Registers in Small Vision Transformers: A Reproducibility Study of Vision Transformers Need Registers

Anonymous authors
Paper under double-blind review

Abstract

Recent work has shown that Vision Transformers (ViTs) can produce "high-norm" artifact tokens in attention maps. These artifacts disproportionately accumulate global information, can degrade performance, and reduce interpretability in these models. Darcet et al. (2024) proposed registers—auxiliary learnable tokens—to mitigate these artifacts. In this reproducibility study, we verify whether these improvements extend to smaller ViTs. Specifically, we examine whether high-norm tokens appear in a DeiT-III Small model, whether registers reduce these artifacts, and how registers influence local and global feature representation. Our results confirm that smaller ViTs also exhibit high-norm tokens and registers partially alleviate them, improving interpretability. Although the overall performance gains are modest, these findings reinforce the utility of registers in enhancing ViTs while highlighting open questions about their varying effectiveness across different inputs and tasks. Our code is available at `https://anonymous.4open.science/r/regs-small-vits`.

1 Introduction

Vision transformers (ViTs) have emerged as a powerful alternative to convolutional neural networks (CNNs) in computer vision, offering flexibility in learning both local and global interactions within images. Originally introduced by Dosovitskiy et al. (2021), ViTs use self-attention mechanisms to capture both local and global interactions within images, enabling strong performance across tasks such as image classification, segmentation, and detection.

Building on the foundational ViT architecture, advancements like DeiT (Touvron et al., 2021), DeiT-III (Touvron et al., 2022), and the DINOv2 model (Oquab et al., 2023) have significantly enhanced the efficiency and effectiveness of ViTs. DeiT and DeiT-III introduced refined training methodologies including new augmentation techniques and optimized regularization strategies, making ViTs more data-efficient and practical for datasets like ImageNet-1k and ImageNet-21k (Deng et al., 2009). Concurrently, DINOv2 reinforced the role of self-supervised learning in ViTs by generating robust, general-purpose visual features from large-scale curated datasets, enabling superior performance in both image-level and pixel-level tasks.

Despite these advancements, ViTs, including enhanced versions like DeiT and DINOv2, face significant challenges. Darcet et al. (2024) identified the presence of artifacts in the self-attention maps of transformer blocks, which encapsulate global information and inadvertently lead to the loss of crucial local details. The authors argue that traditional ViTs face two significant limitations: (1) a heavy reliance on transformer layers to encode all image-level information, which can lead to inefficiencies, and (2) the lack of explicit mechanisms for maintaining state information across layers, potentially affecting long-term coherence in representations.

To address these limitations, Darcet et al. (2024) introduced the concept of registers—special tokens inspired by computer architecture that accumulate and carry information between layers—as an enhancement to the ViT framework. By functioning as memory units within the transformer sequence, registers provide a structured mechanism for managing both local and global information more effectively. This approach reduces reliance on transformer layers to encode all image-level information and offers explicit state maintenance across layers. Integrating registers has been shown to improve interpretability and feature smoothness by mitigating inconsistencies in feature maps and balancing local and global contexts more effectively.

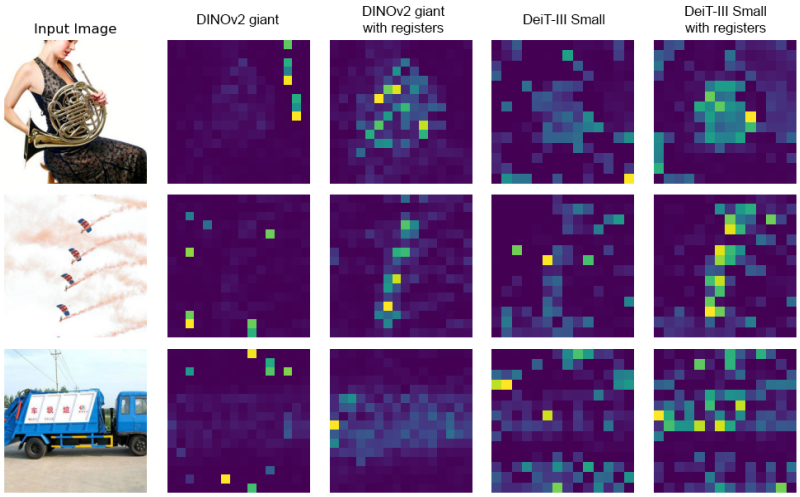


Figure 1: A comparison of attention maps from the DINOv2 giant model with and without registers and our DeiT-III small models with and without register tokens. Register tokens enable interpretable attention maps for most images of the DINOv2 model, while they do not necessarily improve the attention maps of the DeiT-III model.

The introduction of additional tokens for the purpose of retaining abstract information to transformer architectures has been proven effective in the field of NLP. Most relevant for our investigation is the work of Burtsev et al. (2020), who showed that concatenating *memory tokens* to a series of token embeddings improves translation quality in transformer-based machine translation models. The authors observe a similar issue to the artifacts in ViTs when mixing local and global information within token embeddings, inevitably leading to inefficiencies. While the authors extend their analysis to separate memory controllers, their results indicate that simply enabling the model to isolate global sequence representations improves performance and interpretability.

