = D\mathbf{x}$, where $D_{i,k}$ is the 1D DCT basis and orthogonal.

$$D_{i,k} = \alpha_k \cos\left[\frac{\pi(2i+1)k}{2N}\right]. \tag{9}$$

For the reconstruction process, since the DCT expresses a signal as a sum of cosine functions with different frequencies, the IDCT combines these frequency components to recover the original signal. Hence, $\text{IDCT}(\hat{x}) = \tilde{\mathbf{x}} = D^T\hat{\mathbf{x}}$. In our case, we have a token matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and we generalize the 1D DCT to a 2D DCT using the formula $\hat{\mathbf{X}} = D\mathbf{X}D^T$. In natural images, the dominant singular vector, associated with the largest singular value, typically captures the low frequency content of the image. Intuitively, this observation suggests that DCT reconstruction can serve as an approximation for SVD reconstruction. Our DCT token graying pipeline is presented in Algorithm 2.

DCT avoids the expensive computational cost of SVD and has a complexity of $O(nd \cdot \log(nd))$. Tab. 1 shows that training ViT using the DCT algorithm is significantly faster than SVD algorithm.

Algorithm 2 DCT Token Graying

▷ **Input:** training data, amplification coefficient $\epsilon \in (0, 1]$
▷ **Training:**
for each minibatch $\{(\mathbf{x}_i, y_i)\}_{i=1}^{k} \sim$ training data **do**
  Convert image $\mathbf{x}_i$ into token $\mathbf{X}$
  $D\hat{\mathbf{X}}D^T \leftarrow \mathbf{X}$ ▷ Compute DCT
  $\hat{\mathbf{X}}_{\text{norm}} \leftarrow |\hat{\mathbf{X}}|/\max(|\hat{\mathbf{X}}|)$ ▷ Normalize elements $\in (0, 1]$
  $\hat{\mathbf{X}} \leftarrow \hat{\mathbf{X}}^{\epsilon}_{\text{norm}} \cdot \text{sign}(\hat{\mathbf{X}}) \cdot \max(|\hat{\mathbf{X}}|)$ ▷ Amplify
  $\tilde{\mathbf{X}} \leftarrow \hat{D}^T\hat{\mathbf{X}}\hat{D}$ ▷ Inverse DCT
  Patch Embedding: $\mathbf{Z} \leftarrow \text{PatchEmbed}(\tilde{\mathbf{X}})$
  Forward Pass: $\hat{y} \leftarrow \text{Model}(\mathbf{Z})$
  Compute Loss: $\mathcal{L} \leftarrow \text{Loss}(\hat{y}, y_i)$
  Backward Pass: Compute gradients $\nabla_\theta \mathcal{L}$
  Parameter Update: $\theta \leftarrow \theta - \eta\nabla_\theta \mathcal{L}$
end for

6. Experiments

Vision Transformers (ViTs) have emerged as powerful models in the field of computer vision, demonstrating remarkable performance across a variety of tasks. This section is dedicated to validating and analyzing our proposed algorithm for both supervised and self-supervised learning in ViTs. For all experiments, unless specified otherwise, we use $\epsilon = 0.95$.

6.1. Supervised learning ViT

In this subsection, we validate our methods in supervised learning using different types and scales of ViTs. *Cross-entropy* loss objective is used. We use PyTorch and Timm libraries.

– ViT [5]: is the pioneering work that interprets an image as a sequence of patches and processes it by a standard Transformer encoder as used in NLP. This simple, yet scalable, strategy works surprisingly well when coupled with pre-training on large datasets. We train the ViT-Tiny, which has 12 layers and 3 heads, with a head dimension of 64 and a token dimension of 192 on the Tiny-ImageNet dataset [14]. Additionally, we train ViT-Base, which consists of 12 layers and 12 heads, each with a head dimension of 64 and a token dimension of 768 on the ImageNet-1K dataset [2].

– Swin [17] is a hierarchical vision transformer designed to focus on local attention first and expand globally. It improves computational efficiency and locality modeling compared to standard ViTs. We train on the Swin small model with a patch size of 4 and a window size of 7 on the ImageNet-1K dataset.

– CaiT [27] is a variant Vision Transformer that aims at increasing the stability of the optimization when ViTs go deeper. We train CaiT small, which has 24 layers, on the ImageNet-1K dataset.