In this reproducibility study, we examine the main claims from Darcet et al. (2024) and extend their experiments to smaller-scale models. Specifically, we explore whether (1) high-norm artifact tokens are also present in a smaller ViT, (2) registers can mitigate them effectively and (3) the improvements in local-global encoding reported for larger models also generalize to smaller architectures. By replicating and expanding these experiments, we aim to clarify the broader relevance of registers in stabilizing and refining Vision Transformer representations.

2 Scope of Reproducibility

In their study, Darcet et al. (2024) use ground-up-trained DeiT-III base, OpenCLIP, and DINOv2 large models. We replicate these original experiments on larger pretrained ViT models and extend the investigation to a smaller DeiT-III Small model, trained from scratch on ImageNet-1k. Our focus is on determining whether high-norm tokens persist across different architectures and scales and whether registers consistently alleviate these anomalies while improving learned representations. By exploring smaller ViTs, we assess whether the benefits observed in previous work hold in more constrained architectures.

2.1 Model Architecture and Registers Implementation

Artifact tokens, visible as strong peaks in self-attention maps (Figure 1), indicate where the model allocates disproportionate attention, potentially destabilizing its representation. Additionally, these outliers can be identified as having a significantly raised token norm. Throughout this work, we will use the terms high-norm tokens and artifacts synonymously.

Darcet et al. (2024) addressed this by introducing *registers*—additional learnable tokens that act as intermediate memory, reducing the transformer's reliance on attention layers alone for global context. In that work, registers were added to a DINOv2-based ViT (see Figure 2 of Darcet et al., 2024), where they join the input sequence alongside the [CLS] and patch tokens. This approach mitigated high-norm artifacts and improved representations, particularly in dense tasks where patch tokens are retained.

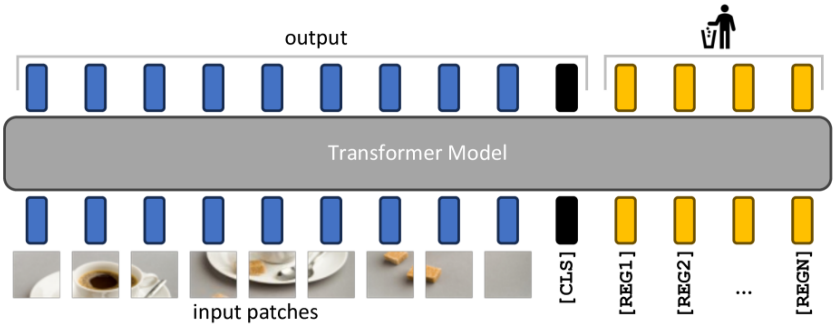


Figure 2: Illustration of the proposed remediation and resulting model. We add N additional learnable input tokens (depicted in yellow), that the model can use as registers. At the output of the model, only the patch tokens and [CLS] tokens are used, both during training and inference. *Adapted from Darcet et al. (2024).*

In our study, we adopt the same core idea of the original paper by adding registers to DeiT-III models. Specifically, we investigate whether a DeiT-III Small architecture benefits similarly from these additional tokens. We focus on confirming the presence of high-norm tokens in smaller models and evaluating whether registers effectively resolve or reduce these artifacts.

2.2 Main Claims

The original paper by Darcet et al. (2024) highlights three key points that we seek to reproduce and validate in this study:

1. **High-Norm Tokens (Artifacts) in Larger Models:** High-norm "artifact" tokens consistently appear in large-scale ViTs.
2. **Register-Based Mitigation:** Introducing registers—additional learnable tokens acting as intermediate memory—reduces or eliminates these high-norm artifacts, leading to more stable and interpretable attention distributions.
3. **Enhanced Local-Global Encoding:** High-norm tokens reflect an imbalance in how local and global information is captured. By segregating global information into registers, ViTs retain a more effective balance of local and global features, resulting in more coherent and expressive intermediate representations.

Together, these claims propose that registers offer a robust architectural enhancement for ViTs, improving both interpretability and representation quality and potentially generalizing across different model scales and training paradigms.

2.3 Broader Context

The incorporation of registers in ViTs makes these models more interpretable and efficient. By reducing inconsistencies in feature and attention maps, the registers clarify the model's internal processes, enhancing trust and understanding. They also streamline information flow, potentially lowering computational overhead and improving scalability, which is a crucial advantage in resource-constrained settings. By incorporating smaller-scale DeiT-III models into this study, we broaden the scope of Darcet et al. (2024) to examine whether their proposed register mechanism also benefits ViTs with fewer parameters. If confirmed, these

findings suggest that registers can offer a robust architectural enhancement to a wider array of ViT variants, potentially leading to more interpretable and efficient models across different scales.

3 Experiments

To address the first two claims outlined above, we conducted experiments on pretrained DINOv2 models as well as DeiT-III models trained from scratch, with and without registers. Our primary analyses include examining attention maps, token norms and final performance metrics on ImageNet-1k.

To assess claim 3, we first evaluate the global information content in the token embeddings of spatial tokens and then assess the LOST performance on our DeiT-III model with registers.

3.1 Model descriptions

In this study, we evaluated the impact of incorporating registers in ViTs using both pretrained DINOv2 models as well as DeiT-III models that were trained from scratch.