– PVT [31] an enhanced version of the Pyramid Vision Transformer that utilizes a hierarchical architecture to ex-

Method	Top-1 Acc (%)	Top-5 Acc (%)
$\frac{1}{k}(\mathbf{QK})$	43.0	62.7
$\frac{1}{k}(\mathbf{QK})$ + SVD	44.3	68.8
$\frac{1}{k}(\mathbf{QK})$ + DCT	44.2	68.8
$\mathrm{softmax}(\mathbf{QK})$	43.0	67.3
$\mathrm{softmax}(\mathbf{QK})$ + SVD	44.7	69.7
$\mathrm{softmax}(\mathbf{QK})$ + DCT	44.8	69.7

Table 2. ViT-Tiny results on Tiny-ImageNet datasets. In the upper part, we show results for linear self-attention, where $k = 16$ is chosen based on [26] w/ and w/o our method. On the lower part, we show regular *softmax* self-attention results w/ and w/o our method.

Method	ϵ	Top-1 Acc (%)	κ_{in}	κ_{out}
ViT-Base	–	81.0	6.72	6.74
ViT-Base + SVDTG	0.9	81.2	6.64	6.66
	0.8	81.2	6.47	6.66
	0.7	**81.4**	6.15	6.17
	0.6	**81.4**	5.73	5.71
	0.5	81.0	5.29	5.25
ViT-Base + DCTTG	0.97	81.1	6.47	6.49
	0.95	**81.3**	6.42	6.43
	0.93	81.2	6.33	6.35
	0.9	81.1	6.25	6.25
	0.85	81.0	6.01	6.02

Table 3. ViT-Base results using SVD and DCT token graying methods with different amplification coefficient ϵ. κ_{in} and κ_{out} are the average token condition numbers in log scale before and after each SAB.

tract multi-scale features for improved performance in computer vision tasks. It incorporates efficient attention mechanisms and refined positional encoding to boost accuracy and reduce computational complexity.

Results. In Tab. 2, we first demonstrate the results for the linear case and the regular $\mathrm{softmax}$ case using our methods on a small model (ViT-Tiny) and the small-scale Tiny-ImageNet dataset. Given that Theorem Proposition 4.2 and Proposition Proposition 4.1 are mathematically explained using linear self-attention, we train the ViT-Tiny model with linear self-attention, which aligns well with our analysis in Fig. 3. Furthermore, as shown in Fig. 4, both SVDTG and DCTTG lead to better conditioning of SAB output embeddings throughout training epochs. Next, we train the model on a larger dataset using the default $\mathrm{softmax}$ self-attention, further validating that our approach generalizes to $\mathrm{softmax}$ self-attention. Moreover, SVDTG is the most direct method to improve the conditioning of a matrix, and as demonstrated in Tab. 3, it leads to notable performance gains.

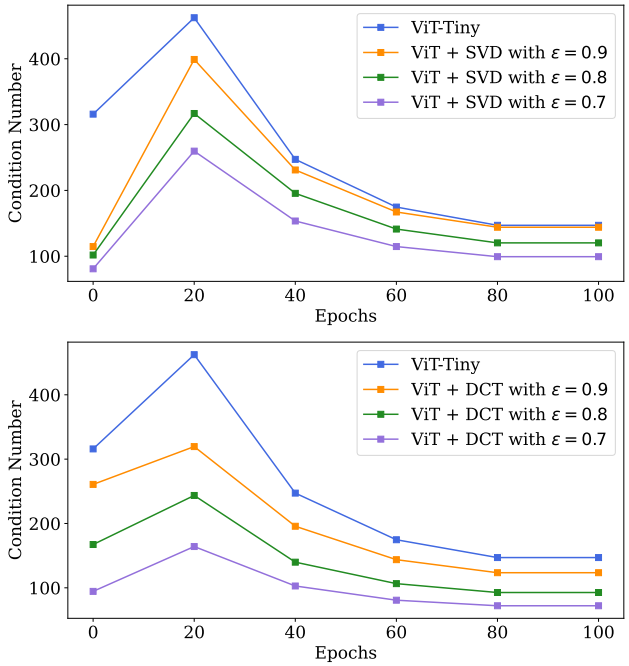


Figure 4. We compute the condition number of the SAB block's output embeddings during ViT-Tiny training using SVDTG and DCTTG with different ϵ.