Pretrained DINOv2 Models: Pretrained DINOv2 models were obtained from the official repository[1](300M parameters). DINOv2 models use a vanilla ViT as a backbone and use two separate networks for self-distillation, one network creates target embeddings, which the second network learns. We tested variants without registers as controls and with registers for intermediate memory Darcet et al. (2024).

DeiT-III Models: We also trained DeiT-III Small models (22M parameters) from scratch to directly measure the impact of registers on performance and feature representation. One model was trained without registers, while the other included four registers, implemented as additional CLS tokens, following the same method used in DINOv2. To ensure consistency, we matched the hyperparameters from the official DeiT-III repository by Touvron et al., detailed in Appendix A. All DeiT-III Small models were trained for 800 epochs using 16×16 patches and standardized with ImageNet mean-std normalization; in configurations that employed registers, four tokens were added.

3.2 Datasets

This study primarily uses ImageNet-1k (Deng et al., 2009) for training and evaluation [2]. ImageNet-1k contains 1.28 million training images and 50,000 validation images across 1,000 classes. We use the official dataset split, which resizes images to 224×224 pixels and applies ImageNet mean-std normalization.

For the LOST experiment, we use the VOC07 dataset (Everingham et al., 2010). Additionally, we employ Imagenette (Torii, 2020) and Caltech101 (Li et al., 2022) for further evaluation of token representations in specific experiments.

3.3 Experimental setup and code

We integrated the register tokens into the DeiT-III codebase by inserting four additional learnable tokens alongside the [CLS] and patch tokens, as shown in Figure 3. We adopted four registers, in line with Darcet et al. (2024), who used four registers throughout their experiments. These registers are randomly initialized from a normal distribution with low variance, following the approach used for the [CLS] token initialization. Unlike DINOv2-based architectures, our DeiT-III setup discards all patch tokens at the output stage, relying on the [CLS] token alone for classification.

4 Results

This section examines our results in relation to the three main claims from Section 2: (1) the presence of high-norm tokens, (2) the effectiveness of registers in reducing artifacts, and (3) their impact on local-global feature balance. We structure our findings as follows: Section 4.1 evaluates *classification performance*,

[1] `https://github.com/facebookresearch/dinov2/tree/main`
[2] `https://image-net.org/download-images`

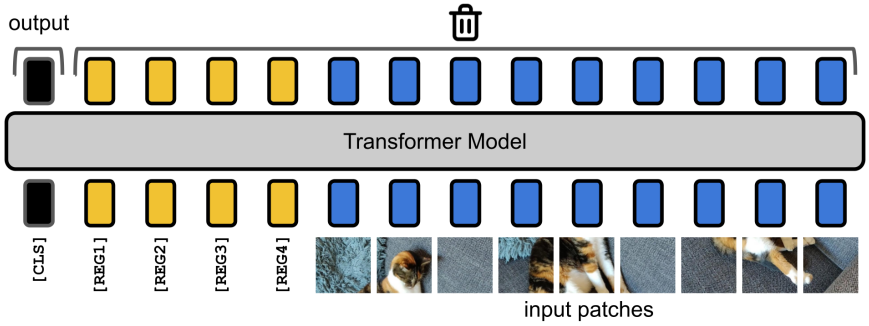


Figure 3: Modified ViT architecture with registers. The 4 additional learnable input tokens (highlighted in yellow) act as registers, while only the [CLS] token is outputted during training and inference.

Section 4.2 analyzes *local features* by measuring high-norm tokens in DINOv2 and DeiT-III, and Section 4.3 investigates *global features*, using LOST for object-level understanding and assessing spatial patch token representations.

4.1 Classification Performance

We begin by evaluating our models on the ImageNet-1k (Deng et al., 2009) validation set. The baseline DeiT-III Small model (without registers) achieves a top-1 accuracy of 81.1%. When four register tokens are added, we observe a slightly lower accuracy of 80.8%. While the addition of registers does not yield an immediate accuracy improvement for image classification, the decrease (0.3%) is small enough to suggest that registers do not substantially impair performance in this setting.

4.2 Local Features

4.2.1 Local information of high-norm tokens

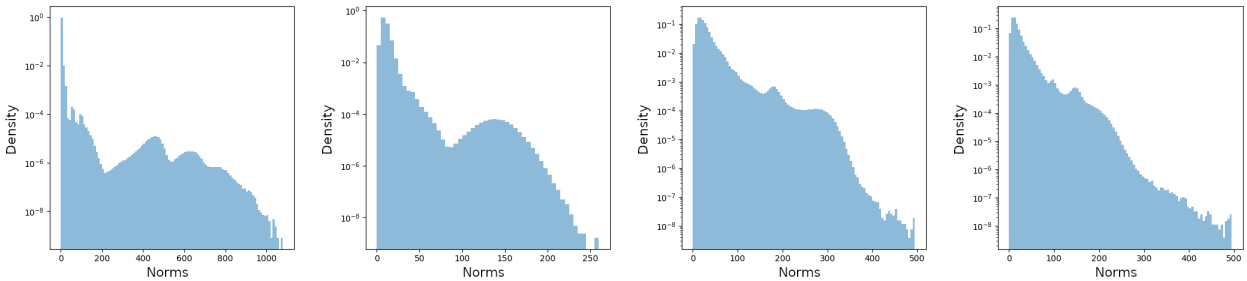
To test whether high-norm tokens primarily carry detailed local information or global context, we train a small Multi-Layer Perceptron (MLP) to reconstruct the original $16 \times 16 \times 3$ image patch from a single token embedding. The MLP has an input size of 384 (matching our DeiT-III hidden dimension), a single hidden layer of size 1536, and an output layer of 768 (for the flattened $16 \times 16 \times 3$ patch). We train this MLP for five epochs on high- and low-norm tokens drawn from our DeiT-III model with no register tokens.