To address the high computational cost—approximately six times slower than regular ViT training—we approximate SVDTG using DCTTG, which achieves better token embeddings conditioning and yields comparable results too. As shown in Fig. 5, our DCTTG method produces condition numbers across all layers that indicate better token conditioning in almost every layer. Finally, we evaluate our methods on different variants of ViTs trained on ImageNet-1K datasets, where our DCTTG method demonstrates superior performance compared to vanilla models in terms of Top-1 and Top-5 classification accuracy, while also achieving better conditioning of SAB output embeddings across all layers in the ViT-B model.

6.2. Self-supervised learning ViT

In this subsection, we validate our methods on self-supervised learning models as well as on fine-tuning pre-trained models [7, 12]. Self-supervised learning is a powerful technique in which models learn representations directly from unlabeled data. Masked Image Modeling (MIM) [12] learns representations by reconstructing images that have been corrupted through masking.

Pre-training. In the pre-training stage, we pre-train our model on the ImageNet-1k training set without using ground-truth labels in a self-supervised manner. We use a batch size of 2048, a base learning rate of 1.5×10^{-4}, and a

Method	Top-1 Acc (%)	Top-5 Acc (%)
ViT-S	80.2	95.1
ViT-S + DCTTG	**80.4**	**95.2**
ViT-B	81.0	95.3
ViT-B + DCTTG	**81.3**	**95.4**
Swin-S	81.3	**95.6**
Swin-S + DCTTG	**81.6**	**95.6**
CaiT-S	82.6	96.1
CaiT-S + DCTTG	**82.7**	**96.3**
PVT V2 b3	82.9	96.0
PVT V2 b3 + DCTTG	**83.0**	**96.1**

Table 4. Top-1 and Top-5 classification accuracy on ImageNet-1K dataset using DCT token graying on different ViTs configurations.

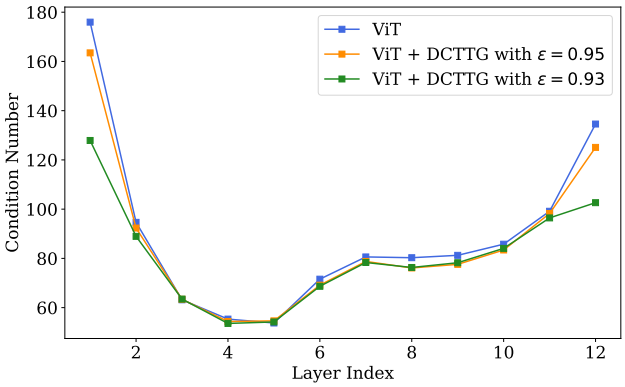


Figure 5. Condition numbers of all layers in the trained ViT-Base model with and without the DCTTG method.

Method	Top-1 Acc (%)	Top-5 Acc (%)
MAE	83.0	96.4
MAE + DCTTG	**83.2**	**96.6**

Table 5. MAE pretrained ViT-Base results finetuned on ImageNet dataset-1k. We evaluate our MSVR method on self-attention using the power iteration method.

weight decay of 0.05, and we train for 400 epochs. Our loss function computes the mean squared error (MSE) between the reconstructed and original images in pixel space.

Finetuning and reuslts. In the finetuning stage, we finetune the pre-trained models on ImageNet-1k for classification tasks using *cross-entropy* loss. We use a batch size of 512, a base learning rate of 5×10^{-4}, a weight decay of 0.05, and train for 100 epochs. As shown in Tab. 5, our DCTTG method outperforms the vanilla MAE model by 0.2% in both top-1 and top-5 classification accuracy.

7. Limitation

Our work establishes a theoretical foundation demonstrating why skip connections act as essential conditioning regularizers within self-attention blocks. Additionally, we introduce a method designed to further enhance the output embeddings of self-attention blocks. However, certain limitations remain. For instance, while our regularizer is effective, training in low-precision settings may pose challenges due to the DCT operation, which involves many multiplications and summations that are sensitive to quantization errors. Additionally, for some variants of ViTs, the performance improvement might be marginal.