We find that high-norm tokens exhibit a substantially larger MSE of 176.54 compared to 120.02 for low-norm tokens, representing a 147.1% increase in error. This indicates that high-norm tokens lose a significant amount of local detail, supporting the original claim by Darcet et al. (2024) that high-norm "artifact" tokens predominantly capture global information and thus encode fewer spatial details.

4.2.2 Artifacts in DINOv2

We further observe how token norms evolve in the pretrained DINOv2 model. In Figure 4a, the distribution of token norms for the DINOv2 model without registers displays a heavy tail toward large values. These "outlier" tokens, which have notably higher norms than the rest, coincide with the "artifacts" identified by Darcet et al. (2024), indicating an accumulation of excessive global information in just a few tokens. Such imbalanced allocation of global context may adversely affect the encoding of local features.

When registers are introduced (Figure 4b), the long tail of the distribution shrinks significantly, signaling that fewer tokens are carrying disproportionately large norms. This aligns with the idea that registers "absorb" or redistribute global context that would otherwise reside in certain patch tokens. Consequently, no single token disproportionately dominates the self-attention mechanism and the overall norm distribution becomes more balanced.

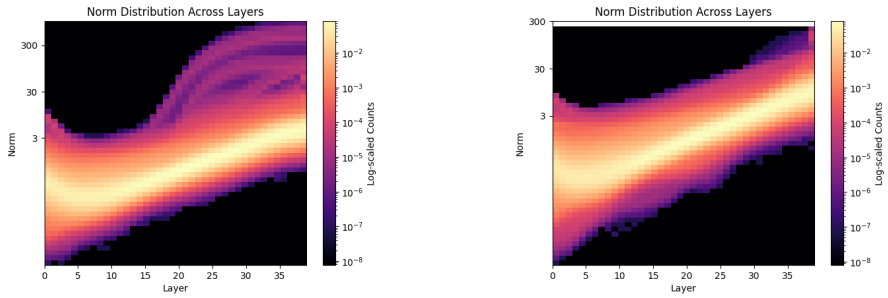


(a) DINOv2 (b) DINOv2 (with registers) (c) DeiT-III (d) DeiT-III (with registers)

Figure 4: Histograms of token norm distributions for DINOv2 (subfigures a–b) and DeiT-III (subfigures c–d), comparing models without registers (left panels) and with registers (right panels).

Looking more closely at how these norms progress across the network, Figure 5 shows layer-wise norm distributions for DINOv2. Without registers (left), the deeper layers produce outlier tokens that far exceed the typical norm range, indicating an over-accumulation of global context. With registers (right), the distribution becomes smoother and high-norm tokens effectively vanish, reflecting a more balanced distribution of information.

Qualitative attention maps (Figure 1) corroborate these findings. Without registers, the model's attention is sharply focused on a few regions, reinforcing the notion that artifact tokens become dominant. In contrast, the inclusion of registers yields more uniformly distributed and interpretable attention. Collectively, these observations validate that registers effectively mitigate high-norm artifacts in the DINOv2 architecture, ensuring a healthier balance between local and global feature representation.



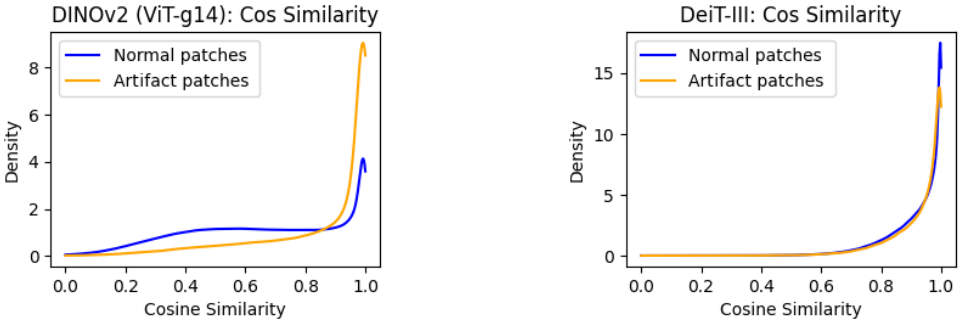
(a) DINOv2 (b) DINOv2 with registers

Figure 5: Distribution of norm values across layers for models. Registers eliminate high-norm outliers in deeper layers.

Finally, we replicate the cosine similarity analysis for DINOv2. In 6a we measure the cosine similarity of neighboring patch tokens. Here, we set the threshold for high-norm (artifact) tokens at the 98th percentile norm value, following Darcet et al. who found approximately 2.37% of tokens to be outliers. Consistent with Darcet et al. (2024), we observe that these high-norm tokens tend to be more internally consistent with some of their neighbors (i.e., close to a cosine similarity of 1.0), suggesting a strong clustering of attention in those areas. The introduction of registers helps redistribute this global context more evenly, diminishing the outlier effect and producing more stable representations.

4.2.3 Artifacts DeiT-III

We then investigated our DeiT-III model, visualized in Figures 4c and 4d. Here, the outliers in the non-register version are less present than in DINOv2, yet they still manifest as a high-norm tail. Adding registers



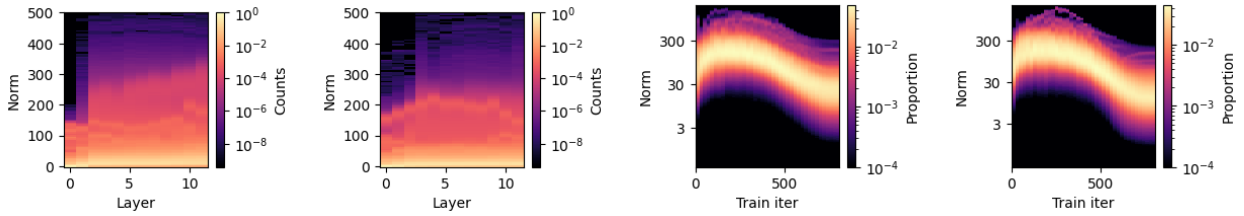
(a) DINOv2 Cosine Similarity

(b) DeiT-III Cosine Similarity

Figure 6: Cosine similarity analysis for DINOv2 and DeiT-III models.

in DeiT-III compresses the distribution, indicating a partial absorption of the global information that these outlier tokens had originally carried.

As shown in Figure 7a, high-norm tokens also appear in DeiT-III, especially in deeper layers where certain tokens accumulate excessive information, mirroring DINOv2. Introducing registers reduces both the frequency and intensity of these outliers (Figure 7b), resulting in a more controlled distribution and fewer norm spikes. However, high-norm artifacts still occur, suggesting that while registers help balance attention allocation, their effectiveness may vary in smaller models like DeiT-III Small compared to larger architectures.



(a) DeiT-III (layers)

(b) DeiT-III with registers (layers)

(c) DeiT-III (iterations)

(d) DeiT-III with registers (iterations)

Figure 7: Distribution of norm values for DeiT-III (subfigures (a) and (c)) and DeiT-III with registers (subfigures (b) and (d)). The left two plots show norms across layers and the right two plots show norms across training iterations.

To explore how these artifacts relate to the broader token space, we replicate the cosine similarity analysis on DeiT-III (Figure 6b), flagging "artifact" tokens as the top 2% highest norms. Like in the larger model, these high-norm patches appear in uniform, low-texture regions. Notably, normal patches peak near a cosine similarity of 1.0, higher than artifact patches, suggesting they form tighter clusters. While outliers capture global cues, they show weaker local alignment with neighboring patches.

This effect is further illustrated in the attention maps of Figure 1. Unlike in DINOv2, where registers consistently improved interpretability, the effect in DeiT-III is more mixed. Some attention maps become more structured with registers, while others retain scattered and concentrated activations. This variability indicates that while registers can mitigate artifacts, their effectiveness may vary depending on specific input images and training dynamics.

In figure 7, we analyze the norm distribution of the last layer throughout the training process. We compare models trained without registers (Figure 7c) and with registers (Figure 7d). Although the distribution of norm values appear visually similar for both models, DeiT-III with registers shows more compact distributions in the last quarter of training iterations. Nevertheless, the model still shows significant outliers in norm values in the first half, indicating that the registers are limited in their ability to combat high-norm tokens.

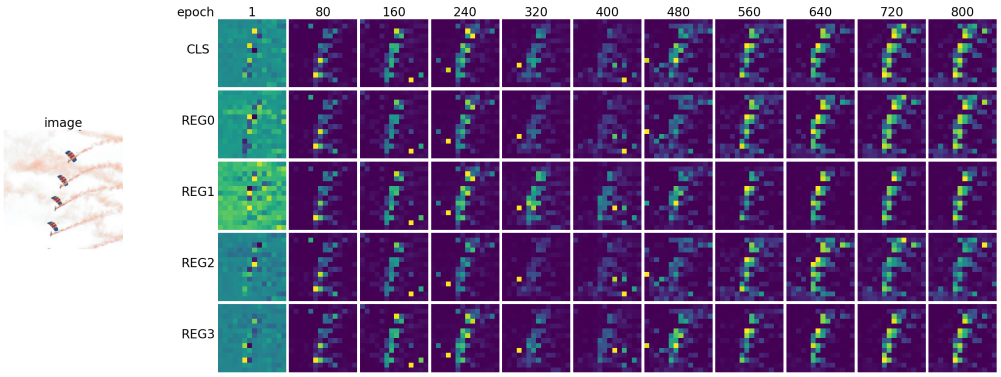


Figure 8: Attention maps of the CLS and register tokens across training epochs where artifacts are successfully removed.

Figure 8 tracks the evolution of attention maps across training epochs. Here, it can be seen that the model alternates between high-norm outliers and well-structured attention maps. This pattern indicates that attention dynamics are not entirely stable throughout training, yet the registers are often able to mitigate persistent artifacts.