8. Conclusion

This paper provides a theoretical demonstration that the output embeddings of linear self-attention blocks, when lacking skip connections, are inherently ill-conditioned, a finding that empirically extends to conventional softmax self-attention blocks as well. Furthermore, we show that skip connections serve as a powerful regularizer by significantly improving the conditioning of self-attention blocks' output embeddings. In addition, we use the condition of the self-attention output embeddings as a proxy for that of its Jacobian, empirically demonstrating that input tokens with improved conditioning lead to a better-conditioned Jacobian. Building on this insight, we introduce SVDTG and DCTTG, simple yet effective methods designed to improve the condition of input tokens. We then validate their effectiveness on both supervised and unsupervised learning tasks. Ultimately, we hope these insights will guide the vision community in designing more effective and efficient ViT architectures for downstream applications.

Acknowledgements. This work was supported by the Australian Institute for Machine Learning (AIML) and the CSIRO's Data61 Embodied AI Cluster.

References

[1] Paul Albert, Frederic Z Zhang, Hemanth Saratchandran, Cristian Rodriguez-Opazo, Anton van den Hengel, and Ehsan Abbasnejad. Randlora: Full-rank parameter-efficient fine-tuning of large models. In *International Conference on Learning Representations (ICLR)*, 2025. 2

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009. 6

[3] Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888, 2018. 3

[4] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas.

Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pages 2793–2803. PMLR, 2021. 5

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 3, 6

[6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3

[7] Maryam Haghighat, Peyman Moghadam, Shaheer Mohamed, and Piotr Koniusz. Pre-training with Random Orthogonal Projection Image Modeling. In *International Conference on Learning Representations (ICLR)*, 2024. 7

[8] Stephen Hausler and Peyman Moghadam. Pair-VPR: Place-aware pre-training and contrastive pair classification for visual place recognition with vision transformers. *IEEE Robotics and Automation Letters*, 2025. 2

[9] Soufiane Hayou, Eugenio Clerico, Bobby He, George Deligiannidis, Arnaud Doucet, and Judith Rousseau. Stable resnet. In *International Conference on Artificial Intelligence and Statistics*, pages 1324–1332. PMLR, 2021. 3

[10] Bobby He, James Martens, Guodong Zhang, Aleksandar Botev, Andrew Brock, et al. Deep transformers without shortcuts: Modifying self-attention for faithful signal propagation. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 5

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 7

[13] Yiping Ji, Hemanth Saratchandran, Cameron Gordon, Zeyu Zhang, and Simon Lucey. Efficient learning with sine-activated low-rank matrices. In *The International Conference on Learning Representations*, 2025. 2

[14] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 6

[15] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 3

[16] Zhuang Liu and Kaiming He. A decade's battle on dataset bias: Are we there yet? In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 6

[18] Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[19] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, et al. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024. 2

[20] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks, 2013. 2

[21] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3

[22] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, et al. cosformer: Rethinking softmax in attention. In *International Conference on Learning Representations*, 2022. 3

[23] Jason Ramapuram, Federico Danieli, Eeshan Gunesh Dhekane, Floris Weers, Dan Busbridge, et al. Theory, analysis, and best practices for sigmoid self-attention. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[24] Hemanth Saratchandran and Simon Lucey. Enhancing transformers through conditioned embedded tokens. *arXiv preprint arXiv:2505.12789*, 2025. 2

[25] Hemanth Saratchandran, Thomas X Wang, and Simon Lucey. Weight conditioning for smooth optimization of neural networks. In *European Conference on Computer Vision*, pages 310–325. Springer, 2024. 2

[26] Hemanth Saratchandran, Jianqiao Zheng, Yiping Ji, Wenbo Zhang, and Simon Lucey. Rethinking softmax: Self-attention with polynomial activations, 2024. 3, 4, 7

[27] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021. 6

[28] Asher Trockman and J. Zico Kolter. Patches are all you need?, 2022. 2, 10

[29] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2, 3

[30] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. *Advances in neural information processing systems*, 29, 2016. 3

[31] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 6

[32] Zhiyuan Zhao, Xueying Ding, and B Aditya Prakash. Pinnsformer: A transformer-based framework for physics-informed neural networks. In *The International Conference on Learning Representations (ICLR)*, 2024. 3

[33] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7641–7649, 2024. 2