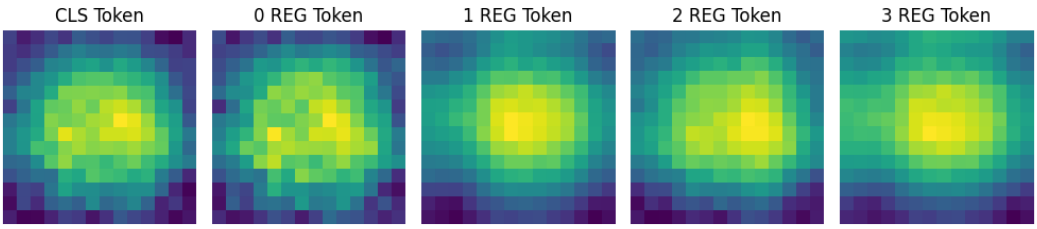


Figure 9: Averaged attention maps for DeiT-III model with registers on the Imagenette dataset. The visualization shows the positional focus of the [CLS] token and register tokens.

Figure 9 shows that in the DeiT-III model with registers, attention remains centered around the [CLS] and register tokens, helping stabilize focus but not completely eliminating local artifacts. We further measured the cosine similarity of attention maps for each register token (Figure 10) and found that, while initially overlapping, these maps diverge slightly over time, indicating register tokens develop specialized roles. This diversification may help mitigate high-norm artifacts and enrich local-global encoding, even in smaller models like DeiT-III.

Overall, these results highlight that while registers help reduce high-norm artifacts in DeiT-III, their impact is more inconsistent than in DINOv2. Some improvements in attention interpretability are evident, while certain cases still exhibit instability, particularly in deeper layers and across training epochs.

4.3 Global Features

4.3.1 Object Detection with LOST

In this section, we evaluated ViTs with registers using the LOST (Siméoni et al., 2021) method on the VOC2007 dataset (Everingham et al., 2010). Table 1 presents CorLoc values for both the pretrained DINOv2 model and our DeiT-III models, with and without registers.

Table 1 shows a small improvement in DINOv2 with registers, increasing CorLoc from 33.77% to 35.24%. Although an improvement in the CorLoc score was found in Darcet et al. (2024), the improvement shown there is more significant than what we were able to reproduce. Similarly, our DeiT-III models saw an increase from 15.13% to 17.92%. The baseline (without registers) aligns somewhat with the original paper,

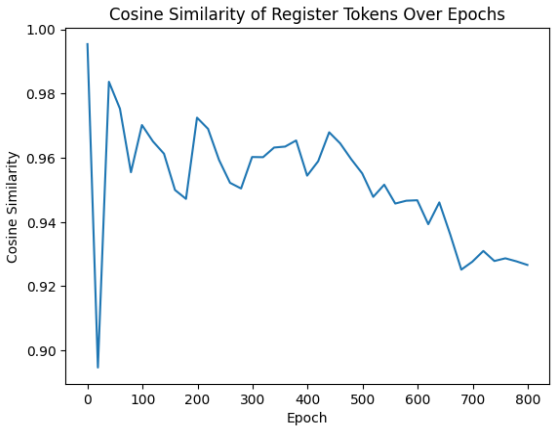


Figure 10: Mean cosine similarity between all register token attention map pairs over epochs. Lower similarity over time indicates increasing specialization among registers.

Table 1: CorLoc values of LOST on DINO, DINOv2, and DeiT-III models with and without registers.

Model	Without Registers (%)	With Registers (%)
DINOv2	33.77	35.24
DeiT-III	15.13	17.92
DINO	61.48	-

indicating that our DeiT-III model performs comparably in that setting. However, the modest improvement with registers is smaller than reported in the original study.

4.3.2 Estimation of global information in spatial patch tokens

Another important aspect of the artifacts described by Darcet et al. (2024) is their potential to encode global information. In their study, the authors theorized that high-norm tokens might carry broad, high-level information crucial for global computations. To investigate this hypothesis, they trained a logistic regressor to classify images based on individual patch tokens. Notably, high-norm tokens consistently outperformed low-norm tokens, indicating that these "artifact" tokens capture more global concepts of the image.

Table 2: Comparison of accuracy across token norm ranges for Imagenette and Caltech101 datasets.

Token Category	Imagenette Accuracy	Caltech101 Accuracy
CLS Tokens	0.6228	0.9145
High Norm Tokens	0.4974	0.8452
Low Norm Tokens	0.4959	0.8233

To replicate this, we conducted a similar experiment on our DeiT-III model. Following Darcet et al. (2024), we identified the top 2% of token norms as high-norm outliers, performing this thresholding on a batch-by-batch basis. For each image, we selected a high-norm token, a low-norm token and the CLS token, training a logistic regressor to predict the original class. Table 2 shows that high-norm tokens achieved higher accuracy than low-norm tokens, supporting the idea that artifact tokens encode more global information. However, the effect was less pronounced than in the original study, likely due to differences in training data and model scale.

5 Discussion

Our primary aim was to examine whether the main findings of Darcet et al. (2024) hold true for smaller-scale Vision Transformers, specifically a DeiT-III Small model trained on ImageNet-1k. This involved addressing three key questions: (1) Do high-norm artifact tokens also emerge in smaller ViTs? (2) Can registers effectively mitigate these artifacts in a low-parameter regime? and (3) Do the local-global encoding improvements observed in larger ViTs generalize to smaller architectures? Below, we contextualize our observations and connect them back to these questions.

High-Norm Artifacts in Smaller Models: Our experiments confirmed that high-norm artifact tokens appear in DeiT-III Small, albeit less frequently than in larger models. Similar to DINOv2 without registers, some deep-layer patches accumulated disproportionately large norms, indicating an imbalance in local-global information processing. This suggests that artifact tokens are not solely a byproduct of large model capacity but can emerge in smaller architectures as well.

Impact of Registers in Mitigating Artifacts: We found that adding registers generally reduced the prevalence and intensity of these outliers in DeiT-III, echoing the trends observed in DINOv2. While the improvement was not as pronounced as in the original paper, our results do show that registers are beneficial for smaller models as well. Specifically, both norm distributions and qualitative attention maps indicated fewer extreme spikes, aligning with the second key question about the role of registers in stabilizing smaller ViTs.

Local-Global Encoding and Downstream Performance: A key question was whether the enhanced local-global encoding capacities reported in Darcet et al. (2024) extend to smaller models. Using LOST (Siméoni et al., 2021) as a proxy for object-level understanding, we observed moderate CorLoc gains when registers were added to both DINOv2 and DeiT-III. Though smaller than the improvements reported in the original study, these findings affirm that registers positively influence local-global encoding across different model sizes.

5.1 Future work

While registers reduce high-norm artifacts in attention maps, their effect is not uniform across all images. A key avenue for future research is understanding why some inputs still exhibit artifact-like behavior while others benefit more substantially. Analyzing the relationship between image complexity, token representation, and attention dynamics could offer deeper insights into when and how registers succeed or fail. Ultimately, this could lead to architectural or training refinements that further enhance the robustness of ViTs.

References

Mikhail S Burtsev, Yuri Kuratov, Anton Peganov, and Grigory V Sapunov. Memory transformer. 2020. URL https://arxiv.org/abs/2006.11527.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024. URL https://arxiv.org/abs/2309.16588.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Mark Everingham, Luß Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge 2007 (voc2007). In *2007 IEEE International Conference on Computer Vision (ICCV)*, pp. 2429–2432. IEEE, 2010.

Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, apr 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. 2023. URL https://arxiv.org/abs/2304.07193.

David Patterson, Joseph Gonzalez, Quoc Le, Chiyuan Liang, Lluis Munguia, David Rothchild, Daiyi So, Mae Texier, and Jeff Dean. Carbon emissions and large neural network training. https://doi.org/10.48550/arxiv.2104.10350, 2021.

Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. 2021. URL https://arxiv.org/abs/2109.14279.

Shintaro Torii. Imagenette and imagewoof: A curated subset of imagenet for rapid prototyping with deep learning. https://github.com/fastai/imagenette, 2020. Accessed: 2025-01-23.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers distillation through attention. *arXiv (Cornell University)*, 2021. URL https://arxiv.org/abs/2012.12877.

Hugo Touvron, Matthieu Cord, and Herve Jegou. Deit iii: Revenge of the vit. 2022. URL https://arxiv.org/abs/2204.07118.

A Model Parameters

Table 3: Hyperparameters used for the DeiT-III Small Model on ImageNet-1k

Hyperparameter	**Value/Description**
Dataset	
Dataset	ImageNet-1k (ILSVRC 2012-2017)
Input Size	224
Number of Classes	1000
Model Architecture	
Model	`deit_small_patch16_LS`
Drop Path Rate	0.05
Training Settings	
Batch Size	256
Epochs	800
Optimizer	`fusedlamb`
Learning Rate	4×10^{-3}
Warmup Learning Rate	1×10^{-6}
Weight Decay	0.05
Scheduler	`cosine`
Warmup Epochs	5
Unscale Learning Rate	True
Data Augmentation	
Mixup	0.8
CutMix	1.0
Color Jitter	0.3
Repeated Augmentation	True
ThreeAugment	True
Regularization	
Reprob	0.0
Label Smoothing	0.0
Drop	0.0
Drop Path	0.05
Loss Function	
BCE Loss	True
Miscellaneous	
Evaluation Crop Ratio	1.0
Seed	0

In Section 3, we described our primary experimental setup for evaluating registers in Vision Transformers. Here, we provide a detailed overview of the training pipeline and hyperparameters used for our DeiT-III Small experiments on ImageNet-1k, closely following the official DeiT-III repository. Table 3 outlines each component of the setup—including data augmentation, optimization and regularization—ensuring transparency and reproducibility. We highlight key choices, such as employing 4 register tokens (when applicable) and adopting an 800-epoch training schedule. These settings closely replicate the strong baselines established by Touvron et al. (2022) while allowing us to isolate and evaluate the impact of registers under consistent conditions.

B Extended Analysis of Persistent Attention Artifacts

In the main text (Figure 8), we illustrated a scenario where the model's attention maps, although prone to transient high-norm outliers, often return to well-structured patterns due to the presence of registers. However, Figure 11 provides a contrasting case. Here, the model remains stuck in outlier states for prolonged periods, leading to persistent artifacts. Despite registers being designed to mitigate such issues, these results emphasize that their effectiveness is not guaranteed under all conditions. Factors such as initialization, optimization strategy or dataset characteristics can undermine the stabilizing role of registers. These observations motivate further exploration into methods for reliably stabilizing attention dynamics in more challenging training scenarios.

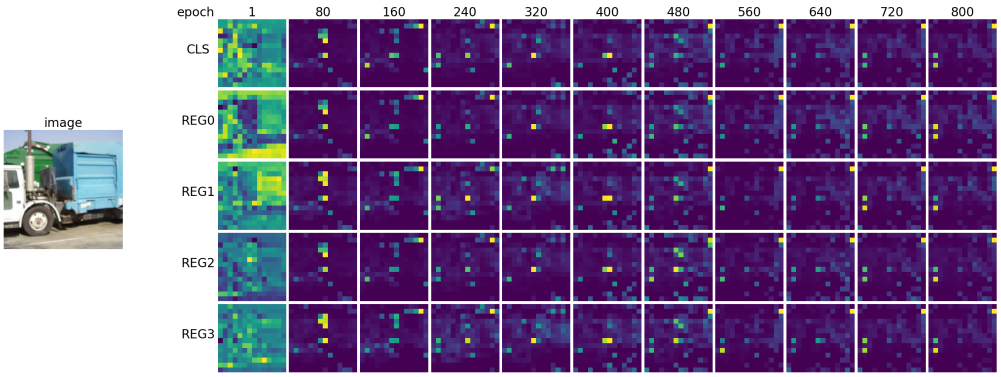
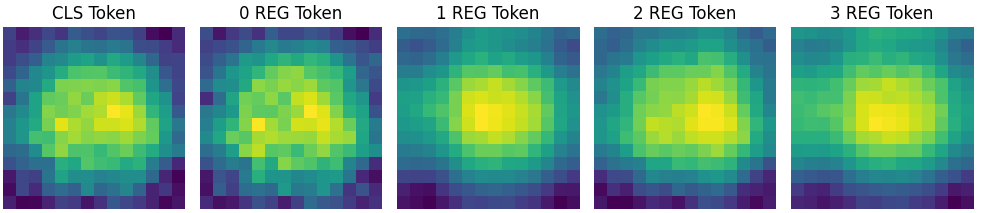


Figure 11: Attention maps of the CLS and register tokens across training epochs in a case where artifacts persist.

C Positional Focus of CLS and Register Tokens on different datasets

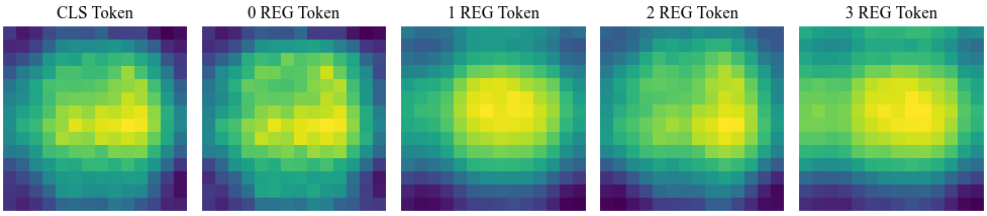
Our attention-map visualizations in Section 4 focus primarily on Imagenette. In Figure 12, we extend this analysis to additional datasets—CIFAR10, Caltech101 and Flowers102—to illustrate how registers influence CLS and patch-token attention across diverse visual domains. The CLS token consistently focuses on a central region, reflecting its role in aggregating global information. The register tokens exhibit distinct yet overlapping focus regions. The presence of registers leads to a more distributed attention pattern, suggesting that they share the burden of global information aggregation. While all registers maintain a broad attention spread, subtle variations emerge, indicating potential specialization in capturing different aspects of the image.



(a) Imagenette

(b) CIFAR10

(c) Caltech101



(d) Flowers102

Figure 12: Comparison of positional focus maps for DeiT-III models on different datasets. (a) Imagenette, (b) CIFAR10, (c) Caltech101, (d) Flowers102.

D Computational requirements and estimated carbon footprint

Each model was trained on a single node with 4 NVIDIA A100 GPUs and 72 Intel Xeon CPU cores, taking about 66 hours per model (132 hours total). Table 4 summarizes the key parameters used to estimate the carbon footprint. The Power Usage Effectiveness (PUE) is taken from the Snellius documentation[3], while the average Carbon Intensity Factor reflects the Netherlands in 2024. In the table, we list only the approximate per-GPU power draw; CPU usage is not separately itemized. Using these values, we arrive at a total carbon footprint of approximately **102.56** kg CO_2eq for training both models (following Patterson et al. (2021)).

[3] `https://servicedesk.surf.nl/wiki/display/WIKI/Snellius`

Table 4: Summary of carbon footprint for reproducing our model training, following Oquab et al. (2023). Here, *GPU-hours* is the product of the number of GPUs (4) and total hours (132). The power consumption (400 W per GPU) is an approximate average draw, excluding CPU usage.

GPU Type	**GPU Power (W)**	**GPU-hours**	**PUE**	**Total Power (MWh)**	**Carbon Emitted (tCO_2eq)**
NVIDIA A100	400	528	1.2	0.23	0.10